# PERCEPTION INCEPTION

## Preparing for deepfakes and the synthetic media of tomorrow

Curtis Barnes
Tom Barraclough

BRAINBOX
EMERGING TECH + LEGAL RESEARCH

The Law Foundation
NEW ZEALAND
SUPPORTING INDEPENDENT LEGAL THINKING.

The mosaic on our cover page contains no photographs. All faces are synthesised from digital data. They were produced by computers, not cameras. None depict real people, and any resemblance to real people is entirely random.

The individual "faces" were produced by generative adversarial networks, a technique in machine learning and artificial intelligence, in an application provided publicly at <thispersondoesnotexist.com>.

The "landscapes" were produced from digital doodles, then processed by similar techniques in the publicly available NVIDIA "GauGAN" application. None depict real places.

The small, constituent images convey the essence of pixels, of which all visual digital information is comprised. Each pixel has a numerical value. Computers can interpret these numerical values to generate new images, like all of those present on our cover page.

The background image apparently showing President Obama is also a synthetic image, produced via a range of artificial intelligence and post-production techniques. It is not a "real" photograph. All credit for the underlying image goes to BuzzFeed and Monkey Paw Productions. This particular synthetic image is taken from a frame of a "deepfake" style video produced by BuzzFeed News as a public service announcement about fake news and disinformation. This video has now been widely disseminated and discussed as evidence of the forthcoming deepfake phenomenon and its risks. The video is available on YouTube at <https://youtu.be/cQ54GDm1eL0>.

ISBN: 978-0-473-48214-5

# Acknowledgements

This work would not have been possible without a grant from the New Zealand Law Foundation – Te Manatū a Ture o Aotearoa and its Information Law and Policy Project. We are grateful for the Foundation's vision and generous support.

Additional support was received from the New Zealand Law Foundation Centre for Law and Policy in Emerging Technologies at the University of Otago.

We are also grateful for the support of the AR/VR Garage, an initiative by the Auckland Tourism Events and Economic Development agency (ATEED).

Our research has also benefitted from discussion with a wide range of stakeholders in industry, academia and government. While we have chosen not to name them here, we are grateful to them for sharing their time and knowledge with us.

Curtis Barnes and Tom Barraclough
Co-authors
Directors, Brainbox Limited

17 May 2019

# Foreword

The camera never lies.

We've long known the aphorism to be, at best, an overstatement. Almost since the birth of photography, evidence has existed that it is capable of distortion and manipulation. Doctored pictures exist from as early as the American Civil War – 'honest' Abe Lincoln was himself a beneficiary of photographic sleight of hand.

Many of these can be benign, playful. Like many children, I delighted in the fun fear of 'ghostly' double exposures. Laughed as forced perspective techniques allowed us to tower over distant landmarks or hold 'tiny' relatives in the palms of our hands.

But 'trick photography' has also had less innocent uses. We worry about the effects of air-brushed and re-touched pictures of models and actors on unrealistic and unhealthy body image among young people. In the political sphere too, the risks are long established. Stalin is believed to have excised evidence of Trotsky's presence from Lenin's speeches.

The technologies discussed in this important report have not emerged from nowhere. They are the direct descendants of the photo-manipulation tricks that have been with us for a century and an half. Yet there is a widespread sense that, in recent years, something has changed. The deceits of the past look simplistic and naïve compared with the sophistication of AI-driven 'deep fakes. Now, words can be put into politicians' mouths as easily as faces pasted onto bodies. The identities of participants can be altered, not only in static photographs, but in moving – and sometimes highly compromising – footage. These can range from the innocently amusing to the reputationally disastrous. And if the editing isn't yet entirely seamless, it's already good enough to fool many casual viewers. And potentially, many voters.

The risks are becoming obvious – not only to individual reputations but to democratic processes.

But not every harmful practice is amenable to a legal solution. And not every new risk needs new law. The role law should play in mediating new technologies and mitigating their dangers is a matter that merits careful evaluation, weighing costs and benefits, figuring out what's likely to work rather than merely look good. There's also the danger of regulatory disconnection, when the technology evolves to outpace the rules meant to govern it.

I was delighted, then, to learn that the Law Foundation was funding this work, the latest in a series of bold, forward-looking and original projects it has supported. Delighted too to learn who would be writing it. Tom Barraclough and Curtis Barnes are two of New Zealand's brightest young legal minds, fascinated by the future, imaginative about the law's potential role in it. They are exemplars of generation of lawyers that I've watched emerge over the past decade or so – tech-savvy, morally invested in New Zealand's future, broad-minded about innovative solutions.

That sort of talent pool, and this sort of work, are going to be massive and indispensable assets for this country, as it adapts to an accelerating wave of technological change, and all that's carried in its wake.

The camera does lie. In more sophisticated, better concealed ways than it ever has before. The question is what we can do about it. This report is a first, vital step to working out an answer.

Dr Colin Gavaghan
Associate Professor of Law, University of Otago
New Zealand Law Foundation Chair in Emerging Technologies

# Executive Summary

## Background

1. New audiovisual technologies can produce increasingly realistic images, sounds and videos by creating and manipulating digital data using computers.

2. These representations can make it look and sound like something happened when it did not, or that it happened differently than it did. Industry-grade visual and audio effects technologies can achieve the same thing, however new audiovisual technologies present new legal and social issues. This is because of:

    a. the rapid speed by which information may be produced;

    b. the lower-cost of production;

    c. the comparatively lesser degree of skill and experience required;

    d. the greater use of automation in production;

    e. the kinds of things that can be represented to a high quality, including the human face and voice;

    f. the new ways in which audiovisual information is being consumed, and the greater volumes of consumption.

3. These technologies have huge potential benefits, but they also have risks. Assessment of these risks will require ongoing cross-disciplinary analysis. From a legal perspective, emerging audiovisual technologies may be used to deceive or mislead. Public awareness of this risk of deception has grown through discussion of one kind of emerging audiovisual technology known as "deepfakes". The existence of such technologies may undermine general trust in audiovisual information to some degree.

4. The rate of development and commercialisation of emerging audiovisual technologies is rapid. New technologies are constantly arising. Deepfakes are only one example within a wider family of such technologies. We refer to this family of technologies as "synthetic media".

5. Some harmful uses of synthetic media technologies are already taking place: for example, the creation of pornography in which non-consenting people are represented in pornographic videos. Other examples include creating convincing fake profile pictures for social media accounts to spread disinformation and gain access to private information. There is also the prospect of deepfake videos causing political confusion.

6. As a result, there is increasing international attention to the risks, threats, and harms of synthetic media. There is increasing support for intervention: technological, social, and legal. There is growing dialogue over the role of law in any intervention.

7. At the same time, there is wide interest in the potential social benefits and uses of synthetic media technologies, which should be taken into account by policymakers.

8. In the course of a 9 month legal research project funded by the New Zealand Law Foundation, we have investigated the technologies, scientific materials, and law associated with synthetic media. This has involved speaking to international and domestic stakeholders and industry leaders to understand the current capabilities of the technology and where it is going.

## Conclusions

9. We anticipate that synthetic media will continue to improve, becoming better and more accessible. We think it likely that in the near future, consumers and citizens will be regularly exposed to audio, images and video that looks or sounds as if it is a reliable representation of factual events, even though it is not. It is information that gives the impression that it was "captured", when in fact it was "constructed" to a greater or lesser extent. Lots of this information will be benign or beneficial, but some of it will be harmful.

10. We believe that the best way to approach synthetic media as a legal subject is through a framework approach, which can help identify the common elements of synthetic media technologies and the existing law that applies to them. Using this approach we were able to identify a large body of law across existing legal regimes in New Zealand that is likely to be applicable to potential harms arising from synthetic media, including:

    a. Privacy Act 1993;
    b. New Zealand Bill of Rights Act 1990;
    c. Broadcasting Act 1989;
    d. Electoral Act 1993;
    e. Films, Videos and Publications Classification Act 1993;
    f. Defamation;
    g. the guidelines of the Media Council of New Zealand;
    h. Human Rights Act 1993;
    i. Crimes Act 1961;
    j. Harmful Digital Communication Act 2015;
    k. Harassment Act 1997;
    l. Fair Trading Act 1986, and the Advertising Standards Authority;
    m. Copyright Act 1994;
    n. Evidence Act 2006;
    o. Social Media platform guidelines.

11. We are not convinced that enacting substantial new law is either necessary or the best way to address the harms that may be generated by synthetic media. We also identify a risk that, where new law goes beyond existing law, it may abrogate rights of freedom of expression. Synthetic media is a means of expression like many others.

12. All digital media is, to some degree, manipulated by digital technologies. For that reason, we think any policy approach that attempts to distinguish between "fake" media and "real" media is unsustainable. Instead, policymakers should focus their attention on the degree to which a synthetic media artefact has been manipulated by digital technologies, and the degree to which that manipulation enhances or undermines its reliability as a record of actual events. Our framework is intended to assist with this inquiry.

13. Importantly, it is possible that enforcement of existing law may be impeded by a range of practical difficulties. These will impinge upon access to justice where a harm or wrong is alleged to have occurred. We note that these difficulties are likely to continue to arise even where any new law is enacted. These practical difficulties include:

    a. a limited supply globally of experts and services necessary to meet the evidential needs generated by much of the existing law;

    b. the costs of legal intervention and other access to justice barriers, particularly where cases are of low financial value;

    c. the comparatively long timeframes for legal investigation and action, and the comparatively rapid speed and scale of harms caused by the dissemination of synthetic media;

    d. the practical difficulty of identifying the agent(s) accountable for a given synthetic media artefact – and their degree of culpability – given the many ways an artefact may be produced.

14. We note that it is not always necessary to show that a piece of synthetic media is "fake" for the law to intervene. Many legal regimes control the use of synthetic media without the need to show that it is an unreliable record of events, for example where the content is objectionable or it is disseminated in harmful ways.

## Recommendations

15. There are a wide range of legal and pseudo-legal regimes touching upon the potential harms caused by the creation, content and dissemination of synthetic media. In particular, we have identified regulation dealing with harms through the lens of privacy law, criminal law, electoral law, property and copyright law, and broadcasting law.

16. Synthetic media can be used in a vast number of ways, both positive and negative. As a result, this report can only be a starting point. We encourage closer ongoing investigation into this area by collaboration between legal and technological subject matter experts.

17. We recommend caution in developing any substantial new law without first understanding the complex interaction of existing legal regimes. Before acting, it is essential to continue to develop an understanding of how these regimes apply to factual scenarios as they arise. Where new law is necessary, it is likely to take the form of nuanced amendment to existing regulation. For now, existing legislation should be given the opportunity to deal with harms from synthetic media technologies as they arise.

18. Any new legislation must take the position that synthetic media technologies and artefacts touch upon individual rights of privacy and freedom of expression, deserving careful attention from policymakers and broad public consultation. There are benefits, risks, and trade-offs to be discussed in deciding whether to allocate responsibility for restricting synthetic media technologies to the State or to private actors. Human rights, the rule of law, natural justice, transparency and accountability are essential ingredients in whatever approach is adopted.

19. There is a risk that the issues resulting from synthetic media will be lost among the wide range of statutes and agencies involved. It may be unclear which agency is responsible for any given synthetic media artefact, and under what legal regime. Accordingly, agencies and stakeholders responsible for the legislation covered here should take the following steps.

    a. First, formulate agreement on the conclusions in this report and their respective responsibilities for the use of synthetic media technologies and artefacts, as defined by our framework.

    b. Secondly, collaborate to issue public statements on their respective responsibilities for the harmful uses of synthetic media technologies, with the goal being to:

        i. provide commercial certainty to actors operating in New Zealand generating artefacts through synthetic media technologies; and

        ii. facilitate access to justice for those people alleging harm or loss by educating legal professionals and members of the public about the remedies available.

    c. Thirdly, in light of their conclusions above, consider how to best publicise the potential impact of synthetic media technologies in a way that:

        i. does not cause undue scepticism about audiovisual information generally; and also

        ii. increases the chance that individuals will exercise appropriate caution before relying on audiovisual information in a way that generates risk of harm.

    d. Fourthly, consider their need for and access to a range of digital forensic services in relation to audiovisual information. In doing so, agencies should note whether private entities can also

gain access to these services. Complaints volumes can be limited by increasing access to evidential services in a way that avoids unnecessary dispute about the reliability of audiovisual information and therefore facilitates dispute prevention.

20. The New Zealand Government, along with New Zealand's technology and visual effects sectors, should consider the opportunities for New Zealand in building capacity for digital forensics and expert evidential services to international markets, given New Zealand's strength in the innovation and use of synthetic media technologies.

21. Pursuant to its functions at s 13 of the Privacy Act, the Office of the Privacy Commissioner should initiate public discussion on the extent to which someone has a reasonable expectation against the creation of synthetic media artefacts about that person without their consent, and the extent to which the creation of such synthetic media artefacts might be considered offensive to a reasonable and ordinary person.

22. The legislature should consider and make amendment to s 216G of the Crimes Act clarifying whether it is primarily an offence against category 1 capture technologies or category 2 manipulation technologies. Stakeholders should be given the opportunity to have input because of the criminal penalties being imposed and the potential infringement on the New Zealand Bill of Rights Act from broad drafting.

23. The review of the Copyright Act 1994 should account for condition 2 of our framework (multiplicity) and the greater use of category 2 digital manipulation technologies in the synthesis of audiovisual artefacts.

24. Apart from existing Copyright protections, New Zealand should not adopt a property-based framework for restricting unauthorised use of an individual's audio-visual profile and should instead prefer a policy response based on individual privacy.

25. Further legal and policy research should be done on the interaction between the law of copyright, privacy and freedom of expression in New Zealand when an individual authorises the use of generative synthetic media technologies to create new synthetic media artefacts about them.

26. The New Zealand government should consider how it can use New Zealand's strengths in effective policy and synthetic media technologies to benefit the international community and facilitate positive international relationships with state and non-state actors.

27. Any individual or agency generating or disseminating synthetic media technologies, or synthetic media artefacts that are highly photo- or phono-realistic, should exercise extreme caution and consider how to affix statements or contextual indicators that make it clear how far category 2 manipulation technologies have been deployed and the extent to which a synthetic media artefact is (or is not) the result of a category 1 capture process.

# Introduction: A new wave of audiovisual information

Today it is possible to use photographs, videos, or voice recordings of an individual to produce models which can generate representations of that individual doing and saying things they never did. These technologies are rapidly advancing. They have many commercially valuable and beneficial applications. They also have obvious potential for harmful or deceptive use.

As citizens and consumers, it is likely that many if not most New Zealanders are unaware of the capabilities of new synthetic media. Few are familiar with the meaning of terms like "deep fakes", even while overseas lawmakers, policymakers, and mainstream media conduct serious ongoing debates on the subject. Few New Zealanders are likely to be aware that at present, from around five to ten minutes of video or twenty to thirty minutes of audio, a skilled person using consumer-level computing technology could create relatively realistic representations of the Prime Minister engaged in entirely untrue behaviour, or saying totally fabricated things. With enough video or photographs, perhaps taken from a personal Facebook or Instagram account, a skilled person could even produce such misleading material of an everyday New Zealand citizen. This adds another layer of complexity to the discussion about misinformation and so-called "fake news", privacy, identity, and a further element to the possibility of foreign interference in domestic politics, already the subject of a Select Committee investigation in New Zealand.

As creators New Zealand companies are world leaders in audiovisual effects technologies. The services of Weta Digital, for example, are highly sought after across the world and have been central to countless blockbuster films. Companies like Soul Machines sit at the forefront of applied computing in their field, pushing the boundaries of animation technologies.

At the same time, New Zealand is often considered a prime location for the development and testing of new policy. It's population size, demographic, and regulatory environment are generally conducive to this purpose. The country has a history of leading the way in unorthodox initiatives, evidenced historically by things like the Accident Compensation Scheme, and more recently, the Harmful Digital Communications Act 2015.

New Zealand is also fortunate to have the opportunity to prepare pre-emptively for this forthcoming technological phenomenon before it arrives in earnest. Other nation states have not been so lucky. For example, overseas jurisdictions are already being forced to confront objectionable phenomena like "non-consensual pornography", whereby the faces of non-participant individuals are realistically synthesised onto the bodies of pornographic performers – a new, challenging slant on phenomena like "revenge porn" or "intimate visual recordings", with which New Zealand law is already concerned.

The fact that New Zealand lags slightly behind with regards to technological phenomenon like this does not mean it will remain unaffected forever. With the pace of advancements in synthetic media and a forthcoming proliferation of the technologies by which such artefacts are produced, it is unlikely to be long before New Zealand is confronted by some of the deceptive and harmful uses that this kind of audiovisual information facilitates.

The combination of these factors puts New Zealand in a unique position to lead in developing robust law and policy for emerging synthetic media so that its creators can continue to innovate, and its consumers and citizens can continue to act with confidence and due criticality. Arguably there is an imperative for New Zealand to lead in this regard given its opportunity to do so, its history of leadership in policy, and its access to several of the world's most preeminent digital audiovisual information creators.

Nevertheless, this does not necessarily mean that New Zealand needs new law, although we do not rule this out. What it first requires is a robust understanding of the technologies and existing law so as to know what harms law can and cannot respond to effectively. Moreover, there may be no reason to believe that new law would be any more effective than what we have now, or at least no reason to believe that it would not be subject to the same limitations. These are predominantly limitations of service delivery and resource availability, rather than flaws or gaps in the law itself. Where gaps do exist, they may sometimes

be intentional so as to ensure that civil and human rights are not unjustifiably limited. Where flaws do exist, these will only be revealed by careful legal analysis of existing statute and common law, and later, through the hammer and tongs of individual cases as new scenarios arise.

There is also a strong imperative to get this right. In responding to the potential of new synthetic media, poor policy generates serious risks. These risks may rival or surpass the very risks that policy is intended to mitigate. Broad, ambiguous, or overbearing legislation threatens to undermine not only the innovative capacity of New Zealand's audiovisual effects industry, but also the ability of its citizens to exercise their right to freely express and exchange information. As such, the pace of policy development must be both proactive and cautious – a challenging feat. This in turn requires a conceptual way of thinking about synthetic media so that it can be analysed and understood in an effective, efficient, and consistent way. Developing this method will allow New Zealand policymakers to respond effectively and proportionately to new synthetic media technologies as they inevitably arise.

Currently, such technologies seem to emerge at a frenetic, almost week-to-week pace. The first impulse is to respond radically to each new technology, treating it as entirely distinct, and therefore not possibly subject to existing rules or dealt with in the same way as we deal with other things. This temptation is doubly strong for audiovisual technologies, which by their nature produce artefacts that are sensational and illusory. Nevertheless, we believe such a course of action would be imprudent. In fact, based on our analysis we believe all digital audiovisual technology shares commonalities with which the law is already deeply concerned. There are touchstones that law can recognise, and actions it can respond to. For this reason we develop a framework for understanding synthetic media artefacts and their interaction with the law. We apply it to New Zealand as a guide to future policy for synthetic media.

# Part 1: Emerging audiovisual technologies and synthetic media

## New technologies, new information, new potential

28. This is a research project about the legal and social implications of technologies that capture, manipulate, display and disseminate digital audio-visual information.

29. The concept of "emerging audiovisual technologies" is deliberately broad. What is currently emerging will eventually become emerged, with no clear indication of when transition occurs. Moreover there will be new technologies tomorrow the seeds of which have scarcely been planted today. What then constitutes an "emerging audiovisual technology" is a matter open to interpretation and subject to change. The phrase is intended to represent a general technological phenomenon, rather than to be used as a rigid yardstick.

30. The essence of emerging audiovisual technologies is that they allow for the creation of remarkably realistic representations, often in ways that are faster, cheaper, and more accessible than previously possible.

31. Looking to the near future, some of these technologies may even disrupt the entrenched ways that we consume audiovisual content: things like "augmented reality" and "mixed reality", for example. In the traditional paradigm of audiovisual information, content is consciously consumed. There is a clear and unambiguous distinction between the real world and the virtual content being consumed. "Liminal" audiovisual technologies like augmented reality actively alter, augment or manipulate the "real" environmental data we detect with our eyes and ears. Through these technologies, some of the light projected onto the back of the consumer's eye will be reflected from the physical world and some will be virtually inserted.[1] With others, some sound waves striking the eardrum will be reflected from the acoustic environment and some will be enhanced or suppressed.[2] In each instance, the consumer may only be partially aware of what has occurred. What was derived from the "real" environment, and what was artificial? Ultimately, these liminal technologies may become a bridge between the real and virtual worlds that we currently occupy to an increasingly equal degree.

32. The immediate focus for policymakers must be the proliferation of new information that makes it look or sound like something happened when it did not happen. Our interest in this project is limited to technologies of "light" and "sound" as humans detect and perceive these things, excluding, for example, haptic (touch) perception. Light and sound energy are the substance of the images people view, the videos they watch, and the music or podcasts they listen to. However, most new technologies do not merely capture a record of a single moment of light or sound energy, but rather 'construct' a composite record through multiple inputs and manipulations. This process of manipulation and synthesis, even in the common digital cameras in our smartphones, makes it difficult to distinguish between "real" and "fake artefacts purely on the basis of digital manipulation alone. "Fake" cannot simply mean "manipulated" and "manipulation" cannot automatically be understood as harmful or deceptive. Most modern audiovisual media is digital, and digital media involves synthesis. For this reason, we refer to such things in a broad way as "synthetic media". We expand upon this definition in our framework.

33. When we say "synthetic media", we essentially mean audiovisual information in digital form. Often that media is a composite of multiple pieces of information synthesised to produce a substantially

[1] See, for example, products like Magic Leap: <https://www.magicleap.com/>; and Google Lens <https://lens.google.com/>.
[2] See, for example, noise-cancelling and enhancing headphones, as reviewed here: <www.theverge.com/reviews/2018/7/12/17032058/>.

new informational artefact. Even individual artefacts require digital processing in order to be useful to human beings, taking the outputs of sensors and reconstructing them into audiovisual outputs. Many people may be surprised at the amount of audiovisual information they consume that can be described as synthetic media – information which is constructed to the same extent it is captured, sometimes more so. This arises predominantly from misconceptions about the way modern digital audiovisual information is created. For the avoidance of doubt, we argue that even devices like modern digital cameras are synthetic media technologies, and that the photographs they produce are synthesised to a greater or lesser degree: they are composites constructed from data which may be collected by multiple sensors. The light energy captured by the sensors is converted to digital data, and in this process, a certain amount of computational manipulation of the digital data is inherent. The device might also apply various manipulations which are intended to alter or enhance the information, filtering out background noise, making it more visually pleasing to the end-consumer, removing things like "red eyes", and so on. The synthetic nature of the technology may be more easily understood to the user when multiple sensors are obvious on the capture device. For example, some smartphones now have multiple cameras working in tandem – as many as four in some cases.[3]

34. An excellent albeit unusual analogy by which to explain synthetic media is the recent "photograph" of a supermassive blackhole, the first of its kind.[4] This photograph is in fact a composite synthesised from a planet-scale array of eight different telescopic sensors located around the globe and acting in synchronicity. Not only that, the sensors used were radio telescopes detecting radio waves that are imperceptible to the human eye, converting electromagnetic information into digital data that was then used to generate a visual image:[5]

> The EHT observations use a technique called very-long-baseline interferometry (VLBI) which synchronises telescope facilities around the world and exploits the rotation of our planet to form one huge, Earth-size telescope observing at a wavelength of 1.3 mm. VLBI allows the EHT to achieve an angular resolution of 20 micro-arcseconds — enough to read a newspaper in New York from a sidewalk café in Paris.
>
> The telescopes contributing to this result were ALMA, APEX, the IRAM 30-meter telescope, the James Clerk Maxwell Telescope, the Large Millimeter Telescope Alfonso Serrano, the Submillimeter Array, the Submillimeter Telescope, and the South Pole Telescope. Petabytes of raw data from the telescopes were combined by highly specialised supercomputers hosted by the Max Planck Institute for Radio Astronomy and MIT Haystack Observatory.

35. This blackhole image is particularly useful for understanding the nature of synthetic media. It illustrates that even where an image is highly synthesised and manipulated, incorporating human interpretation of information from multiple different sensors, this does not make it inherently unreliable, or deceptive, or false. In fact, such a process can still lead to something sufficiently reliable that it informs further scientific discovery. It is not enough to say that audiovisual information is synthesised or manipulated: the real question is the extent of that digital manipulation and how this has affected the artefact's reliability for its intended purpose.

36. Advancements in these computational processes of compositing, manipulating, and synthesising are the basis of a range of new technologies. It is predominantly what allows them to take light and sound energy, and from it generate remarkable new image and audio information. In large part this is achieved because, as discussed above, the vast majority of light and sound information is now stored as digital data. Digital data is, primarily, symbolic language that can be interpreted by computers: generally alphanumeric. It is difficult for computer systems to interpret "images", but it is relatively much easier for them to interpret numbers and letters. Through advanced computer science techniques, including the use of artificial intelligence like machine learning and neural

[3] Chris Welch "Samsung Galaxy's S10 has up to six cameras: here's what they all do" (20 February 2019) The Verge <www.theverge.com/2019/2/20/18233130/>.
[4] "Astronomers Capture First Images of a Black Hole" (2019) Event Horizon Telescope <www.eventhorizontelescope.org/>.
[5] Ibid.

networks, computers can be used to do remarkable things with the digital datafiles containing the numbers and letters that represent audio and images.

37. They can, for example, synthesise new audio and images. Sometimes these may bear remarkably close resemblance to the original subjects – i.e. the person or thing which was the subject of the original image and audio. They can even do this with video, to the extent that video is only a sequence of still images displayed in rapid succession.[6] For policymakers, this elementary description adequately describes so-called "deepfakes", a title often applied to both video and audio artefacts that have been synthesised from existing digital data by means of "deep learning" neural network models, part of the wider family of machine learning artificial intelligence techniques.[7]

38. Technologies like these have a range of very valuable and beneficial uses.[8] At the same time, they have a deceptive capacity to the extent that they can be used to generate high-quality visual and audio representations of things that never happened. An average consumer might think these sorts of representations are impossible, and therefore tend to consume them uncritically. In New Zealand, our experience is that most people are yet to encounter these sort of audiovisual artefacts – deepfake videos, synthetic speech, and so on – despite their increasing commercialisation.

39. The problem arising from synthetic media technologies is summarised in this interview from ABC News with Dr Matt Turek:[9]

> "The challenge is it goes to the heart of our trust in visual media," Dr. Matt Turek, head of the media forensics program at the Defense Advanced Research Projects Agency, run by the U.S. Department of Defense, explained. "We're used to looking at an image or video and putting faith in it -- believing the content of the image or the video. And with the rise of the ability to easily manipulate those, that is going to put our faith in visual media in jeopardy."
>
> Deepfakes began sparking widespread concern last year, when Reddit users began posting fake pornographic videos online, primarily targeting actresses like "Wonderwoman" star Gal Gadot, superimposing the superhero's face onto X-rated content without her permission.
>
> The early fakes were riddled with glitches, but as that technology continues to evolve, some worry they could become indistinguishable from the real deal – potentially swaying elections, triggering widespread panic, riots – or even a war. It's these worst-case scenarios that have caught the attention of many public officials, from lawmakers to the Department of Defense.
>
> "A lot of times there are some indicators that you can see, particularly if you are trained or used to looking at them. But it is going to get more and more challenging over time, so that is why we developed the media forensics program at DARPA," Dr. Turek said.

40. Deepfakes are just one kind of synthetic media technology. We note that the term is being used in a way that expands beyond its narrow, original meaning. "Deepfake" has generally become both a talisman for broader claims of "information apocalypse" and "fake news", as well as a pithy catch-all for any form of audiovisual falsity or manipulation.[10] The *New Zealand Listener* magazine, for example, ran a special issue under the title "Deep Fake" that made only the most cursory mention

---

[6] The number of still images per unit of time of video is referred to as "frame rate". Newer cameras tend to have much higher frame rates than older cameras. Higher frame rates may increase realism, but nonetheless appear less persuasive to consumers who are more used to the distinctive visual effect of lower frame rates.

[7] See Jurgen Schidhuber "Deep Learning in Neural Networks: An Overview" (2015) 61 Neural Networks 85–117.

[8] See, for example "Wavenet and other synthetic voices" by Google <https://cloud.google.com/text-to-speech/docs/wavenet>: "A WaveNet generates speech that sounds more natural than other text-to-speech systems. It synthesizes speech with more human-like emphasis and inflection on syllables, phonemes, and words. On average, a WaveNet produces speech audio that people prefer over other text-to-speech technologies".

[9] Shannon K. Crawford, Kyra Phillips, Allie Yang "Seeing but not believing: Inside the business of "deepfakes" (10 December 2018) ABC News <www.abcnews.go.com/Technology/believing-inside-business-deepfakes/story?id=59731790>.

[10] Charlie Warzel "He Predicted the 2016 Fake News Crisis. Now He's Worried About An Information Apocalypse" (11 February 2018) Buzzfeed News <www.buzzfeednews.com/article/charliewarzel/the-terrifying-future-of-fake-news>.

of the titular subject, instead bringing the entire scope of "fake news" and online misinformation beneath the umbra of the term.[11] This is just one indicator that the boundaries of the term have grown to include most or all digital media, including conventional "fake news", as well as things like "fake text generators", including the application created by OpenAI which was quickly dubbed "deepfakes for text".[12]

41. This creates real problems when it comes to any suggestion that "deepfakes" should be dealt with by new legislation. New, prohibitive regulation for deepfakes risks restricting a much wider range of legitimate, protected expression. This should cause lawmakers to pause. The use of unstable terms to describe developing technologies in contested subject areas is risky. Current terminology lacks the stability and specificity necessary to form load-bearing policy concepts. In essence, legislation which purports to ban "deepfakes" might later be applied or interpreted to include a range of artefacts it was never intended to capture, including many other forms of synthetic audiovisual technologies and media. Alternatively, the law may come to be applied selectively, only enforced to prohibit particular synthetic media artefacts that fall foul of the orthodoxy of a given day.

42. With regards to these technologies, it is easier to show what they can do than to describe their capabilities with words. They are, after all, the stuff of light and sound rather than language. Here are some examples with links provided in the footnotes:

    a. A synthesised representation of President Barack Obama;[13]

    b. NVIDIA's "GauGAN: Changing Sketches into Photorealistic Masterpieces";[14]

    c. NVIDIA's "Image Inpainting" tool for semi-automated rapid editing of images;[15]

    d. Adobe's "Content-Aware Fill",[16] for removing unwanted features from video;[17]

    e. "Do as I Do" motion transfer,[18] and other applications which allow movement to be transferred from a source to a target;[19]

    f. "Face2Face" real-time capture and reenactment,[20] and other methods allowing for real-time animation of one person's facial expression onto a representation of another person.[21]

[11] Gavin Ellis "Deep Fake" *New Zealand Listener* (New Zealand, 16 February 2019) at 14.
[12] Alex Hern "New AI fake text generator may be too dangerous for release, say creators" (14 February 2019) The Guardian <https://www.theguardian.com/technology/2019/feb/14/elon-musk-backed-ai-writes-convincing-news-fiction>.
[13] BuzzFeed "You Won't Believe What Obama Says In This Video!" (17 April 2018) <www.youtube.com/watch?v=cQ54GDm1eL0&feature=youtu.be>.
[14] NVIDIA "GauGAN: Changing Sketches into Photorealistic Masterpieces" (18 March 2019) <www.youtube.com/watch?v=p5U4NgVGAwg&feature=youtu.be>.
[15] NVIDIA "Research at NVIDIA: AI Reconstructs Photos with Realistic Results" (22 April 2018) <www.youtube.com/watch?v=gg0F5JjKmhA>.
[16] Adobe "Remove objects from your videos with the content-aware fill panel" <https://helpx.adobe.com/nz/after-effects/using/content-aware-fill.html>.
[17] Robert Hranitzky "How to remove objects in video with Content-Aware Fill in Adobe After Effects" (3 April 2019) <www.youtube.com/watch?v=gg0F5JjKmhA>.
[18] Caroline Chan, et al. "Everybody dance now." (2018) <arXiv preprint arXiv:1808.07371>.
[19] Caroline Chan "Everybody Dance Now" (22 August 2018) <www.youtube.com/watch?v=PCBTZh41Ris&feature=youtu.be>.
[20] Matthias Niessner "Face2Face: Real-time Face Capture and Reenactment of RGB Videos (CVPR 2016 Oral) 17 March 2016 <www.youtube.com/watch?v=ohmajJTcpNk&feature=youtu.be>.
[21] Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., & Nießner, M. "Face2face: Real-time face capture and reenactment of rgb videos." (2016) *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2387-2395.

g. Lyrebird,[22] which creates synthetic voices from audio data, and other applications by which individual voices can be digitally replicated;[23]

h. "Houdini" procedural world generation, like that used in Ubisoft's "Far Cry 5", to generate more physically accurate, natural-looking virtual environments;[24]

i. Rokoko motion capture suits,[25] which are a step towards democratization of previously less accessible motion capture technology;[26]

j. AI-generated art as a growing artform, including a 'portrait' which sold for $432,500 USD. "AI Art at Christie's Sells for $432,500";[27]

k. Synthesia,[28] and other applications using synthetic media technologies to generate video that allow for things like synchronizing lips and facial movement to speech in different languages;

l. Highly realistic synthetic media images of human faces;[29]

m. "Digital humans", like those produced by Soul Machines (based in New Zealand).[30]

## The problem

43. As the above examples demonstrate, emerging audiovisual technologies can create a digital record that makes it look like something happened when it never did. This does not necessarily cause harm. It also has huge benefits for creative and communications industries. It nevertheless generates the potential for deceptive impacts on consumers and citizens who assume that an audiovisual record is a reliable indication of real events. This risk is even greater for those who are otherwise naive to the possibility that it is synthetic and manipulated, and further amplified by the possibility that the audiovisual record is accompanied by explicit or implicit statements about its reliability, or where presented in a context that suggests authenticity. In terms of the scale of this risk, many commentators are highly concerned:[31]

> "Deepfakes have the potential to derail political discourse," says Charles Seife, a professor at New York University and the author of Virtual Unreality: Just Because the Internet Told You, How Do You Know It's True? Seife confesses to astonishment at how quickly things have progressed since his book was published, in 2014. "Technology is altering our perception of reality at an alarming rate," he says.

44. The phenomenon has been described as "a looming crisis" by American legal scholars Professor Robert Chesney and Professor Danielle Citron.[32] Citron also describes the latent disruption that might arise from a proliferation of realistic but non-veridical media, or what they call the "liar's dividend: when nothing is true then the dishonest person will thrive by saying what's true is fake." Mainstream media has driven this sentiment further, perceiving it as an extension of a wider "fake

[22] See: <www.lyrebird.ai>.
[23] Lyrebird "Lyrebird - Create a digital copy of your voice" (4 September 2017) <www.youtube.com/watch?v=YfU_sWHT8mo&feature=youtu.be>.
[24] Houdini "Ubisoft | Far Cry 5 | Houdini Connect" <www.youtube.com/watch?v=k8ChCR8vBGk&feature=youtu.be&t=93>.
[25] See: <https://www.rokoko.com/en/>.
[26] Rokoko "Online demo of Smartsuit Pro" (17 October 2017) <www.youtube.com/watch?v=Y_9TZHGswVA>.
[27] Gabe Cohn "AI Art at Christie's Sells for $432,500" (25 October 2018) The New York Times <https://www.nytimes.com/2018/10/25/arts/design/ai-art-sold-christies.html>.
[28] See: <www.synthesia.io/>.
[29] See, for example: <https://thispersondoesnotexist.com/> ; <http://www.whichfaceisreal.com/>.
[30] See: <https://www.soulmachines.com/>.
[31] Will Knight "Fake America Great Again" (17 August 2018) MIT Technology Review <https://www.technologyreview.com/s/611810/fake-america-great-again/>.
[32] Robert Chesney, Danielle Citron "Deep fakes: A Looming Crisis for National Security, Democracy and Privacy?" (21 February 2018) Lawfare <https://www.lawfareblog.com/deep-fakes-looming-crisis-national-security-democracy-and-privacy>.

news" issue. The Guardian pronounces that, "deep fakes is where truth goes to die,"[33] while countless other publishers and academics have echoed the sentiment that in the near-future, "seeing will no longer be believing."

45. But the specific effects of new synthetic media technologies must be isolated. Stripped of hyperbole, the technologies generally:

    a. make it easier to create audiovisual information;

    b. make audiovisual manipulation technologies more accessible; and

    c. make the audiovisual information produced more realistic, and thus more persuasive to an information consumer.

46. As Chesney and Citron note, "harmful lies are nothing new."[34] Neither is realistic but untrue audiovisual information, or information presented in a deceptive context. When it comes to dialogue around the risks of deepfakes and synthetic media, there is room for some realism. Moreover, such realism does not preclude a conclusion that these technologies pose some degree of threat to society, even if it is less than an existential one. These are not the first technologies of audiovisual manipulation: For example, Adobe Photoshop has been widely used since 1988, and yet society has not experienced a catastrophic crisis of trust in images. If anything, today's consumers are encouraged to be more critical of the texts they encounter, both audiovisual and written. Moreover, even without digital effects technologies, skilled creators have been able to generate deceptive video and audio through a range of analogue techniques: manipulating lighting, camera angle, framing, depth, as well as traditional editing, and even the simple use of make-up artistry. Thus the deceptive capacities of emerging audiovisual media must be understood in context: as a continuum of a long, historical tradition of creating illusory audiovisual material that tricks the perception of a consumer. Many of these more traditional techniques remain just as effective as their newer, synthetic counterparts: for example, recently three Franco-Israeli conmen fraudulently acquired over eight-million Euro by impersonating the French Foreign Minister, using make-up, and building a replica of his office.[35] No emerging synthetic media technology was necessary. Furthermore, when it comes to political misinformation and propaganda, some experts believe written methods remain cheap, effective, and hard to prove the falsity of.[36]

47. None of this is to deny the existence of a threat arising from new synthetic media, but rather to contextualise that threat. By appreciating the long history of low-tech methods of audiovisual persuasion, policymakers are better situated to examine what aspects of synthetic media are novel. As noted, synthetic media is easier and cheaper to make, more realistic, and able to represent things that have previously been very difficult to represent in a persuasive way. Underpinning this perception of an emerging threat is the ease with which this information can be disseminated to other consumers. The perceived "deepfake" problem is as much an issue of how information may now be disseminated as it is an issue of how information may now be created. Emerging audiovisual technologies coincide with a revolution in publishing and information exchange that has occurred through internet technologies. Without this, things like deepfake videos would largely remain siloed with their creators, little more than curiosities, or matters to be objected to on principle.

---

[33] Oscar Schwartz "You thought fake news was bad? Deep fakes are where the truth goes to die" The Guardian (12 November 2018) <https://www.theguardian.com/technology/2018/nov/12/deep-fakes-fake-news-truth>.

[34] Robert Chesney, Danielle Citron "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security" (2018) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3213954>.

[35] Kim Willsher, Oliver Holmes "Conmen Made €8M By Impersonating French Minister - Israeli Police" The Guardian </www.theguardian.com/world/2019/mar/28/conmen-made-8m-by-impersonating-french-minister-israeli-police>.

[36] Bobbie Johnson "Deepfakes are solvable - but don't forget that "shallowfakes" are already pervasive" 25 March 2019) MIT Technology Review <https://www.technologyreview.com/s/613172/deepfakes-shallowfakes-human-rights/>.

48. Synthetic media is placed in its proper context by Dr Turek, who explains to ABC News that:[37]

> "I think the challenge is that it is easier to create manipulated images and video and that can be done by an individual now," Dr. Turek explained. "Manipulations that may have required state-level resources or several people and a significant financial effort can now potentially be done at home."
>
> "It is significantly easier than it has been in the past -- for instance, to swap faces across video," he continued. "That was a CGI technique that special effects houses in Hollywood were able to do years ago, that now can be done potentially at home on a PC."
>
> The creation of a deepfake is somewhat similar to the state-of-the-art special effects used in today's filmmaking – like the face-mapping used to add the late Paul Walker's likeness to the film "Furious 7."
>
> But, deepfakes also have a lot in common with technology you're probably more familiar with: the photo album on your phone that learns your friends' faces, or "face swapping" on Snapchat.
>
> "If you compare a deepfake to what a person can do on Snapchat on their phone - deepfakes are much more believable," said Jeff Smith, the associated director of the National Center for Media Forensics at the University of Colorado Denver.
>
> "The challenging thing is that you have to have, at this point in time, pretty good skills with a computer," he added.

49. Much of the discussion of synthetic media's deceptive potential has so far occurred in the abstract. If this potential were reduced to its single most threatening element, it would be its capacity to create highly realistic representations of real people,[38] or representations that look as though they must be real people,[39] while obscuring or minimising the impact of digital manipulation technologies. These representations may be static images or animated videos, or even synthesised voices.[40] Moreover, a particularly uncomfortable element of these representations is that they are often generated by using actual audiovisual data about those individuals. Very often this data might have been collected or disseminated for other purposes, either by the subject themselves or some other person. At the time, that individual may have had no idea of the potential resource they were making available to create "fake" audiovisual representations of themselves. Consider the volume of audiovisual data collected, used and processed by social media platforms, whether publicly available or privately held, especially in the context of audio or video calling services. Whether photos, videos, or audio recordings of a person, these sorts of artefacts provide large quantities of digital data, and because digital data can be analysed and processed by computer systems, this allows for models to be trained from such artefacts so as to produce representations that look or sound like real people. It can be used to create both 'linear' artefacts, like a single video, or to create interactive digital assets that can be used over and over again, akin to having an avatar of a person that can be repeatedly re-animated. Some of these representations can even take place in real-time: using only a commercial webcam and appropriate software, a person can be captured and re-animated as entirely different person. The theoretical possibilities are significant, even if the practical realisation of them is limited in the immediate future.

50. The reason for public discussion about "deepfakes" is that they crystallise a harmful potential that is immediately recognisable. This is especially so because many consumers are unaware of the existence of such technology, or its capacity. Moreover, their emergence coincides with a massive increase in consumption of audiovisual information. Today, people consume more images, video, sound and music than ever before, both for entertainment and to be informed. This furthers the opportunity to deceive and to be deceived.

51. As well as consuming, people create and share more audiovisual information than ever before by means of the internet and social media. In their smartphones, people carry with them a tool for

[37] Above n 9.
[38] Above n 13, 18, 19, 21.
[39] Above n 29.
[40] Above n 22, 23.

capturing light and sound energy, manipulating this information, and disseminating it to the rest of the world. Moreover, the capacities of these technologies are rapidly improving. Common applications like Snapchat Filters, augmented reality, and so-called "face swapping" allow everyday people to create and share more and better manipulated audiovisual information: information that makes it look or sound like something happened when it did not. It illustrates the way that these technologies are likely to become commercialised and converted to consumer products requiring little expertise to operate and having largely innocuous impacts. Many of these applications share the same or similar techniques as deepfake videos. These sorts of technologies are becoming subtly but extensively pervasive: In a recent submission via the audiovisual link to Select Committee on the subject on the risks of new synthetic media for political deception, we noted that the conferencing software had an in-built "touch-up my face" function.

52. Therefore, while they are not inherently harmful technologies, the threat or risk of harm by their use is increased by several factors:

    a. There is a greater reliance on audiovisual information for a variety of purposes, including interpersonal communications, data entry, conveyance of meaning, and as a source of fact or truth by which decisions can be made.

    b. There is a common and increasingly incorrect assumption that some kinds of audiovisual information cannot be easily falsified, or that falsification is generally easily detected even by untrained humans, particularly in representations of the human face and voice.

    c. There is a likelihood that a proliferation of the tools to create such "false" audiovisual information will result in greater volumes of such information, which might undermine the reliability of all audiovisual information.

53. It is easy to understand why demos of the potency of new synthetic media are correlated with much wider concerns about phenomena like disinformation, misinformation, and "fake news". It is difficult enough to account for disagreements about opinion or macro-level trends that influence policy discussions, or to point to evidence about facts that is relatively undisputable. But to the extent that audiovisual information has been taken as factual evidence, deepfakes allow us to create wrong evidence about facts. They provide a visceral manifestation of the worst conceptualisations of a "post-fact" or "post-truth" world, and are receiving international attention from mainstream media. Ubiquitously, we are informed that now and in the future, "seeing is no longer believing", a refrain repeated by mainstream publishers including: The Hill;[41] the New Yorker;[42] Vox;[43] Quartz India;[44]

[41] Morgan Wright, "The age of deepfakes: When seeing is no longer necessarily believing" (23 January 2019) The Hill <https://thehill.com/opinion/technology/426536-the-age-of-deepfake-when-seeing-is-no-longer-necessarily-believing>.

[42] Joshua Rothman "In the Age of A.I., is Seeing Still Believing?" (5 November 2018) The New Yorker <https://www.newyorker.com/magazine/2018/11/12/in-the-age-of-ai-is-seeing-still-believing>.

[43] Brian Resnick "We're underestimating the mind-warping potential of fake video" (24 July 2018) Vox <https://www.vox.com/science-and-health/2018/4/20/17109764/deepfake-ai-false-memory-psychology-mandela-effect>.

[44] Aria Thaker "Should India worry about deepfakes affecting the upcoming election?" (26 March 2019) Quartz India <https://qz.com/india/1575860/could-deepfake-videos-spread-fake-news-in-2019-indian-election/>.

CIO;[45] Wired;[46] the Washington Post;[47] ABC News;[48] Radio NZ;[49] CNN and many others.[50] Others also link this to a wider "post-truth" situation being discussed in politics where matters of objective reality are now in dispute.[51] In this environment, .the concern for law and policymakers is not only to respond to a potential social threat, but to respond responsibly, proportionately, and appropriately.

54. Although the possible permutations of synthetic media misuse are endless, some cases that have already occurred include:

    a. Pornography depicting the faces of famous celebrities synthesised onto the bodies of adult performers in pre-existing videos.[52] These deepfake videos make use of the large quantity of publicly accessible digital data which depicts famous female actresses and singers. Originating approximately in the last two years and first distributed on social media, there are now websites dedicated entirely to this kind of pornography. Scarlett Johannson, frequently a subject for deepfakes, has tried and ultimately abandoned her attempts to remove such videos and prevent their further creation and distribution.[53]

    b. False social media accounts attempting to gain access to information or spread information while using photographs produced autonomously by generative adversarial networks (GANs) so as to appear more realistic. For instance, "Maisy Kinsley", a fabricated senior journalist supposedly working for Bloomberg.[54] Kinsley also had a fake LinkedIn profile (with 195 connections) and a personal website speculated to have been written by an algorithm. Kinsley's Twitter account had followed numerous Tesla short sellers, at least one of which reported that the account attempted to gain personal information from him. It is possible that the intention behind the account was to spread information that would affect the value of Tesla stocks.

    c. Amidst civil and political unrest, debate is ongoing about whether or not an official video depicting Gabon's President Ali Bongo was or was not a deepfake.[55] Political opposition claim the video is a falsification. Bongo had been away from Gabon for several months as he received medical treatment, fuelling suspicions that he was gravely ill or perhaps had already died. Ambivalent in their response, the Gabon government promised Bongo would appear for his customary New Year's address, but the video drew extreme scepticism as to its authenticity and its veridicality remains undetermined. The video was part of a number of factors that led to an attempted military coup thereafter. Digital forensic expert Professor Hany Farid of Dartmouth College, who is also working on the DARPA MediFor programme to combat digital

---

[45] J.M. Porcup "What are deepfakes? How and why they work" (1 August 2018) CIO New Zealand <https://www.cio.co.nz/article/644646/what-deepfakes-how-why-they-work/>.

[46] Yasmin Green "Fake video will soon be good enough to fool entire populations" (12 January 2019) Wired <https://www.wired.co.uk/article/deepfake-videos-security>.

[47] Monkey Cage "Fake news is about to get a lot worse" (3 April 2018) The Washington Post <https://www.washingtonpost.com/news/monkey-cage/wp/2018/04/03/fake-news-is-about-to-get-a-lot-worse-that-will-make-it-easier-to-violate-human-rights-and-get-away-with-it/>.

[48] Above n 9.

[49] Max Towle "Deepfakes: When seeing is no longer believing" (18 May 2018) Radio New Zealand <https://www.radionz.co.nz/news/the-wireless/375262/deepfakes-when-seeing-is-no-longer-believing>.

[50] Donie O'Sullivan, et al. "Pentagon's race against deepfakes" (2019) CNN <https://edition.cnn.com/interactive/2019/01/business/pentagons-race-against-deepfakes/>.

[51] Lee McIntyre "Lies, damn lies and post-truth" (19 November 2018) The Conversation <https://theconversation.com/lies-damn-lies-and-post-truth-106049>.

[52] Samantha Cole "AI-Assisted Fake Porn Is Here And We're All Fucked" (12 December 2017) Motherboard (VICE) <https://motherboard.vice.com/en_us/article/gydydm/gal-gadot-fake-ai-porn>.

[53] Drew Harwell "Fake-porn videos are being weaponized to harass and humiliate women: 'Everybody is a potential target'" (3 December 2018) The Washington Post <https://www.washingtonpost.com/technology/2018/12/30/fake-porn-videos-are-being-weaponized-harass-humiliate-women-everybody-is-potential-target>.

[54] See: <https://twitter.com/sokane1/status/1111023838467362816>.

[55] Ali Breland "The Bizarre and Terrifying Case of the "Deepfake" Video that Helped Bring and African Nation to the Brink" (15 March 2019) Mother Jones <www.motherjones.com/politics/2019/03/deepfake-gabon-ali-bongo/>.

disinformation,[56] is suspicious of the video but unable to determine its authenticity – in it Bongo has unusual speech patterns, barely blinks, and is otherwise uncannily static in his movements.[57] The key learning of the case is that the mere existence of these new technologies has created the potential for ambiguity and uncertainty, irrespective of whether the video actually is a fabrication.

55. We cannot only approach these technologies as if they are a threat, however. They have beneficial uses which may lead to other legal issues. New technologies allow for the replication of faces and voices, sometimes characterised as "digital likenesses". This is particularly relevant for persons who trade on their face, voice, or performance in some capacity, or the people who capture and distribute those performances in commercial markets. The advancing technologies of virtual human avatars have potential for use as substitutes (or supplements) to real human actors,[58] as does the use of deep learning neural networks to enhance the flexibility of post-production techniques.[59] When considering what can be done to avoid the threats of these technologies, we also need to consider their benefits, and how far the law may already have existing concepts that allow for protection against financial loss, for example copyright protections for many creative works. Where there is greater ambiguity surrounding law's capacity to protect is in relation to ordinary people who do not trade on their audiovisual profile, but who nonetheless may be synthetically replicated without their consent.

## Mitigating a threat

56. We believe that the potential benefits to be derived from new synthetic media are extensive and unquantifiable. Already we have encountered companies applying these technologies across industries for entertainment, education, customer assistance, healthcare, and mitigating the effects of physical or mental impairments. One example is "Project Revoice", which in conjunction with Lyrebird endeavours to re-create synthetic copies of voices lost by persons suffering from Motor Neurone Disease.[60] The same deep learning techniques by which illicit deepfakes are produced are also used in the creation of major motion pictures, particularly for representations of deceased actors, but increasingly for other creative purposes. It is important to recognise the creative and beneficial potential of new technologies which make is easier and more accessible to make it look or sound like something happened when it did not.

57. Nevertheless, a fair analysis of the impact of these technologies in the short term must acknowledge the risks of disruption or harm. We believe these fall broadly into four categories:

    a. Harm or loss to identifiable persons, including both natural and legal persons;

    b. Harm to less identifiable groups or communities, e.g. religious or ethnic groups which may be the subjects of manipulated audiovisual content that misrepresents them to the wider community;

    c. Disruption or harm at the civic or national-level; e.g. foreign interference in domestic elections by means of audiovisual misinformation; domestic misrepresentation of public figures or politicians; misinformation during civil emergency; or misrepresentation of police and emergency services;

[56] Ibid.
[57] Finances Africa "Le président gabonais Ali Bongo, mort ou vivant ? Vidéo deepfake?" (1 January 2019) <www.youtube.com/watch?v=62vkG7xfC18&feature=youtu.be&t=25>.
[58] See, for example: Idealog "Soul Machines 'Digital Humans'" (7 September 2017) <https://www.youtube.com/watch?v=rRsBMEwflz8>.
[59] Synthesia announced in May 2019 they have raised $3.1m in funding.
[60] See: <www.projectrevoice.org/>.

d. **Disruption of or harm to particular values**, e.g. loss of trust in all audiovisual information; loss of trust in the integrity of news and journalism; increased civic and political mistrust; or loss of trust in interpersonal communications.

58. The further these categories stretch from identifiable harms and victims, the more ambiguous they become. For instance, a loss of trust in audiovisual information can be seen to have both negative and positive impacts. Over-trust in unreliable information is at the heart of concerns about "fake news" and misinformation. At the same time, audiovisual information has also been a tool for positive legal or social change, particularly in recent history. Some organisations, like WITNESS, rely on some appropriate degree of trust in audiovisual records in order to combat human rights abuses.

59. While the scale of these threats can and should be critiqued, there are very few people who disagree with the claim that the proliferation of new synthetic media constitutes no threat at all. Such an argument must rest upon an ignorance towards the concerning examples that have already arisen, failure to account for historical tendencies to use audiovisual information for deceptive purposes, an overconfidence in technological methods to provide a suitable solution in the near-future, or an overconfidence in law to provide a meaningful deterrent effect.

60. The real question is what degree of intervention is necessary, while bearing in mind that the answer may be none, especially if policymakers acknowledge the risks of intervention itself, or are made aware of the extent to which law already intervenes. In overseas jurisdictions there clearly is a will to intervene and an ongoing dialogue over how that intervention should occur. Some prominent voices in the community (including DARPA researchers) see the threat as being sufficiently great as to justify some kind of targeted response. This intervention can be technological or legal or both. The call to action is amplified by the perception that the threat is unprecedented, that the kinds of harms are not recognised by existing law, and that the scale is sufficient that a response is required.

61. We have seen a range of responses suggested to the issue and we frame them as follows at a high level.

    a. **Technological**: "Fake" audiovisual information is framed as a virus for which we need a corresponding antivirus.[61] Researchers can attempt to upscale manual digital forensic techniques, automating where possible, to identify manipulated media at its point of inception, publication, or consumption. The goal may be to either provide guidance or warning signs about information as it is consumed, or to censor and prohibit that information from reaching the consumer.[62]

    b. **Legal:** "Fake" or materially manipulated audiovisual information is framed as a legal wrong. Lawmakers can attempt to draft statutes accurately enough to be an effective deterrent on whatever activity is deemed inappropriate, such as creating "fake" audiovisual information, consuming it, passing it onto others, or any other specific activity. Judges might also apply existing laws to extend their application in ways that make them relevant to new audiovisual technologies.

    c. **Commercial (Publishers and platforms):** Publishers, whether social media or mainstream media, can change the way they do business so as to limit the amount of "fake" audiovisual information that is reproduced through their platforms, or which is reported on incorrectly.

    d. **Commercial (Creators):** Creators of new information and researchers of new technologies can place ethical considerations at the forefront of their programs, taking into consideration the

[61] Jeremy Hsu "Can AI Detect Deepfakes to Help Ensure Integrity of U.S. 2020 Elections?" (28 February 2019) IEEE Spectrum <https://spectrum.ieee.org/tech-talk/robotics/artificial-intelligence/will-deepfakes-detection-be-ready-for-2020>. See also: Deeptrace Labs <www.deeptracelabs.com/>.

[62] Examples of companies working on varying private market solutions to deepfakes and misinformation include Deeptrace Labs (ibid) and Blackbird <https://www.blackbird.ai/>.

potential for unethical uses of their products,[63] and changing their business practices accordingly.

e. **Behavioural (Consumers and Citizens):** Individuals and communities can change the way they interact with audiovisual information by placing a greater emphasis on media literacy, scepticism and critical analysis, including emerging disciplines like open-source intelligence. Government and Non-Government Organisations can help to instigate these behavioural changes.

62. These responses are very general and the specific methods or policies by which they may be pursued will vary. However, even at the general conceptual level, each has particular difficulties, limitations, or concerns:

a. **Technological**: The nature of the virus-antivirus paradigm, or forger-detective paradigm, is such that new forensic methods are immediately reciprocated by newer, better "fakes". For example, some of the indicia which supposedly give away fakes are already becoming obsolete: lack of blinking or eye movement, surreal backgrounds, etc. Reliance on these indicia is already unjustifiable mere months after they were first identified. Furthermore, ongoing research programs have not yet succeeded in producing the kind of technological solutions that would 'solve' the audiovisual deception problem. Moreover, it is unclear whether they ever will: even within the largest publicly-funded programs, it is clear there will be no silver bullet approach. Others have said publicly that technology alone cannot solve the problem.[64]

b. **Legal:** It is not yet clear whether new laws for "fake" audiovisual information are necessary, desirable, or plausible. Statutory language is imprecise and may not be able to generate norms which respond to the "fake" audiovisual information phenomenon in a way that does not capture more audiovisual information than is justified. The risk of harmful unintended consequences is high. Moreover, there is reason to be sceptical of law that will impinge upon civil and human rights, and which is drafted in an environment of urgency so as to respond to the perceived threats of various new technological phenomena. Furthermore, legal mechanisms for redress are slow moving. They may struggle to identify culpable actors, or even identifiable victims, or to respond proportionately to nebulous kinds of harms. There is reason to be sceptical of the central claim of crisis requiring urgency, given how old the phenomenon of "harmful lies" is generally and the variety of laws we identify that already deal in deception (in New Zealand at least).

c. **Commercial (Publishers and platforms):** Besides the same technical difficulties faced by technological researchers, there are legitimate concerns as to the extent to which publishing platforms ought to be censoring information,[65] and in the case of social media platforms, whether it is sound to oblige them to take on the role of censor over the interactions between their users,[66] how their processes will be scrutinised for fairness and justice and even how any emergent rules and regulation would be enforced.[67] The same questions persist as to whether they can effectively distinguish between "good" and "bad" manipulated information in a way that is not gameable, or prone to partisan interpretations.

---

[63] In extreme cases creators may actually refrain from releasing products, as is alleged to have occurred with an AI-assisted text generator developed by OpenAI: Zack Whittaker " OpenAI built a text generator so good, it's considered too dangerous for release" (February 2019) TechCrunch <www.techcrunch.com/2019/02/17/openai-text-generator-dangerous/>.

[64] Including Hany Farid, above n 9.

[65] See, for example, Niam Yaraghi "Regulating free speech on social media is dangerous and futile" (21 September 2018) Brookings Institute <www.brookings.edu/blog/techtank/2018/09/21/regulating-free-speech-on-social-media-is-dangerous-and-futile/>.

[66] See, for example, David French "The Social Media Censorship Dumpster Fire" (1 March 2019) National Review <www.nationalreview.com/2019/03/the-social-media-censorship-dumpster-fire/>.

[67] See, for example, Bryce Edwards "Jacinda Ardern's 'Christchurch Call' might not be so simple" (29 April 2019) New Zealand Herald <https://www.nzherald.co.nz/nz/news/article.cfm?c_id=1&objectid=12226256>.

d. **Commercial (Creators):** There are limits to which creators, particularly those producing products and technologies, should be obliged to have concern for potential fringe risks arising from the tools they distribute. Just because it is feasible that a technology could be used for a harmful purpose by a determined bad actor, that does not automatically mean the technology is maleficent, or that we should prohibit it and thereby forgo the many valuable uses it may have. Many relatively benign or beneficial technologies can and have been misused. We do not ban knives or scalpels only because they are occasionally used as a weapon, and there are limits to which we require knife-makers to future-proof their tool against harmful use. We note, however, there is precedent for this kind of action in Photoshop and photocopiers when it comes to currency duplication.

e. **Behavioural (Consumers and Citizens):** Changes to consumer behaviour may be difficult to instigate. It could take a major geopolitical event before citizens alter the way they consume and share audiovisual information,[68] by which point significant harm might already have occurred. Furthermore, all behavioural changes will necessarily have trade-offs. For example, the line between healthy criticality and unhealthy distrust may be slim. Loss of trust in audiovisual information is considered to be one of the major threats of a proliferation of fake content. Encouraging criticality may accelerate loss of trust, rather than prevent it. Further, the "fake news" issue illustrates that relying on the critical analysis of everyday people may not be sufficient, particularly in the face of determined and sophisticated bad actors.

63. With regards to legal intervention, some overseas jurisdictions are already examining proposals for new law. Some of these are directed specifically towards the phenomenon of "fake" audiovisual artefacts. Others address harmful information generally, and therefore necessarily capture audiovisual information. Some of these include:

a. *Bill S.3805* introduced in the United States Senate by Sen. Ben Sasse (R-Neb). The "Malicious Deep Fake Prohibition Act of 2018" defines a deep fake as:[69] "an audiovisual record created or altered in a manner that the record would falsely appear to a reasonable observer to be an authentic record of the actual speech or conduct of an individual".[70]

b. *Bill A08155* introduced in the New York State Assembly as,[71] "An act to amend the civil rights law, in relation to the right or privacy and the right of publicity", which purports to establish the right of privacy and right of publicity for both living and deceased individuals, provided that an individual's persona is the personal property of the individual and is freely transferable and descendible.[72] If this element is satisfied, "the use of a digital replica for purposes of trade within an expressive work shall be a violation."[73]

c. *Bill 564* introduced in the California Senate by state Sen. Connie Leyva (20th District), with the goal of enhancing performers' protections when they are involved in sex scenes and their rights to control their likeness, including digital depictions.[74] The legislation would give Californians

[68] See, for example, Robert Chesney, Danielle Citron "Deepfakes and the New Disinformation War" (January 2019) Foreign Affairs <www.foreignaffairs.com/articles/world/2018-12-11/deepfakes-and-new-disinformation-war>.
[69] Malicious Deep Fake Prohibition Bill 2018 (S.3805) <www.govinfo.gov/content/pkg/BILLS-115s3805is/pdf/BILLS-115s3805is.pdf>.
[70] Ibid at § 104(a)(2).
[71] New York Assembly Bill A08155 2018. <nyassembly.gov/leg/?default_fld=&leg_video=&bn=A08155&term=2017&Summary=Y&Text=Y>.
[72] Ibid at § 50f(2).
[73] Ibid at § 51(3).
[74] Depiction of individual using digital or electronic technology: sexually explicit material Bill 2019 (Bill 564) <www.legiscan.com/CA/text/SB564/id/1926323/California-2019-SB564-Introduced.html>,

the right to sue creators of "deepfake" pornography or fake sex tapes, and is supported by the Screen Actors Guild - American Federation of Television and Radio Artists.[75]

d. The "Protection from Online Falsehoods and Manipulation Bill" introduced in Singapore, intended to combat perceived misinformation and disinformation online.[76] Publishers would be forced to display "correction notices" that direct readers to "correct facts" as claimed by the government.[77] The law would also grant government authorities power to issue "take-down" order that require the removal of content posted by social media companies, news organizations, or individuals[78]. The Bill is near to passing and likely to become law within one or two months.

e. Online Harms White Paper,[79] under consultation in the United Kingdom, which is to form the basis for a framework to sanitise the online or virtual world in the UK. Notable among its proposals is the establishment of a new statutory duty of care to make companies take more responsibility for the safety of their users and tackle harm caused by content or activity on their services, to be overseen and enforced by an independent regulator.[80]

f. Australia has drafted the Abhorrent Violent Material Bill.[81] According to Engadget, "Under the new law, content hosting platforms have to "expeditiously" pull down audio and video recordings or streams depicting "abhorrent violent conduct." In addition, they must notify authorities whenever they find illegal materials. If the companies fail to remove violent content in time, they could face fines up to 10 percent of their annual profit. The law could even, "slap individuals running hosting services with a $2.1 million fine or send them to prison for up to three years."[82]

64. Almost all of these proposed laws are the subjects of ongoing criticism,[83] the common thread of which is their reliance on overbroad, ambiguous language. This generates a risk that they may be used for undue censorship or limitation on rights of speech and expression. Frequently the persons most likely to be subject to the law emphasise the impossibility of compliance. There is also scepticism of the claims of crisis that supposedly justifies new legislative action.

65. Others, particularly those arising in the United States, have been criticised for overreacting, and for creating "unprecedented" new rights to control the use of so-called "digital replicas", while generally criticizing the claims of urgency which are stimulating new legislative response. Opposition has arisen from the creative industries, including organisations like the Motion Picture Association of America, Inc.,[84] Disney,[85] and NBCUniversal.[86]

---

[75] SAG-AFTRA "SAG-AFTRA Backs Legislation to End Nonconsensual Digital Sex Scenes and Nudity" (28 March 2019 <www.sagaftra.org/sag-aftra-backs-legislation-end-nonconsensual-digital-sex-scenes-and-nudity>.

[76] Protection from Online Falsehoods and Manipulation Bill 2019 (10/2019) <https://sso.agc.gov.sg/Bills-Supp/10-2019/Published/20190401?DocDate=20190401>.

[77] Ibid at s 11 "Correction Direction".

[78] Ibid at s 12 "Stop Communication Direction".

[79] <https://www.gov.uk/government/consultations/online-harms-white-paper>.

[80] Ibid p 42 at 3.1.

[81] Criminal Code Amendment (Sharing of Abhorrent Violent Material) Bill 2019.

[82] Maiella Moon "Australia's new law threatens social media companies with jail, fines" (4 April 2019) Engadget <www.engadget.com/2019/04/04/australia-laws-social-media-fines-jail/>.

[83] See, for example: Alex Hern "Internet crackdown raises fears for free speech in Britain" (8 April 2019) The Guardian <www.theguardian.com/technology/2019/apr/08/online-laws-threaten-freedom-of-speech-of-millions-of-britons> ; Human Rights Watch "Singapore: Reject Sweeping 'Fake News' Bill" (3 April 2019) <https://www.hrw.org/news/2019/04/03/singapore-reject-sweeping-fake-news-bill> ;

[84] MPAA "Memorandum in Opposition to New York Assembly Bill A.8155B (Morelle, Right of Publicity)" <www.rightofpublicityroadmap.com/sites/default/files/pdfs/mpaa_opposition_to_a8155b.pdf>.

[85] Letter from Lisa Pitney (Vice President of Government Relations at Walt Disney Corporation) to various Senators (of the New York State Assembly ) requesting their opposition to Bill A.8155B (8 June 2018) <www.rightofpublicityroadmap.com/sites/default/files/pdfs/disney_opposition_letters_a8155b.pdf>

[86] NBCUniversal "Memorandum in Opposition to New York Assembly Bill A08155B (Right of Publicity)" 8 June 2018 <https://www.rightofpublicityroadmap.com/sites/default/files/pdfs/nbc_opposition_a8155b.pdf>.

66. There is also good reason to be skeptical of the idea that new law to curtail 'wrong' information is urgently necessary, given how long the debate about what constitutes 'wrong' information has been occurring and the constant flux of these concepts. Law and policymakers therefore must be cautious of claims of novel threat justifying novel intervention, which very often rides on the coattails of new technological developments, used as a justification for the abrogation of various rights and privileges that would otherwise be difficult to roll-back. In many ways, things like deepfakes simply oblige the reconsideration and renewal of much older debate on the proper role of information, stories, and illusion in a civil society – a discussion dating back at least as far as Plato.[87] The discussion also overlooks the potential that large volumes of existing law already govern both the content of audiovisual media, the distribution of it, and misrepresentation.

## The challenges for legal intervention

67. The focus of this Report is the proper role of law in responding to new synthetic media technologies. This requires an understanding of:

    a. the present and future capabilities of the technologies;

    b. their potential for misuse;

    c. the ways they are consumed and also perceived;

    d. the kinds and scale of harms that might arise from them;

    e. the readiness of existing law; and

    f. the risks of overbearing legal interventions.

68. The conclusions for several of these points may be inferred directly from an understanding of the technologies, their developmental trajectory, and their present and future potential. For example, based on our consultations within both the creative and detective cohorts of the industry, we believe that in the near-future it will be possible for everyday people to create synthetic media artefacts that are highly persuasive, being realistic, and otherwise depicting representations of things that are traditionally very difficult to "fake". We anticipate production of the sorts of video and audio representations that currently can be produced only by highly-resourced special effects studios and expert training, although quality in each case will vary. By extension, in the near-future, the sort of artefacts that may be produced by highly resourced actors - like large corporations or nation states - will be of unprecedented realism. Many of these artefacts will be able to be produced at a rapid scale, often with limited input or oversight from human agents.

69. With this in mind, the potential for harmful misuse is significant. Suffice to say that highly realistic video and audio is very persuasive, and that persuasive audiovisual content may be used to deceive, mislead, or misinform a naive consumer to a great extent. Over time, consumers may become guarded to the possibility that a realistic video or audio clip does not truly evidence the information that it seems to represent. In the meantime, wrongly trusting this information may have harmful consequences, with the level of harm varying depending on any number of factors. Moreover, harms might be of a sort that are difficult for existing law to deal with: there may not be a single obvious victim, or a readily identifiable culpable actor. Alternatively, the harms might be as novel as the technologies by which they are wrought, in which case, speculating about them is hazardous at best.

70. There are significant challenges to using law as a tool for intervention. They are both conceptual and pragmatic. Understanding and acknowledging these challenges must underpin any

[87] Consider the attack on poets and poetry, Plato *The Republic* (2nd ed, Penguin Group, London, 2007) 53 to 76 and 335 to 349.

forthcoming policy response. What then are the challenges for effective legal intervention? We identify a number of preconditions which must be accounted for:

a. **Audiovisual information is expression** and thus protected by rights of free expression, subject to normal limitations. These rights and concomitant limitations will vary depending on the legal jurisdiction. In New Zealand, free expression is protected by the New Zealand Bill of Rights Act 1990 ("**NZBORA**"). In spite of any of the limitations placed upon it, the right to freedom of expression holds significant weight. As such, there must be good reasons that are demonstrably justifiable in a free and democratic society for repressing any given expression or means of expression, including the audiovisual.[88]

b. **Limitations on expression do not apply to "falsity" in general**. In other words, lying or deceptive expressions are not inherently illegal, and thus not (or ought not to be) generally illegal when occurring by means of audiovisual artefacts. Of course, falsity or deception may be illegal when arising in particular contexts or particular relationships, or where leading to particular kinds of hardship or loss. But falsity in itself is not generally illegal, whether occurring in audiovisual information or otherwise. It would therefore be a significant departure from existing norms to make false audiovisual information generally illegal.

c. **"Fake" or "manipulated" audiovisual information is not inherently harmful**. In fact, this kind of information frequently is benign, even beneficial. Harmful uses generally are an exception and are peripheral in comparison to the quantity of video and audio that is readily sought out by citizens and consumers. Very often, audiovisual information is consumed specifically because it is a fabrication - both realistic but non-veridical - and because it has been manipulated.

d. **The quality of "fake" or "false" is predominantly contextual,** by which is meant that an audiovisual artefact cannot be "fake" simply because it is non-veridical. Many audiovisual representations will be sufficiently realistic so as to make it look like something happened when it did not. But much less of this information will be "false", to the extent that a sensible definition for falseness predominantly invokes contextual elements which go beyond the content of the text itself. A definition of false or fake which is entirely reliant on an audiovisual artefact itself would inevitably capture huge swathes of benign audiovisual information.

e. **"Manipulation" is pervasive in digital audiovisual technologies**, and therefore is not a load-bearing concept for law or policy in a binary sense. Restrictive norms applied generally to manipulated audiovisual information, or to the processes by which this is created, will include almost all of such information. This is particularly true of digital artefacts, which are the products of ineluctable processes of manipulating, changing, supplementing, and removing data. Of course, manipulation may come by degrees, and some artefacts will be 'more manipulated' than others, or manipulated in ways more material to the case at hand than others. But determining the materiality of these manipulations diverts to assessment of contextual factors, away from the text itself or the means and methods by which it was made. We submit that very quickly, narrow focus on "manipulation" absent context loses all utility.

f. **"Falsity" or "fakeness" is pervasive in all media,** and therefore is not a load-bearing concept for law or policy. For one, all audiovisual media is untrue to the extent that it is only an approximation of whatever light or sound energy was first "captured" in audiovisual form, if ever captured at all. Moreover, media of all kinds necessarily omits certain information, or directs the consumer's attention in a way that rarely represents the "true" scene as it unfolded. Similarly all mediums by which information is presented have their own effect on the way the information is perceived and understood by the consumer - the medium itself being a part of the message. Many of these decisions are made consciously by the creator or publisher so as to influence the consumer according to their needs. As above, we accept that there are 'degrees of fakeness' that are greater or lesser than others, but attempting to capture this nuance within

[88] New Zealand Bill of Rights Act 1990, s 5.

imprecise conceptual terms is very difficult, and perhaps impossible, especially in the abstract in advance.

71. At its simplest, 'manipulation' can be largely understood as an objective matter. Whether or not audiovisual information is or is not manipulated is a matter of fact, although given the pervasiveness of "manipulation" in the creative process, it is of little value. Conversely, 'falsity', or whether it misrepresents the subject matter it appears to have captured, can be understood largely as a subjective matter, accounting for factors like the relationship between one or more persons in an information exchange, the claims implicit in that information exchange, the medium in which information is presented and perceived, and so on. All these factors may influence the subjective belief of a person perceiving an audiovisual artefact, and indeed the subjective mindset of the person who is presenting the information and potentially being false. While it may be somewhat easier to assess for the presence or absence of manipulation in an audiovisual artefact at the outset, assessing for falsity or misrepresentation is much more difficult, and for the most part, only possible via a retrospective analysis of the contextual relationship between two or more people involved in an information exchange that has already occurred. Perhaps the chief risk for the design of law or policy is conflating the former with the latter.

## Issues arising regardless of specific legal regime

72. There are also a number of pragmatic difficulties inherent in any feasible legal intervention. Several of these are not necessarily limited to the problem of audiovisual technologies, but are general problems familiar to the application of law as a solution to a social or economic harm, especially wherever digital and internet-related phenomena are at issue:

    a. Difficulty in identifying responsible actors in relation to each action, and separation of actionable harms;

    b. Limited existence of and access to evidential, detection and verification services, whether human or automated, to justify claims of manipulation or falsity;

    c. Difficulties in access to justice given the potentially high volumes of synthetic media products causing harms, and the potentially non-financial harms (or low-level harms) arising from their use;

    d. Cross-jurisdictional issues arising from a globalised communications environment, including the increased prominence of media platforms' terms and conditions of use as an often more effective remedy than attempting to apply domestic legal regimes.

73. We characterise these issues as pragmatic because they are problems predominantly related to the enforcement of law and service delivery of the legal system, which are already particularly challenging where digital and internet-based media is concerned. These problems arise separately from problems related to the existing law itself, or any flaws or gaps therein. What makes these pragmatic problems pernicious is that they are unlikely to be resolved by the introduction of any new substantive law. In fact, new substantive law may exacerbate them by placing further burdens on service delivery and creating barriers for new competitors to enter markets. At the same time, new law may suffer from similar limitations in terms of access to the necessary expert evidential services that will be integral to the enforcement of existing law relating to realistic but non-veridical synthetic media.

## Conclusion to Part 1

74. New synthetic audiovisual media technologies are already rapidly proliferating overseas. Their eventual arrival in New Zealand is likely, if not inevitable, and some of them even originated here.

This has implications for creators, consumers, and citizens. Based on our investigation of the technologies and consultation within industry, we believe that almost any audiovisual representation will be possible in the near-future, even those which have typically been very difficult to create, or prohibitively expensive. It seems likely that within five years, these kinds of realistic synthetic media artefacts with no relationship to reality will be the subjects of regular consumption by most internet users. Based on the definition in our framework, we consume these already all the time.

75. There are enormous benefits to be gained from these technologies, but achieving these benefits requires the development of a sound and load-bearing policy foundation which can facilitate confident innovation. In short, creators need to know that the technologies and informational artefacts they invest huge resources into are legal and likely to remain that way. Where they are illegal or likely to become illegal, creators must be aware of this. Where particular uses of synthetic media are to be regulated or prohibited, this needs to be established in advance and incorporating wide consultation.

76. Preemptive preparation is the best method to guard against unsound reactionary policy with regards to synthetic media. There is potential for overreaction due to the affecting nature of state-of-the-art audiovisual representations: things which can present real people doing and saying things they never did, or which can represent events that did not take place. This overreaction may be dangerous if it drives policy towards indiscriminate or irrational restrictions on audiovisual expression.

77. Subsequently, there is a need for sound law and policy in preparation for this phenomenon. We believe this begins with **effective analysis of new technologies** as they arise, and **close attention to existing law** so as to identify its readiness, any deliberate gaps, and any potential flaws.

78. For this reason, we propose a **framework approach** to the investigation and analysis of synthetic audiovisual artefacts. In particular we note that, despite their ostensible novelty and uniqueness, all such artefacts share **common features, traits, and uses**. Because of these commonalities, they can be readily identified, and more easily regulated to the proper degree.

79. Part 2 of our report articulates that definitional framework. Part 3 applies it to New Zealand law to illustrate how many of the conceivable harmful implications of synthetic media technologies are already subject to legal oversight, or established legal norms. We call for close attention to the way that identifiable synthetic media technologies already sit within this landscape before entertaining any calls for regulation.

80. Despite the rhetorical fears of so many across mainstream media and academia, seeing is still believing. We are not close to "the collapse of reality," despite what has been written by Franklin Foer at the Atlantic and others (Foer's article has been run alternatively with the headlines "The Era of Fake Video Begins" and "Reality's End").[89] We appreciate that these titles are editorialised. Nevertheless, they conflate: first, the novelty of fakeness in video; and second, perception of audiovisual artefacts with sense and perception generally. It is not our capacity to trust in our senses that is compromised, but our capacity to trust in the reliability of audiovisual artefacts. This may stimulate a higher level of criticality when consuming audiovisual information, and subsequently the ways that we rely on it. This is not necessarily bad, but of course results in trade-offs. Audiovisual information has been a powerful vessel for conveying complex truths. To evidence this claim, look to the examples of recent history: audio recordings of the Nixon administration, photographic record of the War in Vietnam, CCTV footage of the police abuse of Rodney King, the livestreaming of police shooting unarmed civilians in America and countless international examples where human rights abuses have been documented. But the opposite has also been true, whereby people have been misled by their trust in deliberately misleading audiovisual information. Ultimately consumers and

[89] Franklin Foer "The Era of Fake Video Begins" (May 2018) The Atlantic <www.theatlantic.com/magazine/archive/2018/05/realitys-end/556877/>.

citizens must adapt - whether self-directed or through education or both - to understand that just because a video or audio clip sounds as if it must have captured actual light or sound energy of a real event, that does not mean this ever occurred. This may be an easy transition: most modern people understand the limits to which what they see and hear through video and audio recording is only a partial representation of reality. But we have in-built biological trust in the data derived by our eyes and ears, and on top of this have built both interpersonal and institutional reliance on audiovisual artefacts. Emerging audiovisual technologies do create change and therefore risk, as well as positive potential.

## Summary of Part 1

81. In summary:

    a. We already consume synthetic media artefacts in high volumes. These volumes will increase. The kinds of information being produced may lead to new challenges.

    b. Synthetic media artefacts and technologies create opportunities, benefits and risks. They are not inherently bad. They do have positive uses.

    c. Many of the concerns about synthetic media are as much about dissemination of misleading media than the creation of factually false media.

    d. All digital media entails a degree of synthesis and manipulation. If “fake” or “false” is understood as “manipulated”, then all synthetic media is manipulated and fake. This is not a sustainable basis for public policy.

    e. Whether the content of a synthetic media artefact is misleading or unreliable is a more complex question than simply whether it has been manipulated.

    f. The threat of deception posed by synthetic media technologies is not a new one, simply a new opportunity to cause harms of a very old kind.

# Part 2: A framework for synthetic media and the law

## Why develop a framework?

82. Many emerging audiovisual technologies are sensational artefacts. They appear new and difficult to critically assess. Arthur C Clarke wrote that, "Any sufficiently advanced technology is indistinguishable from magic."[90] Frequently, one's first impulse is to trust this impression, treating new technologies as if they are beyond the parameters of all existing rules. Audiovisual technologies generate illusions that, to the naive consumer, are magical.

83. In the course of our research, almost on a daily basis, we have identified new technologies that broadly 'make it look like something happened when it didn't happen' using visual or auditory outputs. We have included a range of examples throughout this report.[91] In each case we have asked ourselves whether the technology sits within our research scope. That has proved to be a useful exercise because it reflects the basic task at the heart of any investigation into hypothetical uses of that technology.

84. Subsequent analysis of those technologies - both social and legal - risks proceeding on an ad hoc basis with little consistent structure. There is a need to distinguish between the specific harms being contemplated: harms of capture, creation, content or dissemination. Each time a new technology arises, so does the need for analysis. This is particularly true where public perception becomes hostile towards a particular example of a general kind of technology, often fueled by catastrophic events.

85. This phenomenon is also stimulated by marketing efforts of the creators themselves, who seek to differentiate their product from the market and embellish its novelty. Even in academic circles, there is a need to compare and contrast the developments made from previous technologies, illustrating how a problem has been solved or improvement made. Thus the impression builds that each new successive technological artefact possesses unencountered characteristics, poses never before seen challenges, and requires entirely new ways of thinking, as well as new legal and social norms. The ostensible novelty of an artefact may distract from the fact that it contains many of the same elements as other audiovisual artefacts. These elements are recognisable and familiar to law. Far from being outstripped or outmoded, these elements of emerging technologies are already the subjects of extensive legal attention.

86. We believe there are essential commonalities between most or all synthetic audiovisual technologies and the information they produce, even where they may seem quite different. With these commonalities comes predictability and consistency, and the ability to see new technologies not as entirely new artefacts that outstrip existing law, but as things comprised of the sorts of elements with which law is already deeply concerned. Further, when it comes to audiovisual technologies, the legal standards we apply to that information already incorporate a degree of flexibility and media-neutrality to account for technological development and the complex social balancing exercises involved in restricting its use: we illustrate this in the next chapter of our report.

87. New audiovisual technologies are always arising and advancing, therefore the responsive task of lawyers and policymakers is unending. By approaching this task analytically, law can remain focussed on the essential elements of new technologies as they arise, rather than becoming distracted or overwhelmed by the impressive illusory effects of those technologies.

88. We considered whether it would be preferable to conduct our analysis by reference to a series of case studies or thought experiments. In these we would generate a hypothetical fact pattern about the deployment of a mixture of synthetic media technologies in a given social context that calls attention to the kinds of harms that might be caused.

---

[90] Arthur C Clarke "Hazards of Prophecy: The Failure of Imagination" in *Profiles of the Future: An Enquiry into the Limits of the Possible* (1973) 14 to 36.

[91] See, for example, above n 35.

89. We have elected not to rely on hypothetical fact patterns and instead developed a framework approach that can be applied to new situations as they arise. The framework approach accounts for the difficulties involved in generating useful hypothetical cases, which have limited use because of the following factors.

a. For any situation we construct that uses synthetic media technologies to achieve nefarious ends, there is often an equally plausible way to achieve those ends without the use of synthetic media, making the hypothetical case study somewhat redundant. For instance, the same or greater deception achieved by 'low-tech', analogue means.

b. There is a wide range of technologies that perform synthetic media functions. Further, these technologies are used together in interchangeable ways to generate unanticipated results. The range of potential social, legal and technological variables is innumerable, and thus 'solving' a given fact pattern does not necessarily assist in better decision-making beyond where future fact patterns happen to involve the same set of variables.

c. New technologies are emerging constantly, which means that any exhaustive attempt to 'solve' for today's particular set of technologies is ultimately not exhaustive, and of limited future value.

d. Effectively any digital media that appears to capture and broadcast light or sound energy can be described as synthetic media.

e. The sheer range of conceivable uses of the technology is vast and subject to unanticipated technological and social disruption. There is speculation that augmented reality, mixed reality, and spatial computing, for example, will be the future of smartphone technology. Building hypothetical cases for us to knock down based on the particular artefacts of today has limited future value.

f. There are a wide variety of legal regimes that deal in audiovisual media specifically, or generally with regards to deception in commercial, domestic, or criminal contexts. Emphasising the role of synthetic media technologies in such examples is unnecessary for the overall task of demonstrating that they might apply in appropriate fact patterns.

g. Assessment of each hypothetical fact pattern would be both ad hoc and unnecessarily detailed, with as much variety in criteria and terminology as there is in potential future researchers or fact-finders. Producing a consistent framework from which to work will help future work proceed on common, consistent ground.

90. For these reasons, in our own work, we have found it much easier to develop and implement the framework we articulate here.

91. The framework generates a vocabulary for answering the following questions:

a. Is this technology a synthetic media technology?

b. To what extent does this technology differ from existing technologies?

c. How far is this technology already subject to legal regimes that facilitate or restrict its use in particular contexts?

d. How can it its use be detected and any legal regime enforced?

e. What are the legal or social impacts it may generate and how far do existing legal mechanisms anticipate these potential impacts?

## Explaining the framework

92. We think that the pace of research and development of synthetic media technologies makes any static list of them obsolete almost as soon as it is written. A framework that defines synthetic media technologies and can be used to guide policy and legal responses to any perceived issues is more durable. The framework articulates a broad definition that treats synthetic media as a family of technologies that perform certain functions. They share the common capacity to manipulate human experience of audio-visual perception - they can "make it look like something happened when it didn't", both for positive and negative purposes.

### Summary of Framework elements

93. Synthetic media is a 'family of technologies' united by a series of common traits.

94. The framework emphasises digital storage and processing, and therefore excludes more traditional technologies such as eyeglasses or oil painting (for example), which can also arguably manipulate perception.

95. Identifying common features of synthetic media technologies allows for a consistent and principled definitional exercise to take place. It allows us to identify how far the features of new synthetic media technologies are already covered by existing legal regimes. In turn, this will identify entry points for any potentially harmful actions that are not already covered. The purpose is to break down synthetic media into recognisable functions, components and actions that can be analysed through a consistent and logical set of parameters. This enables us to consider how law and society could or should respond to their potentially harmful use by comparing them with other technologies and preserving their capacity for social benefit when used in a desirable fashion.

96. The framework describes three categories of technologies that capture, manipulate and display audiovisual information. It also has three "conditions", which explain how humans interact with the artefacts produced by synthetic media technologies.

    a. There are three categories of technologies (Figure 1). One or more of these is always present in any given synthetic media example. The categories are not mutually exclusive. A technology that meets the description in any of 1-3 can be described as a "synthetic media technology" ("SMT"). When each of 1-3 (though not always 1) are used to create a category 3 output, we describe that category 3 output as being a "synthetic media artefact" ("SMA").

        i. Category 1: "Capture" technologies capture light or sound energy and convert it to digital data.

        ii. Category 2: "Manipulation" technologies change digital data that is either: captured by category 1 technologies; or capable of being displayed by category 3 technologies.

        iii. Category 3: "Display" technologies include digital data that can be displayed for human consumption as light or sound energy, and the technologies required to facilitate that display.

    b. There are three conditions, that explain the interaction of the three categories of technologies that tend to be seen in any given synthetic media example:

        i. Condition 1: "Veridicality": the extent to which a category 3 product appears to be a reliable representation of something that happened in the real world.

        ii. Condition 2: "Multiplicity": the possibility that multiple actors have been involved in the various steps of creation and dissemination of synthetic media artefacts.

        iii. Condition 3: "Dissemination": referring to the way that the harms from disseminating an SMA are very different from an SMA that is never disseminated.

c. There are three main **kinds of harms** that result from SMA and SMT. These are caught by very different kinds of legal mechanisms.

i. Creation: harms caused by the process of creating synthetic media or using synthetic media technologies.

ii. Content: harms caused by the content of SMA.

iii. Dissemination: harms caused by disseminating an SMA, to the extent that these are different from the harms arising from an SMA that is never disseminated.

97. The framework lets us consider, for example, why and how deepfake videos are different from digital recordings made with a camera. One way is that an individual creating a deepfake deals in digital files only. They do not require the use of category 1 capture technologies, or at the least, they can make use of pre-existing digital data captured in another time and place, by other people and for alternative purposes. This guides us towards more relevant legal regimes, away from those more concerned with "capture", and perhaps towards those more concerned with data protection and data manipulation, or the harms caused by audiovisual information as it is displayed, rather than captured. Further, the creation of a deepfake for research purposes with consent of the subject may be completely without harm: the dissemination of that deepfake without context by another actor may be especially harmful.

## Harms

98. In most cases, the first question to ask should be "what is the harm that I am seeking to avoid or remedy?"

99. We think there are three broad groups of harms that arise with synthetic media technologies.

a. Harms can arise from the creation or capture process and the use of SMT.[92]

b. Harms can arise from the content of the SMA.[93]

c. Harms can arise from the dissemination of the SMA.[94]

## The Categories

100. The three categories can be used to answer whether a particular kind of technology is a synthetic media technology, and whether the product can be described as a synthetic media artefact.

### Category 1: "Capture" technologies

101. Category 1 technologies detect light and sound energy and convert this to electrical energy that is recorded as digital data. For example, a digital camera or a microphone with a digital converter.

102. As an example, smartphones contain multiple sensors that detect and capture light or sound energy: devices such as the iPhone X and Huawei P20 have multiple rear-facing cameras. One model of

[92] Examples of legal regimes that target harms from the creation process include: making an intimate visual recording as defined by s 216G of the Crimes Act; the tort in *Holland*; collection by unfair means in breach of the Privacy Act; the use of deceptive filming methods covered by the Broadcasting Guidelines by the Broadcasting Standards Authority; breach of copyright by incorporation or modification of copyrighted works.

[93] Examples of legal regimes targeting harms from the content of an SMA include: objectionable material under the Films, Videos and Publications Classification Act; misrepresentation in trade under the Fair Trading Act; altering a document contrary to the Crimes Act; the offence of perjury and manipulating evidence to mislead a tribunal.

[94] Examples of legal regimes targeting harms from the dissemination of an SMA include: defamation; posting a harmful digital communication; publication of private facts per the tort in *Hosking*; distribution of Intimate Visual Recording contrary to the Crimes Act.

smartphone by LG has up to 5 cameras enabling various photographic functions. Smartphones also commonly contain at least two microphones, which act to differentiate between ambient background noise and human speech.[95]

103. A more complex example of technologies of this kind is LIDAR, which "is a surveying method that measures distance to a target by illuminating the target with pulsed laser light and measuring the reflected pulses with a sensor. Differences in laser return times and wavelengths can then be used to make digital 3D representations of the target."[96] LIDAR is used by companies such as Staples VR to create 3D digital assets for use in virtual and augmented reality applications.

104. Augmented reality headsets also conduct a scanning exercise of the external environment using light energy in order to display computer generated visual effects over that environment, and therefore have category 1 features.

105. We note that the definitions of "photograph" and "sound recording" in the copyright Act both refer to light and sound respectively, and so a definition of this kind is not unusual in this area of the law.

### Category 2: "Manipulation" technologies

106. Category 2 "manipulation" technologies alter or manipulate digital data, rather than detecting or "capturing" light or sound energy in the environment. Our use of the term "manipulation" is entirely practical, in the sense of manipulating digital data. Frequently, manipulations to digital data are fundamental to making it fit for human consumption.

107. Many Category 1 technologies incorporate Category 2 as a matter of course, so that manipulation of the digital data becomes an inherent part of the "capture" process. Alternatively, some Category 2 technologies operate independently as stand-alone products: e.g. modern video and image editing software.

108. Software for manipulating digital data does not capture or detect light or sound energy, but rather makes changes to digital data files.

109. Category 2 anticipates software technologies that enable human actors to process digital data that has been generated by category 1 technologies from light and sound transduction. They also allow someone to generate digital data that makes it look as if a capture technology has been used (for example, animation).

110. The classic example of a Category 2 technology is Adobe Photoshop. While the end-user of the software operates a graphical interface to alter an image in most cases, the reality is that edits are being made to the digital data comprising an image file. There is a wide range of software that is capable of making augmentations to image files. For example, Instagram (and even simple PDF file viewers) enable users to operate sliders that brighten, darken, or enhance static images.

111. An example of category 2 technologies oriented toward sound energy is the rise of digital audio workstations (DAWs) such as Apple's GarageBand or Ableton's Live suite that enable anyone with a computer to simulate the effects of analogue audio engineering studios. These will include both sample banks and audio effects units that either reproduce or mimic real-world instruments and effects units, for example the drum kit used in the Beatles' Abbey Road recordings, an 808 drum machine used by J Dilla, or a particular kind of guitar pedal.

112. In the process of capturing light or sound energy using category 1 technologies, inevitably a degree of information loss occurs and a degree of information or data creation can also occur. For this reason, t can be difficult in practice to distinguish between category 1 (energy capture into data)

[95] Thomas Thorn "Background noise reduction: one of your smartphone's greatest tools" <www.techradar.com/au/news/phone-and-communications/mobile-phones/background-noise-reduction-one-of-your-smartphone-s-greatest-tools-1229667>.
[96] Wikipedia "LIDAR" <www.en.wikipedia.org/wiki/Lidar> accessed 11.50am 17 March 2019 NZT.

and category 2 (data manipulation) technologies. We do not think this is fatal to our framework. Many, if not most, will contain technologies captured by more than one category of the definition. The categories do not need to be mutually exclusive in order to generate useful insights about the multiple functions being performed by individual devices or products. Each case should be considered as a question of fact in the circumstances.

113. The boundary between categories 1 and 2 does serve an important purpose. Category 2 technologies are important for taking account of synthetic media technologies which do not capture any light or sound, but lead to products that give the impression that such a capture has taken place. A key example is the use of animation techniques whereby a model is animated in ways that are highly photorealistic, for example use of animation techniques in video games as applied to gaming engines, as well as procedural generation of game-world features.

114. Many virtual reality products incorporate "virtual assets" which have been purchased in online marketplaces, where those assets may have been captured through a category 1 process by an entirely separate agent. Accordingly, someone may produce a piece of synthetic media purely through the use of category 2 technologies even though a separate agent originally deployed a category 1 technology (further engaging condition 2 below).

115. Technologies such as deepfakes create the impression that a category 1 technology has been recently deployed, when it may be more accurate to describe the audiovisual product as being a result of technologies better situated within category 2.

116. We see many synthetic media technologies as automating and democratising category 2 technologies.

117. Two examples of legal provisions that regulate the use of category 2 technologies are information privacy principle 5 of the Privacy Act 1993 and s 258 of the Crimes Act 1961. We detail these in Part 3 of our report.

### Category 3: "Display" technologies

118. Category 3 "display" technologies convert digital datafiles into light or sound energy, producing images and audio information for human consumption, the digital data having been produced by Category 1 or 2 technologies. The exact manner or form of display may vary, but inevitably involve an output of light or sound energy: such as videos displayed on a monitor, audio through a speaker, or even a hologram. The key output from a category 3 technology is light or sound energy for human consumption.

119. Category 3 also includes the digital datafiles with potential to display images or sounds. In this respect, the boundaries of our category 3 definition mimics the idea of reproducibility found in the definition of a "document" in the Privacy Act 1993 and the Crimes Act 1961 (described in part 3 of our report). The inclusion of digital files within the definition of category 3 is also consistent with the incorporation of reproducibility in definitions of "photograph", "film" and "sound recording" in the Copyright Act.

### Summary of the categories

120. One or more of these three categories of technologies will be present in any given synthetic media technology, and any synthetic media artefact will be the product of at least one (and usually more than one) category. Many modern devices contain all three. The three categories can be used to answer the question, "Is this an example of a synthetic media technology?"

121. "Capture" or "display" may occur in different ways from technology to technology. Because of this, it may not be immediately apparent that the same essential phenomena are taking place. Nevertheless, even seemingly dissimilar audiovisual technologies are often fundamentally alike. The

categories therefore also enable us to answer the questions, "how far is this a new technology?" and "how far is this technology or device already regulated by the law?"

## The Conditions

122. The framework is incomplete without reference to three conditions that explain the way in which the three categories of synthetic media interact in ways that generate differing kinds of harms.

### Condition 1: The Appearance of "Veridicality"

123. Condition 1 takes Category 3 synthetic media artefacts as its starting point. When people consume an SMA it creates the impression that it has been produced by the use of a Category 1 capture technology. When we see a realistic "photo", we tend to assume it has been taken by a camera, though it may in fact be entirely synthesised by Category 2 technologies.

124. The capture process is never "perfect". Some light or sound information is always lost, created, or amended. There is a "camera pipeline" of sensors and processors that can be analysed forensically. Notably, a photograph is a two-dimensional representation of three-dimensional space. A sound recording inevitably adopts a degree of tone from the kind of microphone used.

125. This information loss during the capture process does not mean that all Category 3 products are useless: we rely on synthetic audio-visual information (category 3 products) produced by capture processes (Category 1 products) all the time. Instead, what this means is that the quality of a synthetic media product is judged in practice by whether it is fit for a particular purpose. What makes a synthetic media product fit for a particular purpose?

126. In our view, a Category 3 product is fit for purpose when there is a sufficiently reliable correlation (or relationship) between the light or sound energy captured (by a Category 1 process) and the light or sound energy broadcast for human consumption (through Category 3 technologies). The question of whether the relationship is close enough to make it sufficiently reliable depends on the context in which the synthetic media artefact is being consumed, and for what purpose.

127. Condition 1 takes this analysis into account from the perspective of someone consuming a category 3 product. It states that, when consuming category 3 products:

    a. there may be an explicit or implicit communication to the consumer that the Category 3 product is the result of a category 1 (capture) process.

    b. there may be an explicit or implicit communication that there is a reliable relationship between the light or sound energy captured by Category 1 technologies and the light or sound energy produced by Category 3 technologies.

    c. That the role of Category 2 technologies can be obscured, overlooked or misunderstood by the end consumer of the synthetic media product.

128. Condition 1 is fundamental for assessing the harms that may be caused by using synthetic media products in a deceptive way. Without any further context, if I see a photorealistic image of a human face, there is a risk I will assume that it was taken by a camera. I will overlook the fact that it is not perfect on the basis that it is recognisable or good enough for a particular purpose. I will overlook the role of the sensors in that camera in "creating" the image file and the role of data processing technologies in rendering the information captured into a useful form. I may also overlook the way that my particular smartphone has a different kind of screen than another smartphone that makes the image look slightly different than it would through another category 3 technology (a computer screen or data projector), or that an audio recording loses fidelity when played through a smartphone speaker as opposed to high-end studio monitor speakers.

129. We think that the history of audiovisual media leads people to assume, in certain contexts, that the correlation between 1 and 3 is very close in all cases. This is because it was historically difficult to modify Category 3 products or use Category 2 technologies in ways that were persuasive. Frequently, film-makers from Charlie Chaplin to Peter Jackson would use visual effects based on altering the physical environment such that category 1 technologies (non-digital in Chaplin's case) would generate illusory effects by capturing light in innovative ways that take advantage of the forced perspective generated by moving from a 3D to a 2D environment.

130. The law recognises that deception can arise from contextual circumstances where someone is reckless as to the truth of a representation, or omits a material particular in circumstances where there is a duty to disclose it (see Crimes Act 1961, s 240(2)(a)(ii) and (2)(b) in Part 3).

131. Condition 1 can be coupled with an additional act of deception, being an explicit affirmative statement that the category 3 product is a reliable representation of a category 1 technology. This tends to arise at the point the SMA is disseminated and can be separate from the SMA itself.

132. Condition 1 ascertains the perceived similarity of a given synthetic media artefact to an external phenomena. It can be stated in objective terms: does the virtual representation look or sound sufficiently similar to something in the physical world, so that a reasonable person observing would believe the representation was produced via "capture" or Category 1 technologies? In even simpler terms, does the image, video, or audio clip look or sound as if it must have been recorded by means of a camera or microphone? In saying that, the objective inquiry cannot be separated from the context in which the artefact is consumed: in video calls, for example, audiovisual quality frequently drops to account for fluctuations in data connections but the relative realism is maintained.

133. The presence or absence of Condition 1 in a given synthetic media artefact - or the degree of uncertainty in this assessment - is important across many areas of law that are common to most legal systems. Realistic audiovisual information has the potential for deception, harm, or loss that unrealistic information generally does not.

134. Frequently, we are conscious that there has been some degree of information loss at category 1, data modification at category 2, or deception arising from the implied representations inherent in Category 3 technologies. We think that our condition 1 describes the potentially deceptive aspect of SMT generating the most public concern. But not all media is consumed uncritically. Consumers critically assess the following kinds of Category 3 products based on how far we believe their audio-visual aspects have or could have been modified from an original state and how reliably they capture the relevant light and sound information at the point of capture. Consider how the implied representation at condition 1 varies in the following cases:

    - advertising;
    - Instagram photos and videos;
    - virtual reality environments;
    - CCTV camera footage;
    - traffic cameras positioned over congestion hotspots;
    - telephone conversations;
    - a documentary film;
    - a news broadcast on radio or television;
    - a smartphone camera using "selfie mode", "panorama mode", or "night mode".

135. Many legal regimes in New Zealand take a Category 3 product and require Courts and other enforcement bodies to assess the express and implied meaning of it. They do so to consider how it communicates express and implied messages to a consumer. Examples include the Films, Video and Publications Classification Act, the tort of defamation, the Privacy Act, the Fair Trading Act and the Harmful Digital Communications Act.

### Condition 2: The Effect of "Multiplicity"

136. In generating an SMA the relationship between the technologies at categories 1-3 is not necessarily linear or direct in the way a traditional photograph or sound recording might have been.

137. In producing an SMA, there may be multiple actors using multiple technologies in tandem in multiple ways from multiple sources to produce a single (or practically limitless volume of) synthetic media product. Further, the technologies, devices or software described by each category (1-3) will likely be capable of multiple functions, not just those described by 1-3.

138. Condition 2 accounts for the possibility that audiovisual information may move backwards and forwards through a synthetic media 'pipeline', with different actors involved at various points, and with Category 1, 2, and 3 technologies being repeatedly used at different stages and potentially by different persons.

139. Multiplicity accounts for many of the legal complexities of synthetic media. In any given analysis, there will be a question of whether the condition of multiplicity is engaged.

140. For example, a single digital photograph may be created in a benign context. This image may then be made available for consumption on the internet. At this point, Category 1, 2, and 3 technologies have all been used by one or more persons for a certain purpose. Sometime later, the digital data of the image may be manipulated again, and then re-displayed, potentially by a different person and for a different purpose, in a different context. The manipulated image might later be used to form part of a training dataset used to create a deepfake video or GAN-generated "human", which then is used for another purpose by another person. Supposing that this purpose is deceptive and causes harm or loss, there are questions as to the degree of involvement, accountability, and culpability of all actors active in this ongoing and potentially limitless pipeline.

141. We think multiplicity is anticipated by the Privacy Act and the Copyright Act, among other legal regimes, and discuss this in Part 3.

### Condition 3: The Act of "Dissemination"

142. The purpose of this condition is to draw a distinction between:

    a. the technologies by which synthetic media is generated, including synthetic media artefacts themselves; and

    b. the technologies by which synthetic media artefacts may be disseminated.

143. Stable definitions for separating these two things is important for law and regulation. Often an objection to synthetic media which is ostensibly about the artefact will actually be about the dissemination of the artefact - the fact it has been shared with or communicated to someone else.

144. This distinction is also important because there is a suite of different legal regimes that deal with the communication of media as distinct from the creation on that media. This reflects the fact that mere creation of a synthetic media artefact, in the absence of communicating it, may cause different (and lesser) harms.

145. For legal and regulatory purposes, it is important to identify where regulation is concerned with the dissemination of synthetic media and where it is concerned with the simple existence of it, or private consumption. It is also important to appreciate that the kinds of harms which may arise from either instance may be different, requiring different degrees of legal intervention or redress.

146. In any given analysis of a synthetic media technology or artefact, a central question must be the presence or absence of dissemination as distinct from the artefact itself.

147. We are conscious that, in the process of disseminating an SMA, the digital data comprising the SMA may be manipulated by a category 2 technology. For example, uploading a video to a social media platform often results in the use of compression algorithms that change the data comprising the SMA . We still think it is worthwhile to draw this distinction. In fact, the framework allows us to focus our attention on the ways in which dissemination technologies do manipulate SMA , and the extent to which manipulation itself can be benign or difficult to regulate.

148. Examples of legal regimes that deal specifically with harms from dissemination include the Privacy Act, defamation and the Harmful Digital Communications Act.

## Applied examples

149. Below, we briefly apply our framework to some technological examples to help illustrate how we have applied it in the course of our research. We also briefly refer to legal regimes in an illustrative way before explaining the next chapter precisely how we think they apply.

    a. A smartphone: contains technologies from all three categories. It can capture sound or light energy, convert this information into digital datafiles, manipulate that data, and display those files again as light or sound energy to be seen or heard. It enables the dissemination of SMAs and facilitates input by multiple actors into the creation process, including by making SMAs available for use by others to produce more SMAs. It produces SMAs that are immediately realistic enough that they create the impression a capture technology was used and that the use of manipulation technologies was minimal or innocuous, such that the records it produces are reliable. State-of-the-art smartphones tend also to contain multiple sensors.

    b. An oil painting: does not fall within our framework because it involves no digital technologies.

    c. Active noise cancelling headphones: are SMT containing Categories 1 and 3, and very often Category 2 also. They include microphones which detect environmental sound data and play frequencies over the top of a consumer's chosen SMA to enhance the user experience and eliminate background interference.

    d. A digital photograph: is an SMA because digital photography involves inherent manipulation processes intended to enhance picture quality. The digital camera is an SMT usually containing all of Categories 1-3.

    e. Music production software, such as Ableton Live: is an SMT containing Category 2 technologies that can make it sound like capture processes have been used in a way that satisfies Condition 1, either drawing attention to or minimising the role of those manipulation technologies. In practice, these deceptions create little risk of harm and are instead highly enjoyable forms of artistic expression. Software such as this also facilitates the conditions described by Condition 2 (multiplicity) as a way of enhancing the creative process between individuals.

    f. A GAN-generated face: is primarily the product of Category 2 technologies that manipulate digital data to produce a Category 3 output. GANs will both be trained on and produce from large datasets of category 3 outputs that may be the result of category 1 processes, but the person who deploys the GAN itself involves no category 1 capture. That GAN generated face is not harmful until it is disseminated in a deceptive way (per condition 3), but it contains significant potential for deception (per condition 1) because it is photorealistic to the point where a consumer would reasonably assume it is a reliable representation of the product of a capture technology and that any category 2 manipulations are either obvious or innocuous.

    g. Virtual reality: is an SMT involving primarily Category 2 and 3 technologies. That is because any capture that takes place requires such heavy manipulation before it appears veridical. However, virtual reality does capture user data by eye-tracking technologies, and contains some Category 1 features. Some virtual reality technologies may also coincide with technologies that record light or sound data based on a user's movement or voice. At this point, the computational processing power required to accurately render photorealistic virtual reality assets is so high that photorealism is sacrificed in many situations. Virtual reality also requires the use of

headsets that convey to the user that the content they are consuming is heavily manipulated. Therefore, its capacity for deception in terms of Condition 1 is low with current technologies.

h. In-painting: is a Category 2 technology trained using artificial intelligence paradigms trained from many Category 3 products. There is a high capacity for deception in terms of condition 1 and the product is intended to obscure the role of manipulation technologies.

i. Augmented reality or mixed reality: involves a Category 1 capture process to blend virtual environments with physical environments. Depending on its realism, there is a capacity for deception per Condition 1 because it can make it look as if there is a physical object in the environment being captured by a Category 1 technology. In practice users will be wearing a device that makes the deception and role of manipulation technologies obvious to them. The current state of technology is that the processing power required to render highly photorealistic environments is seldom justifiable for the end use case, and as a result, realism suffers.

j. Surrounding cameras in a new car: late model vehicles include a kind of reversing camera that displays an image in a centre-console as if the car was being viewed from a birds-eye perspective by another camera. In reality, no such birds-eye camera exists. Instead, the image of the car's surroundings is produced by synthesising the inputs from multiple cameras around the vehicle. They are therefore heavily manipulated. The technology is good enough and novel enough that it may not occur to the consumer that no birds-eye camera is in operation. It could therefore be deceptive, however it is reliable enough that you will use it to avoid property damage to yourself and others. The context in which you are viewing the Category 3 product is also relevant because it contains an implied representation from the manufacturer that the audiovisual product can be relied upon for that purpose.

k. Computer animation in an animated film: is an SMT involving Category 2 technology. It can create products that do not appear veridical at all in terms of Condition 1, because while they are capable of creating the impression of audiovisual perception, they do not create the impression that a Category 1 technology has capture light and sound in a real environment. Further, the role of Category 2 technologies is readily apparent.

l. A deepfake: an SMA produced by means of multiple Category 3 artefacts on which a Category 2 technology is trained, some of which may be the result of Category 1 captures. Again, the human deploying the category 2 technology on the Category 3 product does not require any Category 1 technology. It may be harmful because of the way it was created by impinging on privacy or copyright of others, or its contents may be harmful in the sense that it creates content restricted by the law. It might also be harmful at the point it is disseminated, regardless of whether it is claimed to be veridical or not. Deepfakes are capable of being sufficiently photorealistic that they are highly deceptive in terms of Condition 1: they create the illusion that a Category 1 capture took place and obscure the significant level of category 2 manipulation that has occurred.

m. A synthesised voice: an SMA generally produced via Categories 1-3 and which has a high risk of being deceptive per Condition 1. Users might deliberately provide audio data on which the voice can be trained, or it may be trained on pre-existing audio data created and collected in a different context. Category 2 manipulation technologies then process this data to create a new Category 3 display product that makes it sound as if a capture technology (a microphone) is relaying a person's speech directly. The technology's intended realism is meant to minimise consumer awareness that Category 2 technologies are being deployed. This was the reason for concern about the use of Google's Duplex, which did not identify itself as an artificial intelligence technology when conversing with a human.

n. A hologram: is a Category 3 technology that, because of the novelty of hologram technology, makes it readily apparent that Category 2 technologies play a heavy role in the media being consumed by the consumer. Depending on the representations made to the consumer and the context in which it is being consumed, there may or may not be Category 1 technologies involved, although in the creation of the hologram, Category 1 technologies may have provided the building blocks for the hologram product (through the use of motion or performance capture suits).

o. A 3D printer or printed artefact: does not fill well into our framework. Because it is a static object printed in a medium, it does not display light, but merely reflects it from other ambient sources. However, there is a low capacity for deception in terms of Condition 1: any reliance on the reliability of a 3D printed object's resemblance to the external physical environment is difficult to conceive. We think that, in any event, any situation where someone is relying on the relationship between a 3D printed object and a capture process would also involve repeated representations by a human actor, which would be caught by other legal regimes without the need to refer to emerging synthetic media issues.

p. Twitter, Facebook Live, Youtube Live: are primarily dissemination technologies. However, they also deploy capture technologies and manipulation technologies that process the data being disseminated. Social media platforms use compression codecs that alter the digital information in a video without materially changing the apparent relationship between the light captured by the camera and the light broadcast to the consumer. Notably, drops in streaming quality that hamper photorealism do not undermine user perception of the reliability of the stream as a reflection of events in the physical environment. The apparent difference between Facebook Live and videoconferencing technology is primarily the breadth of the audience, and therefore a dissemination issue. The immediacy of Facebook Live can be an indication that it is reliable because the opportunity for the use of deceptive category 2 technologies is ostensibly minimised because of the "live" nature of the stream. However increasingly, many manipulations may occur even in real-time. Further, the interactivity of the stream and the way that unanticipated events can interrupt it can also be contextual or content-related indicators of its reliability.

## Summary of Part 2

150. In summary, we conclude from Part 2 that:

a. the features of synthetic media technologies and the wide range of conceivable uses of them are so broad that generating and analysing hypothetical fact patterns is of limited value. For this reason, an iterative framework-based approach has been preferred.

b. The framework creates a broad definition of synthetic media technologies and artefacts that acknowledges that digital manipulation technologies are pervasive and often innocuous.

c. The framework allows policymakers to isolate the features of a given technology and assess how far those are new or similar to what already exists.

d. The framework also calls attention to ways of describing the particular act that is alleged to be harmful: whether creation, content or dissemination. In this way, we can identify existing analogues in the law.

e. We think that any discussion about the harms of SMT and SMA can be answered by reference to the elements of our framework.

151. In Part 3, we apply the framework to examine how the law catches the kinds of actions and technologies described in Categories 1-2 and the way that the law anticipates the impact of Conditions 1-3 on the way that the categories of technologies are used.

# Part 3: New Zealand law applied to SMA and SMT

## Summary of our findings when applying the framework to NZ law

152. When understood through the lens of our framework, we believe that there are multiple legal regimes in New Zealand that apply to the use of synthetic media technologies to produce synthetic media artefacts.

153. Our goal is only to show that such legal regimes could apply or be affected by the rise of SMT. We cannot anticipate all the ways in which SMA will be used to generate harm or dictate what the consequences would be in the abstract. Instead, anyone concerned about the use of synthetic media will have to identify the specific harm they are seeking to remedy and work backwards through these existing legal regimes according to the specific fact pattern of a given scenario.

154. This section of our report illustrates how SMA and SMT could be caught by identifiable legal and non-legal regimes that restrict the conduct of actors in New Zealand in terms of creation, content and dissemination of SMA. We therefore focus heavily on legal definitions to demonstrate that the subject matter of our report is caught in some way. Because of the factual indeterminacy of the subject, we restrain ourselves from adopting a position on whether the legal regime is adequate or not, although we point to obvious gaps where appropriate. The point is that the identified legal regimes represent a limitation on dealing in synthetic media in some way that has already been through a democratic process, and should therefore be considered before new restrictions are introduced.

155. Most if not all of these legal regimes acknowledge the difficult interaction between freedom of expression and other legal values, including privacy. They adopt a case-by-case approach and express caution about the idea that an exhaustive universal standard can be articulated in the relevant area. There is frequent use of principle-based frameworks which can only be applied to specific facts.

156. At a high level, we conclude that the law generally does not require us to 'go behind' the SMA itself and assess the process by which it has been created, unless that is explicitly called for by the relevant law. This considerably simplifies much of the discussion about synthetic media.

157. We are not confident that any legal or policy process could achieve a more certain or universally agreeable standard for intervention than what is already articulated in law. Accordingly, it is preferable to leave that existing law to be applied in appropriate cases by expert application of law to proven facts.

### Common issues regardless of legal regime

158. We note some key issues are likely to be faced under any legal regime utilised:

    a. some issues are likely to be particularly difficult to show from an **evidential perspective**, including **digital forensic issues** and the **identification of relevant agents**. Some companies such as DeepTrace specialise in technological and software solutions in this emerging market.

    b. **jurisdictional issues** will arise from the role of international companies at various stages of the SMT process.

    c. **access to justice issues** are inevitable in some form, including the potentially low value of disputes in relation to the cost of pursuing them, the complexity of identifying which part of the legal system to work through, and the nuance of the legal issues involved.

d. **law has a retrospective orientation** that means it generally only intervenes in events that have already occurred. This means prevention is primarily achieved via deterrence rather than direct intervention. Pre-emptive intervention would increase the risk of unjustified censorship, but would help prevent harmful events occurring.

e. it can be difficult to establish **causative connections** between the kind of harm alleged to have resulted and the identified use of SMA or SMT alleged to have caused that harm. Further, the law can find it difficult to recognise certain kinds of **diffuse or disparate harm**, such as "loss of trust" or generic impacts on the democratic process.

## Structure of Part 3

159. The regulation of synthetic media entails a complex interaction between the law as it relates to privacy, freedom of expression, property rights in original works through copyright, and the use of digital media to inflict criminal deception or harm.

160. Consistent with this overall conclusion, we have grouped the legal regimes as follows:

a. Individual privacy and data protection;

i. Privacy Act 1993;

ii. the analysis by the Court of Appeal in *Hosking v Runting* and *C v Holland*.

b. NZBORA and limitations on freedom of expression;

i. New Zealand Bill of Rights Act 1990;

ii. Broadcasting Act 1989;

iii. Electoral Act 1993

iv. Films, Videos, and Publications Classification Act 1993

v. Defamation

vi. the Media Council of New Zealand Guidelines

vii. Human Rights Act 1993

c. Interpersonal harms that are Criminal or approaching criminal;

i. Crimes Act 1961

ii. Harmful Digital Communication Act 2015

iii. Harassment Act 1997

iv. Fair Trading Act 1986 and Advertising Standards

d. Copyright and the rights of creators

i. the Copyright Act 1994;

ii. indigenous intellectual property.

# Individual privacy and data protection

## Summary of privacy and data protection

161. Synthetic media artefacts and technologies are and should be dealt with through an individual privacy and data protection lens. This can be done in combination with the other areas we identify in this part of the report when harms of a different nature arise.

162. We conclude that the Privacy Act 1993 applies to the creation and use of synthetic media. Without exhaustively detailing our analysis, we have no reason to believe that the Privacy Bill (in its current form) will change this conclusion.

163. While the Privacy Act applies, there are limitations on the specific restrictions the Act applies to personal information.

164. We refer to the discussion of the Court of Appeal in *Hosking v Runting* [2004] NZCA 34 (25 March 2004); [2005] 1 NZLR 1; (2004) 7 HRNZ 301, despite its age:

    a. as an authoritative legal statement on the role and sources of privacy in New Zealand law and the way that privacy interacts with other legal instruments. We draw support from that approach for our own analysis of how a range of enactments can touch on the specific harms arising from the creation, content and dissemination of SMA by the use of SMT.

    b. *Hosking v Runting* also explicitly notes and rejects a cause of action based on misappropriation of image in New Zealand.

165. Because of those conclusions, we reject any argument that the concept of "personality rights", "publicity rights", or any kind of property right in one's audiovisual profile is of any use in New Zealand. The interests intended to be protected by these doctrines are already covered by New Zealand law. To the extent there are gaps, they should be filled by extensions to statutory privacy frameworks.

## The Privacy Act 1993

### The purpose of the Privacy Act 1993 and the Privacy Bill

166. Many of the hypothetical harmful uses of synthetic media relate to the way that they can show an identifiable individual doing or saying something they never did.

167. We already have a framework for dealing with information about identifiable individuals in New Zealand – the Privacy Act.

168. The Act is currently undergoing relatively significant amendment. There is only so far we can or should take that amendment into account. Generally speaking, any reference here is to the Privacy Act as currently enacted unless otherwise stated.

169. We note, however, that the current form of the Privacy Bill states its purpose at clause 3 acknowledges the need to balance a "right to privacy" with other "rights and interests", and referring to international human rights frameworks.

> to promote and protect individual privacy by—
> (a) providing a framework for protecting an individual's right to privacy of personal information, while recognising that other rights and interests may at times also need to be taken into account; and

(b) giving effect to internationally recognised privacy obligations and standards in relation to the privacy of personal information, including the OECD Guidelines and the International Covenant on Civil and Political Rights.

170. The Privacy Act is open-ended to allow for developments in technology and privacy practices. We think it applies to current SMT and will govern the bulk of SMT created in the future to the extent they interact with respect for privacy as the autonomy and dignity of an individual to control their presentation to the world.

171. Further, we think that, to the extent there is any doubt about the application of the Act, it should be clarified by legislative amendment in favour of its inclusion. It is a good, comprehensive regime for dealing in digital information about identifiable individuals.

### Does Privacy fit?

172. There are some potentially counterintuitive aspects to the application of privacy law to synthetic media. We note them here for completeness and deal with each of them in kind throughout this report.

a. Almost by definition, a synthetic media artefact like a harmful deepfake does not depict or broadcast 'correct' information about the person involved. It is not a disclosure of private information. The concern is that the SMA is deceptive or misleading. In terms of condition 1 of the framework, it is possible that no capture whatsoever occurred in the creation of information about that individual. The artefact is highly manipulated.

b. Given Condition 2 (multiplicity), the source data for an SMA may be drawn from a wide range of sources, some or all of which may be "publicly available". To the extent that capture technologies are used in the generation of the synthetic media, the product will, to a greater or lesser extent, reflect the product of those capture technologies, being audio or visual recordings of humans.

c. The human face and voice are, in Western cultures, generally publicly available. Many of the leading cases in tort question the extent to which a photograph taken in a public place of a person's face can be the subject of litigation on the basis of privacy concerns, and how privacy can be a workable concept in that situation.[97] There is concern about giving individuals undue control over what are, essentially, publicly available materials.

### Synthetic media and the definition of "personal information"

173. In order to be the subject of a complaint by an individual, there must be, at some level, a risk that the individual will be identified as the subject represented in the SMA. In many cases, we think this will avoid any need for a threshold test or analysis to be applied by any agency monitoring synthetic media. Effectively, in order to have stimulated a complaint, the majority of complaints made by identifiable individuals about representations in synthetic media will already bear a passing resemblance to that individual, with the exception of acutely sensitive, vexatious or unreasonable individuals.

174. The key issue under the Privacy Act is whether SMAs can be "personal information". It will be personal information governed by the Act if it is information "about an identifiable individual".

175. The Act imposes limitations on the way that agencies (including individuals) can deal in personal information through a series of information privacy principles ("IPP"). An action can be an interference with privacy if it breaches an IPP and causes a specific kind of harm. The principles only apply to "personal information".

[97] This issue is canvassed in *Hosking v Runting* in particular.

176. The question of whether a particular SMA is "about an identifiable individual" is heavily fact-oriented. It must be examined on the evidence in the circumstances. It would be fruitless to speculate in advance for all cases and the definition is intended to be broad to allow for application to future cases:[98]

> "… there is no 'bright line' test which separates that which is obviously personal information about an identifiable individual from that which is not. Much will depend in any given case on the context in which the information is found. There may be particular factors in different settings that compel a conclusion that, … there is a sufficient connection between the information and the requester to justify a conclusion that the information is personal information..."

177. We note that the Law Commission has also stated that: "It seems to be undisputed that "personal information" covers information collected or held in a wide range of forms, including audio and visual recordings."[99]

178. Consistent with our earlier conclusions, we think there is no way to draw a sustainable distinction for all cases between a common digital audio or visual recording and synthetic media products only on the basis of the level of manipulation involved, or its "fakeness", apart from through the nuanced application of Condition 1 of our framework. Accordingly, if digital audiovisual recordings are caught by the Act, then so is synthetic media, so long as it is "about an identifiable individual".

179. The extent to which the Privacy Act will provide an effective remedy in relation to the particular SMA or SMT will depend heavily on the nature of the synthetic media in question. The Privacy Act may also raise legal issues at all stages of the synthetic media creation process: capture, manipulation, display, dissemination and verisimilitude.

180. We refer to the definition of a "document" in s 2 of the current Act because we think it includes various kinds of SMA and also note that other enactments such as the Crimes Act also employ a definition of "document".

> "document means a document in any form; and includes—
> (a) any writing on any material:
> (b) any information recorded or stored by means of any tape recorder, computer, or other device; and any material subsequently derived from information so recorded or stored:
> …
> (e) any photograph, film, negative, tape, or other device in which 1 or more visual images are embodied so as to be capable (with or without the aid of some other equipment) of being reproduced"

181. Regardless of the relative sophistication of an SMA, from virtual human to digital photograph, it will be digital information stored by means of a computer or other device per para (b), and almost certainly will be able to be described as "material subsequently derived from information so recorded or stored".

182. If para (b) were not enough to indicate the legislature's intention, then its explicit reference to "1 or more visual images" in para (e) puts this beyond argument. We note that (e) includes the notion of reproduction from a "device in which … images are embodied so as to be capable … of being reproduced". We have incorporated a similar notion into Category 3 of our framework in the way that Category 3 includes digital files capable of generating light or sound energy for display to human senses, "with or without the aid of some other equipment".

183. We think that the definition's reference to "material subsequently derived from information so recorded" also accounts for condition 2 of the framework, in the sense that it anticipates that multiple documents could be created from a single document in a kind of creative chain across

[98] *CBN v McKenzie Associates* [2004] NZHRRT 48 (30 September 2004) at [41]. Followed by *Taylor v Corrections* [2018] NZHRRT 35.

[99] New Zealand Law Commission "Review of the Privacy Act 1993: Review of the Law of Privacy Stage 4" Issues Paper 17, March 2010 at 3.5, available from: <www.lawcom.govt.nz/>.

individuals. To the extent that creation process is digital, the definition anticipates the use of Category 2 manipulation technologies.

184. The definition is materially unchanged in the Privacy Bill,[100] and both legislative instruments make it clear that personal information may be stored in documents such that access to the document may be required under the Act by the identifiable individual.[101] The fact that personal information is contained in a document gives both a requester and an agency varying ways to negotiate access to that personal information depending on the context with specific allowances for sounds or visual images. We note the exception for trade secrets at s 28 of the Act.

185. The Law Commission also had this to say about the definition of personal information in the Act as drafted, and it is one of the reasons we call for guidance from the Office of the Privacy Commissioner on our conclusions about the Act:[102]

> "Leaving the meaning of personal information to be clarified through opinions and decisions in particular cases has the advantage of flexibility. There are also some issues (such as the meaning of "about") that can probably only ever be resolved in relation to the facts of specific cases. However, it takes time for a consensus to develop in the jurisprudence, or for a suitable case to lead to an authoritative court decision. Clarifying the meaning of the Act through jurisprudence is also less accessible to users of the Act than stating matters in legislation or official guidance."

### Wrong personal information is still personal information

186. One possible objection to the use of privacy law to govern synthetic media artefacts is that, by virtue of being synthesised, emerging or novel, SMAs will not show real personal information. Put bluntly, they are not "about" that individual at all. They may be verisimilar and persuasive, but they are non-veridical. The things represented in an SMA never actually took place.

187. A similar argument can be made that the law tends to leave publication of wrong facts about an individual to the law of defamation, suggesting it is a poor fit with privacy.[103]

188. We deal with arguments of this nature in our discussions of this Act, of the analysis in *Hosking v Runting*, and the content of the Harmful Digital Communications Act elsewhere in this report. We do not think it is persuasive, to the extent that it attempts to exclude non-veridical SMA from the ambit of privacy law.

189. In relation to the Privacy Act specifically, we think that wrong information about an identifiable individual can still be information about that identifiable individual. That must be the case because of principle 7 of the Act, one of its cornerstone principles in connection with the principle 6 right to request access.[104]

190. If wrong personal information is not personal information, then principles 6 and 7 are rendered ineffective, or at least unworkably complex.

191. Further, a key purpose of the Act is to allow individuals to identify situations where an agency may have relied on incorrect or misleading information about them to their detriment (principle 8), and seek redress.

192. In this sense, the definition of personal information could be read as being "information [that purports to be] about an identifiable individual". The question of whether it is correct only comes

[100] Privacy Bill 2018 (34-2) Clause 6.
[101] Privacy Act s 42, Privacy Bill at Clause 62.
[102] New Zealand Law Commission "Review of the Privacy Act 1993: Review of the Law of Privacy Stage 4" Issues Paper 17, March 2010, at 3.30.
[103] Noting *Hosking* at [138].
[104] See the way that the Human Rights Review Tribunal linked Principles 6 and 7 in paras 96 and 130 of *Watson v Capital & Coast District Health Board* [2015] NZHRRT 27 (7 July 2015).

later. If read this way, the amended definition reflects Condition 1 of our framework, because it takes into account the complex interaction between the content and context of an SMA in creating meaning from initial impressions, explicit statements and context, including whether it looks like a Category 1 technology was used and whether the role of Category 2 technologies are apparent. If the definition is read in this way, we think there can be very little dispute that synthetic media information is personal information governed by the Privacy Act, and the fact that the synthetic media information is not veridical is immaterial.

193. To the extent that the privacy of other people may be infringed by misidentification of one individual as another individual, then the Act allows for redactions to be made to documents or for information to be withheld for this purpose in a way that can be challenged by complaint if necessary (s 29(1)(a)).

194. It is also important to note that, in practice, there will be a number of indicators linking an SMA to an identifiable individual, for example through data collection and organisation practices and unique identifiers. The situation where an individual is identifiable solely from the SMA at the point of consumption is likely to be rare unless it has been widely disseminated. We also note the prospect that facial recognition algorithms could be used to enhance individual privacy by making individuals aware of information "about" them as identifiable individuals in datasets they would otherwise be unable to process.

195. One interesting question is whether someone could use principle 7 to request "correction" of (what appears to be) audio or video recordings, or have a video statement of correction attached to a video alleged to be incorrect. That would depend on the nature of the incorrectness involved: for example, whether it has been heavily manipulated, or whether it is non-veridical, or whether it is simply unreliable when taken out of context. Does a person have, for example, the ability to request that an image be photo-shopped in a more attractive way on the basis that the image is not an accurate reflection of their appearance? This illustrates the limitations of "truth" or "fakeness" as a boundary standard. In line with condition 1, the "fakeness" of an SMA is better assessed by reference to its context or purpose rather than in an abstract sense. There is a degree of "manipulation" inherent in digital media during the capture and processing stages.

### The Privacy Act and Condition 2 of the framework (multiplicity)

196. Condition 2 acknowledges that it can be difficult to identify and assign culpability to the range of actors who may be involved in the process of generating and disseminating synthetic media.

197. We think some definitions in the Act indicate that Condition 2 of our framework can be taken into account by the Privacy Act as a legislative framework. In particular, it acknowledges that different agents may have different roles in relation to the use of synthetic media technologies and artefacts, entailing different obligations.

198. The definition of "action" includes a failure to act as well as the role of any policy or practice. This could embrace a degree of carelessness or recklessness when it comes to dealing in personal information in breach of an IPP.

199. Agency is defined to include a body of persons or any person, but does not include a news medium in relation to its news activities. It also exempts certain agencies through s 3(4):

> (4) For the purposes of this Act, where an agency holds information—
> (a) solely as agent; or
> (b) for the sole purpose of safe custody; or
> (c) for the sole purpose of processing the information on behalf of another agency,—
> and does not use or disclose the information for its own purposes, the information shall be deemed to be held by the agency on whose behalf that information is so held or, as the case may be, is so processed."

200. We note that s 3(4) will undergo relatively substantial amendment if clauses 3A, 8 and 9 of the Bill for example are enacted, but simply argue that this reflects further support for our conclusion that the legislature's intent is that the kinds of issues anticipated by condition 2 of the framework can be dealt with using privacy legislation as a framework.

### Collection and creation under the Privacy Act

201. Further discussion is required on the implications of treating "collection" as including "creation" or "generation" under the Act.

202. Principles 1-4 of the Act regulate the collection of both the data necessary to create SMAs and the collection of SMAs themselves, however simply reading them as if the word "generated" or "created" was substituted for "collection" leads to some awkward phrasing (eg "creation from the individual concerned" in principle 2(1)) as well as some easy substitution ("person information shall not be created or generated by any agency unless the creation or generation of the information is for a lawful purpose and is necessary for that purpose" in principle 1).

203. We think that, based on a purposive approach, "collect" could also be read to include the action of "creating" or generating. In support we refer to *Armfield v Naughton* [2014] NZHRRT 48 (6 October 2014) at [39]-[45], and in particular at paras 41.2 and 44.3:

> [41.2] … Surveillance usually results in the collection of personal information and information collection is one of the main purposes for which surveillance is used. In fact the Group of Experts state in their Explanatory Memorandum to the OECD Guidelines at [52] that the second part to the Collection Limitation Principle is directed against the use of surveillance devices:
>
> > 52. The second part of Paragraph 7 (data collection methods) is directed against practices which involve, for instance, the use of hidden data registration devices such as tape recorders, or deceiving data subjects to make them supply information. The knowledge or consent of the data subject is as a rule essential, knowledge being the minimum requirement …
>
> As stated by the Law Commission in its June 2011 Report at [2.81], the current definition of "collect" is not intended to exclude the obtaining of personal information by means of surveillance devices. The purpose of and background to the Act suggest that surveillance should be considered to be a form of collection
>
> ...
>
> [44.3] Individual privacy will be promoted and protected by giving to the term collect a broad meaning. The term is not a synonym of "solicit". It is to be given the purposive meaning of "gathering together, the seeking of or acquisition of personal information".

204. Collection will often require the act of recording, through surveillance devices or otherwise, and the generation of a new record from old records using digital technologies. The broad definitions of "personal information" and "document" give support to this idea of "collection" as generation, particularly by reference to information "subsequently derived" at para (b) of the definition of "document".

205. The alternative would be that the Privacy Act primarily controls the use of Category 2 manipulation technologies or technologies of dissemination. The shortcoming of adopting this approach is that the restrictions imposed by privacy principle 5 are relatively limited and rely heavily on the original purpose of "collection". This would mean that, having authorised collection for a purpose, subsequent rights of control would be relatively limited.

206. For completeness we note that an agency who stored large quantities of publicly available information from the internet would still be collecting personal information. Instead, the Act applies different restrictions on that information given its public availability.

207. Notably, s 2 of the Act excludes the unsolicited receipt of information from the definition of "collect" under the Act, but unsolicited receipt of personal information does not avoid an agency's obligations to deal with it in light of other privacy principles.

208. In summary, the Act may impose a degree of control for users in the following ways, however both involve some awkwardness or gaps in the framework set out by the Act, and because of their commercial and private implications, would benefit from wider discussion:

    a. Generation or creation of synthetic media artefacts as "collection" (privacy principles 1-4); and/or

    b. Generation or creation of synthetic media artefacts as "access, use or modification" (principle 5).

### Publicly available information and the Privacy Act

209. To the extent that the Privacy Act will govern SMAs about identifiable individuals, it is highly likely these SMAs will be recordings of that individual's face or voice. In many cases, that face or voice is unavoidably public.

210. Privacy law has a history of struggle with the boundary between public and private facts: we deal with this in greater detail in the context of our discussion in *Hosking v Runting* as an illustrative example.[105] The logic goes that there can be no expectation of privacy in something that could be observed in a public place, because this would make privacy law unworkable and allow unacceptable limitations to be imposed through law on freedom of expression.

211. Here, we deal with this issue in the context of the Privacy Act specifically, which includes the following definitions at s 2:

> publicly available information means personal information that is contained in a publicly available publication
>
> publicly available publication means a magazine, book, newspaper, or other publication that is or will be generally available to members of the public; and includes a public register.

212. The Privacy Act does not exclude such information from the definition of being "personal information", but it does limit someone's entitlement to control that information via the privacy principles. For example, Principle 2(2)(a) means that an agency is not required to collect information directly from the individual concerned if the agency has reasonable grounds to believe it is publicly available. Similarly, (b) allows agencies with reasonable grounds to believe that the individual authorises collection of the information from another source to collect it from that source.

213. Similarly principle 10(1)(a) states that:

> "An agency that holds personal information that was obtained in connection with one purpose shall not use the information for any other purpose unless the agency believes, on reasonable grounds,—
> (a) that the source of the information is a publicly available publication and that, in the circumstances of the case, it would not be unfair or unreasonable to use the information; or
> (b) that the use of the information for that other purpose is authorised by the individual concerned …"

214. Principle 11 states that:

> "An agency that holds personal information shall not disclose the information to a person or body or agency unless the agency believes, on reasonable grounds,—

[105] *Hosking v Runting* [2004] NZCA 34 (25 March 2004); [2005] 1 NZLR 1; (2004) 7 HRNZ 301.

(a) that the disclosure of the information is one of the purposes in connection with which the information was obtained or is directly related to the purposes in connection with which the information was obtained; or

(b) that the source of the information is a publicly available publication and that, in the circumstances of the case, it would not be unfair or unreasonable to disclose the information; …"

215. These principles illustrate how publicly available information may still be subject to rights of control by an identifiable individual. Apart from these legal responses, there is another answer to the suggestion that the Privacy Act will not produce meaningful remedies for synthetic media on the basis that a person's face or voice are publicly available, or produced from publicly available materials. That answer draws attention to the distinction between the thing depicted and the artefact depicting it.

216. While a person's appearance or "sound" in the abstract sense of their audiovisual identity may be generally public, the artefacts that are produced through SMT create a separate "document" or record that was not publicly available. This document itself – purporting to demonstrate a set of factual events at a certain place and time – was not publicly available at the point of creation, even if somebody on the street could observe that person doing a similar act if they chose to do so. This is an important difference between the general question of whether someone has a reasonable expectation of privacy in public (as discussed in *Hosking*) and their ability to control "personal information" (including SMAs) about themselves.

217. The Privacy Act regulates information, including in recorded form. In this way, it avoids some of the fraught questions of identity definition discussed in relation to publicity rights by regulating the SMA itself rather than the thing it appears to depict.[106] In other words, if publicity rights protect an individual's distinctive audiovisual profile, what exactly is required to draw a consistent boundary around this profile in order to exclude someone from it? There are no simple answers to this question, as outlined by Zapparoni. We acknowledge some of these questions of identity will simply be absorbed into a wider question about whether the artefact is "about" the individual in question. They are unavoidably fact based, and just as likely to arise in relation to the Fair Trading Act, Broadcasting Act, or Advertising Standards Authority provisions that we identify elsewhere in this report.

### Implications of retaining privacy in generated artefacts

218. A result of our conclusions is that an individual will retain a degree of control over their personal information even where that personal information is in the form of a generative SMA that has been sold. For example, if I sold my audiovisual likeness to a company in a way that would enable the company to deploy my profile as a chat-bot or digital human, or for repeated use in advertising, it would still be information about me as an identifiable individual broadly speaking. In that case, when does an individual lose their ability to control SMAs about them? For commercial purposes, how can the connection be broken between an individual's control over SMA's about them as a privacy right, our broad definition of an SMA, and the strong link we have drawn between those things?

219. There are several answers to this which merit further investigation.

220. One answer is that the connection between an individual and their personal data can never be absolutely severed. We think this must be accurate and the eventual policy position reached in New Zealand given the essential nature of the right to privacy as a human right. The European Data Protection Board is an agency responsible for applying the General Data Protection Regulation (GDPR).[107] On 9 April 2019 it adopted a document entitled "Guidelines 2/2019 on the processing of

[106] See Rosina Zapparoni "Propertising Identity: Understanding the United States Right of Publicity and Its Implications - Some Lessons for Australia" (2004) 23 MelbULawRw; (2004) 28(3) Melbourne University Law Review 690.

[107] General Data Protection Regulation 2016/679.

personal data under Article 6(1)(b) GDPR in the context of the provision of online services to data subjects". At para 55 of those guidelines, it states:[108]

> Considering that data protection is a fundamental right guaranteed by Article 8 of the Charter of Fundamental Rights, and taking into account that one of the main purposes of the GDPR is to provide data subjects with control over information relating to them, personal data cannot be considered as a tradeable commodity. Data subjects can agree to processing of their personal data, but cannot trade away their fundamental rights.

221. As acknowledged by the very existence of the guidelines, however, this does not make commercial dealing in data impossible.

222. While we are conscious that the GDPR operates in a different legal environment to that of New Zealand, we observe that:

    a. The Courts in *Hosking v Runting* and *C v Holland*[109] both refer to international legal instruments in support of their findings on torts of privacy in New Zealand meaning that it may be relevant to interpretive and other inquiries conducted by Courts where appropriate.

    b. The Privacy Act 1993 was itself adopted in order to provide compliance with international OECD guidelines. Accordingly, there is a need for New Zealand's privacy regimes to keep pace with international trading partners.

    c. Clause 3 of the Privacy Bill, which would act as an updated interpretive guide to New Zealand courts about Parliament's intent, records a parliamentary purpose of "giving effect to internationally recognised privacy obligations and standards in relation to the privacy of personal information, including the OECD Guidelines and the International Covenant on Civil and Political Rights." The GDPR is also a defined term in the Bill.

223. We also think that the law does not permit agencies to contract out of the Privacy Act entirely, although clearly there is scope for agreement within its overall application. In *Director of Human Rights Proceedings v Schubach* [2015] NZHRRT 4, the Tribunal held:

> [66] For the reasons given we are of the clear and firm view that the text and purpose of the Privacy Act do not permit its terms to be circumvented by an agency contracting out of its statutory obligations. The protest to jurisdiction is dismissed.

224. This case has not been tested other than in respect of the level of awards made by the Tribunal. We think that the current drafting of clause 3A in the Privacy Bill will support the Tribunal's conclusion.

225. A second answer is that people can give their consent through contract to a broad discretionary use of the information for vague purposes and a long duration, including wide rights of disclosure. This could remain operative even if they retain ultimate rights of control over their personal information. This will substantially affect the reasonableness of their expectation to have their information dealt with in certain ways, including perhaps the quantum of any damages available to them. Further, any attempt to withdraw consent for something promised as consideration in a contract could be treated as repudiation and dealt with according to orthodox legal principles. It would also be important to account for the varying allocation of copyright and property interests in the information in question.

226. There is no right to erasure under the Privacy Act. Further, once consent has been given for collection, generation or use for a wide purpose, there is no obvious way to amend or withdraw the terms of this consent through the Act.

---

[108] Ibid at article 6(1)(b) para 55.
[109] *C v Holland* [2012] NZHC 2155; [2012] 3 NZLR 672 (24 August 2012)

227. A third answer arises from condition 1 of our framework. We suggest that the connection between a person and information ostensibly "about" them can be severed by taking obvious steps to undermine the veridicality of the SMA, such as through structuring of context and the use of explicit statements that the SMA is not to be taken as evidence of the truth of its contents. While this would not mean the information ceases to be "about" an identifiable individual, it would drastically limit that individual's entitlement to assert control over it or allege harm has been caused by it. In other words, issuing a statement accompanying the SMA that it is not the product of Category 1-style capture of "real events", or does not actually represent the person or events it purports to depict.

228. To illustrate, consider the way that the Privacy Act would treat audiovisual recording of a theatrical performance. At first instance, the identifiable individual (the actor) has given their consent for their audio-visual profile to be represented in a media artefact. There may be express or implied contractual terms that go to purpose, duration, disclosure, use, etc. There will also be matters of commonly accepted practice, such as the idea that a theatrical performance will not be produced as evidence of the truth of its contents – ie that the actor is a person called Julius Caesar who was murdered by a group of Roman Senators. However, the audio-visual footage may be useful evidence that a person with the actor's features appeared in a theatrical performance of Shakespeare's Julius Caesar that was recorded at a particular date and time, and therefore could not have been present at another location where an alleged crime was committed. The information itself will be "about" an identifiable individual, but what the information says about that individual is a matter of context and communication existing apart from the media artefact itself.

229. If we take the example of a digital human, who resembles a real identifiable individual, the resulting synthetic media artefact will be, at some level, personal information under the Act. However, with careful contextual information, any subsequent use of that personal information will only be "about" them in the sense that it records that their audio-visual profile was once captured through a synthetic media technology. It may also be information about them in the sense of their consent to have their audio-visual profile used in a particular way, as in the case of a celebrity who has licensed the use of their audio-visual appearance for recreation as a digital human. Through other design elements, it can be made clear that the digital human is not information "about an identifiable individual". We understand that companies such as Soul Machines are moving toward a situation where elements from identifiable individuals will be blended to create new profiles rather than attempting to reproduce identifiable individuals.

### "Personality rights" or "publicity rights"

230. Because of our conclusions, we do not see any merit in the development of legal doctrines of personality rights (typically a privacy right associated with German or European rights of privacy flowing form personhood under civil codes) or publicity rights (a pseudo blend of property and privacy concepts giving control over public profile) in New Zealand.[110]

231. The substantive interests represented by those "rights" of personality and publicity are already protected in New Zealand law.

232. Under all three of the Privacy Act, publicity rights, and personality rights doctrines, a common legal issue is the need to draw boundaries around what exactly a person's identity or audiovisual profile

[110] Consider the analysis of Petra Butler: "A Dworkinian Right to Privacy in New Zealand" in Salman Khurshid, Lokendra Malik and Veronica Rodriguez-Blanco (eds) *Dignity in the Legal and Political Philsophy of Ronald Dworkin* (Oxford University Press, India, 2018) pp 433-465; "The Case for a Right to Privacy in the New Zealand Bill of Rights Act" (2013) 11(1) *New Zealand Journal of Public and International Law Special Issue - 21st Birthday of the New Zealand Bill of Rights Act 1990* pp 213-256; and Zapparoni above n 106. See also the finding of the US Copyright Office on the inconsistency in publicity rights across states in America in United States Copyright Office "Authors, Attribution, and Integrity: Examining Moral Rights in the United States" (April 2019) <https://www.copyright.gov/policy/moralrights/full-report.pdf>.

is. It is preferable to focus on the use of that identity, rather than attempting to draw an exclusionary boundary around it for all future purposes.

233. New Zealand's legal system generally protects definable interests rather than creating general rights. Further, where it does create general rights (as in the NZBORA) these always face a balancing exercise, and so the terminology of rights does not assist when it comes to the balancing exercise inherent in a person's ability to control their audio-visual profile in particular contexts.

234. We think it is unnecessarily complex to attempt to ascribe a property framework to an individual's audiovisual profile. Property is frequently (though by no means exclusively) conceptualised as "a bundle of rights", as well as a set of relationships which can be enforced against other individuals. In practice, this "bundle of rights" is exactly what New Zealand law provides through existing legal regimes.

235. We do acknowledge that a property framing through publicity rights in audiovisual profile would provide a degree of descendability to the heirs of identifiable individuals. While we note this as a policy factor, we think this can be achieved through other means, including by copyright or broader developments in the law of privacy as it relates to deceased individuals: for example, "the right to be forgotten". We think that privacy is already being forced to deal with the concept of the rights and privacy of deceased individuals, and thus it would be better to deal with that discussion using privacy concepts rather than attempting to transition New Zealand towards greater emphasis on property concepts.

236. We note that property is already a difficult concept when it comes to digital data in a criminal context. We note the disagreement between Court of Appeal,[111] and Supreme Court,[112] in the Dixon appeals on whether digital CCTV footage can be property, including subsequent commentary on those decisions.[113] Most audiovisual representations of a person's profile will be in the form of digital datafiles, capable of being broadcast as Category 3 technologies. There is therefore the prospect of multiple property interests coinciding in the same SMA, particularly when copyright is also incorporated into the analysis.

### Reliance on information without recourse to subject

237. One of the key harms that could arise from SMA is by the reliance upon an artefact as being veridical to the detriment of the subject. The Privacy Act has answers to this problem too.

238. Information privacy principles 1, 2 and 3 are intended to avoid a situation where a piece of synthetic media is relied upon without recourse to the apparent subject of it. Any SMA (and perhaps the data from which an SMA is created) should generally have been collected directly from the subject with their consent to a particular timeframe and purpose. Clearly this is not always the case, and in the case of generative SMT producing novel SMAs, that may be impossible except by a kind of delegated arrangement.

239. Principle 8 is an important bulwark against reliance on synthetic media to the detriment of an individual in an evaluative process such as a job interview:

> "An agency that holds personal information must not use or disclose that information without taking any steps that are, in the circumstances, reasonable to ensure that the information is accurate, up to date, complete, relevant, and not misleading."

240. Essentially, any decision-maker receiving personal information in audiovisual form about an identifiable individual must take reasonable steps in the circumstances to assess the quality of the

[111] *Dixon v R* [2014] NZCA 329, [2014] 3 NZLR 504.
[112] *Dixon v R* [2015] NZSC 147, [2016] 1 NZLR 678.
[113] David Harvey, "Digital Property - Dixon v R [2015] NZSC 147, [2016] 1 NZLR 678" [2017] NZCLR 195." [2017] New Zealand Criminal Law Review 195.

information. We think that over time, the standard of what is reasonable in the circumstances may change in light of access to superior quality SMAs, but the simplest way to identify a misleading video is simply to put it to the person in question.

241. An individual's ability to challenge the correctness of synthetic media in such circumstances may be limited. When requesting access to personal information, an agency can decline to provide it based on the exception for evaluative material under s 29(1)(b) and (3) of the Act:

> (3) For the purposes of subsection (1)(b) of this section, the term evaluative material means evaluative or opinion material compiled solely—
> (a) For the purpose of determining the suitability, eligibility, or qualifications of the individual to whom the material relates—
> (i) For employment or for appointment to office; or
> (ii) For promotion in employment or office or for continuance in employment or office; or
> (iii) For removal from employment or office; or
> (iv) For the awarding of contracts, awards, scholarships, honours, or other benefits; or
> (b) For the purpose of determining whether any contract, award, scholarship, honour, or benefit should be continued, modified, or cancelled; or
> (c) For the purpose of deciding whether to insure any individual or property or to continue or renew the insurance of any individual or property.

242. In any case where an agency relies on this section to decline access, section 87 of the Act puts the onus of proof on the person declining to provide the information.

243. There is a distinction between factual material on the one hand and evaluative or opinion material on the other, and there can be no "mixed purposes" when it comes to the requirement that material is "compiled solely" for evaluative purposes.

244. Synthetic media artefacts are most likely to be provided as factual evidence, not opinion material, purporting to show the truth of its contents per Condition 1 as a capture of light and/or sound energy that is veridical. Accordingly, we doubt that any SMA given to a decision-maker can justifiably be withheld by that decision-maker if requested under the Privacy Act.

245. The basis for withholding the information is that it would identify the person who supplied it. It is possible that an SMA could be forensically analysed to identify the person who provided it, however we think it is unlikely such a situation would arise and merely note it here for future cases.

246. It is not for us to comment on the adequacy of the regime, only to note that there is existing law that governs any situation where a piece of synthetic media is provided to influence the judgement of a decision-maker about the candidacy of a person per s 29(3).

### Impermissible manipulation and disclosure

247. Information privacy principle 5 requires that agencies protect information by security safeguards that it is reasonable in the circumstances to take, including against "access, use, modification, or disclosure that is not authorised", and "other misuse". We think this principle protects against unauthorised use Category 2 technologies, as well as display, dissemination, and unwarranted statements about reliability and veridicality. However, the limitation is linked to authorisation of the agency, and so once generative authorisation is given, then principle 5 will provide little protection.

248. Principle 11 of the Act limits the disclosure, including publication, of SMA. We have already explained why we think that the "publicly available" exception at principle 11(b) is of limited relevance. We note that there is an additional requirement in 11(b) that "in the circumstances of the case, it would not be unfair or unreasonable to disclose the information" and that this should provide added security against harmful use. This was inserted by the Harmful Digital Communications Act 2015 and accordingly the purpose of that statute will be relevant to interpreting this provision.

## What is the harm?

249. Section 66(1)(b)(iii) states that a breach of a privacy principle will be an interference with privacy if it also "has resulted in, or may result in, significant humiliation, significant loss of dignity, or significant injury to the feelings of that individual." The reference to dignity appears significant, and matches closely the way that the Court of Appeal assessed the harms that varying torts of privacy in *Holland* and *Hosking* were intended to remedy.

250. Section 66 also accounts for causation of loss, detriment, damage or injury to the individual, or adverse effects (or potential adverse effects) on the rights, benefits, privileges, obligations or interests of the individual. The section is drafted in a broad manner in ways that can account for a wide range of harms arising from synthetic media technologies.

## Another authentication tool available to Privacy Commissioner

251. The Privacy Commissioner can also use more orthodox evidential techniques to establish the extent to which a piece of synthetic media is the result of Category 2 manipulation technologies. The Commissioner can seek corroborating evidence in any situation where there has been an allegation that an SMA misrepresents the truth of its contents or has been manipulated in an impermissible manner.

252. Section 91(1) allows the Commissioner to summon and examine on oath any person who is able to give information relevant to an investigation of a privacy act complaint. By doing so, that examination attracts the character of a judicial proceeding and the criminal offence of perjury per s 108 of the Crimes Act. The Commissioner can also compel production of documents or things in the possession of that person relevant to the subject matter of the investigation.

253. The scope of these powers is defined by reference to the Commissioner's opinion, therefore allowing a degree of latitude for the investigation. This power could be used to compel an individual to answer questions about the provenance and manipulation of a digital artefact produced or altered by SMT.

254. Importantly, an examination of that kind would need to acknowledge the varying extent to which digital technologies can manipulate SMAs without necessarily making them untruthful or non-veridical. The Commissioner has existing expertise in dealing in digital information, making the Commissioner an appropriate entity to undertake such investigations.

## Application of the Privacy Act to individuals in connection with personal affairs

255. One significant limitation on the Privacy Act's ability to deal with things like deepfakes, particularly non-consensual pornography, is s 56 of the Act:

> 56 Personal information relating to domestic affairs
> (1) Nothing in the information privacy principles applies in respect of—
> (a) the collection of personal information by an agency that is an individual; or
> (b) personal information that is held by an agency that is an individual,—
> where that personal information is collected or held by that individual solely or principally for the purposes of, or in connection with, that individual's personal, family, or household affairs.
> (2) The exemption in subsection (1) ceases to apply once the personal information concerned is collected, disclosed, or used, if that collection, disclosure, or use would be highly offensive to an ordinary reasonable person.

256. This issue has been avoided by the insertion of sub 2 by the Harmful Digital Communications Act, indicating that it has been statutorily limited in order to prevent its use as a defence against the harms of intimate visual recordings. Notably, the exception applies to collection, disclosure and use, and therefore extends beyond the narrow limitations on dissemination imposed by the HDCA. Again,

we think this indicates a parliamentary intention that harmful generative synthetic media could be located within this statutory framework. This section would appear to anticipate both a "highly offensive use" as well as a "highly offensive creation", which further supports the notion that the Privacy Act is the logical home for generative synthetic media.

257. We note the adoption of the test of "highly offensive to an ordinary reasonable person". On the one hand, this is concerningly broad and may prove difficult for the Commissioner to apply when considering whether a complaint against the way an individual has dealt in personal information is an interference with privacy. But in response, we refer to the analysis of similar tests conducted in two leading cases: *Hosking v Runting* and *C v Holland*. Both of these consider the history of the "highly offensive" to a "reasonable and ordinary person" aspects of the test. They illustrate a justification for its open-ended drafting and some guidance as to the level and nature of harms being considered.

258. In both cases, we think that interpretive questions of this nature must be answered by reference to the essential harm anticipated by privacy doctrines: protection of human autonomy and dignity in the way an individual is presented to the world. We deal with this in greater depth in the next section of our report.

### Conclusion on Privacy Act 1993 (and Privacy Bill)

259. The law should focus on the artefact itself (the Category 3 SMA) when it comes to control of personal information and personal data, not always the content of the video in the sense of what it appears to depict (the human face or voice). Otherwise the use of other SMT of capture, for example digital surveillance cameras, would also be excluded from the principles governing collection of personal information, and the information they disclosed – the face and voice of a person – would be "publicly available".

260. We think that emerging synthetic media will pose significant issues for the Privacy Act. This is because audiovisual information of varying kinds that can be said to be about identifiable individuals will be able to be generated at potentially massive volumes with or without the use of capture devices.

261. One advantage of our suggested approach is that the question of whether a video is falsified will be of secondary importance. As long as, from the perspective of an end-user, it is about an identifiable individual, then it will be caught by the Privacy Act. The accuracy of the video will fall to be dealt with by other principles, and will be relevant to the question of how the video is used in a particular context. That will require close attention to its use and disclosure, and may have a secondary impact on the quantum of damages available.

262. Significant gaps in the Act include the absence of a right to erasure (including in the Privacy Bill), difficulties in ascertaining whether generative SMT constitute a "collection" or a "use" under the Act, and how individuals can limit the scope of their authority to use information for a particular purpose once it has been given.

263. We think that the inclusion of SMA – whether highly veridical or completely untruthful – is consistent with the treatment of appropriating someone's likeness in the history of privacy law and the attention given by Privacy to the dignity and autonomy of an individual to control their presentation to the world around them. We expand on this in the context of two torts discussed next.

# Hosking v Runting and C v Holland

## Relevance of these torts

264. We think that the potential availability of civil torts is unlikely to be of much assistance to victims of harmful uses of SMT. That is because of access to justice barriers not limited to this area of the law. A large number of harms arising from SMT will simply be of a financial level that do not justify the law's intervention through a judicial process of the level involved in order to bring tortious action.

265. Despite that, the decisions by the Court of Appeal in *Hosking v Runting* (then New Zealand's highest Court with judges who later joined the Supreme Court of New Zealand) and High Court in *C v Holland* [2012] NZHC 2155; [2012] 3 NZLR 672 merit relatively extensive analysis. We are conscious that *Hosking* dealt primarily in publication, which sits within Condition 3 of our framework emphasising harms of dissemination. It also analyses harms related to content, whether there was a reasonable expectation of privacy in the circumstances in relation to the facts disseminated. The case was also decided in 2004, the same year that Facebook was founded,[114] and preceded the first iPhone, which was only released in 2007.[115] *Holland* was decided in 2012 and deals primarily with Category 1 technologies. It was accepted that no dissemination of the images in question had occurred and civil action followed conviction under the Crimes Act for intimate visual recording offences.

266. The cases are included here because they demonstrate key propositions which we rely upon for our own conclusions, including in relation to the Privacy Act above. We rely on them for the following points.

    a. There is a long association between the law of privacy and the kinds of harms to human autonomy and dignity associated with increased use of SMT, including misappropriation of someone's image or unauthorised use of their likeness.

    b. Privacy has relevance in New Zealand law beyond the Privacy Act itself, and privacy interests are protected by a range of apparently unrelated statutes as well as international and domestic legal instruments. The Court in both cases accepted privacy was a value protected by the New Zealand Bill of Rights Act.

    c. The fact that a right to privacy cannot be exhaustively defined in all future cases does not mean that it cannot be recognised by the law and developed to fit new technological developments. Privacy can be justiciable even though a complex weighing and balancing of policy factors is required. Privacy is not an absolute right and must be weighed and balanced against other important legal values. When it comes to law restricting dissemination of SMAs, freedom of expression is a significant concern.

267. The decisions are therefore a crucial part of our argument that existing legal mechanisms should be left to govern the use of SMT until a clear gap in the law is identified justifying legislative intervention. Further, it provides support for our conclusion that extreme caution should be taken before suggesting that any such legislation limits fundamental rights in the NZBORA, such as freedom of expression.

268. Much of the analysis in both decisions could be inserted into this report without much amendment and as statements by the Judiciary carry significant weight, however we have done what we can to limit extensive quotation.

[114] See: Wikipedia "Facebook" <https://en.wikipedia.org/wiki/Facebook>.
[115] See: Wikipedia "iPhone (1st generation)" <https://en.wikipedia.org/wiki/IPhone_(1st_generation)>.

## Association between privacy, wrong facts and misappropriation of image

269. A significant issue for us has been whether the idea of appropriating someone's likeness through synthetic media technologies better sits in the context of the law of defamation or privacy.

270. The two do not have to be mutually exclusive, both dealing in very similar policy considerations around freedom of speech and the dignity of an individual in the community. Despite that, on first impression it is not necessarily obvious that privacy could deal in the publication of wrong facts.

271. Privacy is commonly associated with intrusion into a private spatial zone with the result that true but intimate facts are disclosed to the public.[116] The Court of Appeal in *Hosking* attributes recognition of the importance of privacy to the right against unreasonable search and seizure in the NZBORA.[117] Privacy is also closely linked to the action for breach of confidence,[118] and Courts in the United Kingdom have developed the action of breach of confidence rather than recognising a separate tort of privacy.[119]

272. By contrast, the law commonly associates the deliberate publication of false statements about a person injuring their dignity and reputation in the community with the law of defamation: "To the extent that a remedy in damages is awarded arising from publicity given to private information it may be seen as constituting a remedy for damage to reputation which hitherto has been the almost exclusive realm of defamation."[120]

273. These areas of the law collide in the Harmful Digital Communications Act 2015, which includes principles dealing with the disclosure of confidential information as well as the making of false allegations.[121] There is explicit recognition that a digital communication can either be truthful or untruthful so long as it is about an individual who can be identified and suffers harm.[122]

274. It is important therefore to note that the appropriation of someone's likeness has a long association with the law of privacy.[123] Scholz notes that Prosser (a formative scholar on Privacy as an area of law) "only broaches the issue of whether privacy is property in the context of appropriation of likeness,[124] noting: "It seems quite pointless to dispute over whether such a right is to be classified as 'property.' If it is not, it is at least, once it is protected by the law, a right of value upon which the plaintiff can capitalize by selling licenses."[125]

275. Notably for the law of synthetic media in New Zealand, the Court agreed that there is no cause of action in New Zealand law, "directed to unauthorised representation of one's image" (para 171), which should be seen in the context of the appellants relying on "alternative claims for misappropriation of image".

276. At para 66 the Court in Hosking turned to the restatement of the Law of Torts in America,[126] and notes that part of that tort is appropriation of a person's likeness, as drawn in part from Prosser's 1960 article.

---

[116] See, for example, *C v Holland* [2012] NZHC 2155; [2012] 3 NZLR 672.
[117] NZBORA s 21.
[118] *Hosking* at [25] to [26], referring to *Coco v A N Clark (Engineers) Ltd* [1969] RPC 41.
[119] See, for example: *Douglas v Hello! Ltd* [2005] EWCA Civ 595.
[120] *Hosking* at [138].
[121] HDCA s 6 Principles 6 and 7.
[122] Ibid s 4, definition of "posts a digital communication", ss (a)(i).
[123] William L Prosser "Privacy" (1960) 48 Cal LR 383. See also the discussion of 652C in Scholz, L. H. (2016). Privacy as quasi-property. Iowa Law Review, 101(3), 1113-1141, including conceptual confusion about the privacy versus property distinction.
[124] Ibid in "Privacy as quasi-property" at FN 23, citing Prosser (1960) at 423.
[125] Ibid.
[126] Reproduced in *Holland*, above n 109 at para [13].

[66] Causes of action for invasion of privacy have their origins in United States jurisprudence. The Restatement of the Law, Second, Torts 2d (1977) at pp 383 – 394 refers to the general principle relating to the tort of privacy as follows:

> "§652A. General Principle
> (1) One who invades the right of privacy of another is subject to liability for the resulting harm to the interests of the other.
> (2) The right of privacy is invaded by
> (a) unreasonable intrusion upon the seclusion of another, as stated in § 652B; or
> (b) appropriation of the other's name or likeness, as stated in § 652C; or
> (c) unreasonable publicity given to the other's private life, as stated in § 652D; or
> (d) publicity that unreasonably places the other in a false light before the public, as stated in § 652E."

That law has developed with the experience of numerous cases over more than a century. Such experience is of real value, but it must be considered in its full context.

[67] The formulation in the Restatement is adopted from William L Prosser's article entitled "Privacy" (1960) 48 Cal LR 383. In it, Prosser considered the developments in the law since Warren and Brandeis' highly influential article ("The Right to Privacy" (1890) 4 Harvard LR 193), and concluded that the existence of a right of privacy (in fact four separate torts) was recognised in the great majority of the American jurisdictions that had considered the question.

277. In *Hosking* at para 99, the Court identified that §652C was anticipated by the Code of Ethics used by the Advertising Standards Authority at the time under the heading of "Privacy". We cannot locate this now although note similar concerns addressed under current advertising standards elsewhere in this part of the report:

> [199] The second self-regulatory regime which includes within its code of ethics a rule on privacy is that set up by the Advertising Standards Authority:
>
> "10. Privacy – Unless prior permission has been obtained an advertisement should not portray or refer to any persons, whether in a private or public capacity, or refer to any person's property, in a way likely to convey the impression of a genuine endorsement."
>
> The authority's website indicates that this rule is almost never invoked. It can be related to the provision in §652C of the American Law Institute's Restatement on torts making the appropriation of the name and likeness of another one of the privacy torts. Parliament in 1993 expressly deferred to the self-regulatory functions of the authority and its complaints board in ss 8(2) and 21(3) of the Broadcasting Act: the functions of the Broadcasting Standards Authority do not include advertising where the broadcaster and advertiser have accepted the board's jurisdiction.

## Relevant harms: privacy, human autonomy, dignity and SMT

278. One persuasive comment by Tipping J links the notion of privacy to personal autonomy and dignity. We note it here because the loss of control over one's visual or vocal profile – a real prospect with the use of generative synthetic media technologies – is clearly linked to the core of what privacy is intended to protect: human dignity and personal autonomy.

> [239] … It is of the essence of the **dignity and personal autonomy** and wellbeing of all human beings that some aspects of their lives should be able to remain private if they so wish. Even people whose work or the public nature of whose activities make them a form of public property, must be able to protect some aspects of their lives from public scrutiny. Quite apart from moral and ethical issues, one pragmatic reason is that unfair and unnecessary public disclosure of private facts can well affect the physical and mental health and wellbeing of those concerned. Their effectiveness in the public roles they perform can be detrimentally affected to the disadvantage not only of themselves, but of society as a whole.
>
> [emphasis added]

279. This notion of privacy as protection of autonomy and dignity was also a strong feature of the judgment in *Holland*. We think when privacy is seen as the means of autonomy, control, and dignity in one's presentation to the world, there can be no hesitation in ascribing a privacy framing to synthetic media in this way. In *Holland* at [67] it was noted that this has been drawn from international instruments:

> [67] Privacy's normative value cannot be seriously doubted, with various expressions of a right to personal autonomy affirmed in international conventions on human rights,[127] and in various domestic constitutional arrangements and human rights charters.[128] While these domestic instruments do not expressly affirm a general right to privacy, they have been interpreted as protecting rights which are central to autonomy aspects of privacy,[129] ...

280. We do not think that the transition in the next paragraph ([68]) to quotations about information should be seen as material in the context of the wider discussion of privacy in the judgment. The Court analysed the varying policy arguments raised in Hosking about the differing roles of the Courts and Parliament and its discussion cannot be taken too far from its context in acknowledging a tort of intrusion upon seclusion, however we think the Court repeatedly reverts to the terminology of autonomy, and that this is significant, particularly in the same paragraph that the tort itself is acknowledged: [emphasis added]

> [86] … Privacy concerns are undoubtedly increasing with technological advances, including prying technology through, for example, the home computer. The affirmation of a tort is commensurate with the value already placed on privacy and in particular the protection of personal autonomy. ...

281. A similar reference is made at para [95], immediately after the Court articulates the elements of the tort.

282. We are conscious that there is a degree of dispute about the way that both of these torts were recognised and we acknowledge that dispute, as did the judiciary in each case. But, in the context of a report intended to facilitate discussion about the law of synthetic media, we take the law as it is and note that personal autonomy is a significant value when it comes to judicial assessment of the extent to which law should play a role in the ability to control information about oneself. We think this is one of the central harms anticipated by commentators concerned about synthetic media: that people will lose their ability to control information about them, whether public or private. We refer again to the recognition of similar values in s 66 of the Privacy Act.

### Privacy interests are protected by a range of apparently unrelated statutes

283. The Court of Appeal conducted an analysis much like that we have attempted in this report by identifying the extent to which New Zealand law recognises and values privacy in its statute and case law. The wide range of law that it touched upon in its analysis should be an indication to policymakers of the complex policy factors to be weighed and balanced when it comes to limiting freedom of expression in the name of privacy. These policy factors cannot be dictated in advance for all conceivable circumstances. Like the definition of personal information under the Privacy Act, it is preferable to allow the law to develop over time. It will always be a case of balancing competing factors as well as assessment of the facts of each case. As put in *Hosking*:

> [116] The question is how the law should reconcile the competing values. Few would seriously question the desirability of protecting from publication some information on aspects of private lives, and particularly those of children. Few would question the necessity for dissemination of information albeit

[127] Citing International Covenant on Civil and Political Rights, art 17; Universal Declaration of Human Rights, art 12; European Convention on Human Rights, art 8; American Convention on Human Rights, art 11(2).
[128] Citing Canadian Charter of Rights and Freedoms, art 8; United States Constitution, First, Third, Fourth, Fifth and Ninth Amendments; New Zealand Bill of Rights Act 1990, s 21.
[129] Citing David Feldman Civil Liberties and Human Rights in England and Wales (2nd ed, Oxford University Press, Oxford, 2002) at 517-518.

involving information about private lives where matters of high public (especially political) importance are involved. Just as a balance appropriate to contemporary values has been struck in the law as it relates to defamation, trade secrets, censorship and suppression powers in the criminal and family fields, so the competing interests must be accommodated in respect of personal and private information. The approaches adopted by the Privacy Act and in the jurisdiction of the BSA provide informative examples.

284. In particular, the Court in *Hosking* rejected the contention that mere exclusion of a right to privacy from the NZBORA should be seen as determinative of privacy claims:

> [91] The legislative landscape is important. As already mentioned, when enacting the Bill of Rights Act to affirm New Zealand's commitment to the international covenant Parliament did not include among the provisions affirming specific rights and freedoms a provision corresponding to art 17 of the covenant. That provides:
>
> > 1. No one shall be subjected to arbitrary or unlawful interference with his privacy, family, home or correspondence, nor to unlawful attacks on his honour and reputation.
> >
> > 2. Everyone has the right to the protection of the law against such interference or attacks.
>
> Article 8 of the European convention is to similar effect.
>
> [92] We do not accept that omission from the Bill of Rights Act can be taken as legislative rejection of privacy as an internationally recognised fundamental value. It is understandable that, in an enactment focused more on processes than substantive rights, privacy law, which has a very wide scope, would be left for incremental development. The breadth of matters encompassed by privacy had been emphasised by Geoffrey Palmer in his article "Privacy and the Law" [1975] NZLJ 747. Issues of definition, scope of protection and relationship with other societal values clearly would have been such as to defeat any attempt to comprehensively delineate the legal principle.
>
> [93] The White Paper on the proposed Bill of Rights showed that Parliament was concerned not to entrench a vague and uncertain privacy right in the current New Zealand social climate.
>
> [94] As Richardson J said in R v Jefferies at p 302:
>
> > "The nature and significance of a privacy value depends on the circumstances in which it arises. Thus privacy values relied on in search and seizure cases under the Fourth Amendment range from security, to secrecy, to the broad right to be let alone. ... It is not surprising that there is no single readily identifiable value applying in all cases.
>
> [95] The Law Commission's preliminary paper "Protecting Personal Information from Disclosure" (NZLC PP49, February 2002) also highlights the diverse nature of privacy rights in New Zealand. Privacy is seen to include such varying rights as freedom from surveillance (whether by law enforcement or national security agents, stalkers, paparazzi or voyeurs); freedom from physical intrusion into one's body, through various types of searches or drug-testing procedures, or into one's immediate surrounding; control of one's identity; and protection of personal information.
>
> [96] We do not draw from the absence from the Bill of Rights Act of a broad right of privacy any inference against incremental development of the law to protect particular aspects of privacy (or confidence) as may evolve case by case.
>
> [97] It is appropriate to look at legislative provisions that have been enacted to ascertain whether there can be discerned any policy indications in respect of the protection of privacy and whether statutory protections so far enacted amount to a comprehensive treatment.

285. We deal in more depth with the dissenting judgments of Anderson and Keith JJ in our analysis of the NZBORA, however we think that these dissents did not reject the value of privacy itself, merely the ability of the Court to articulate a tort with sufficient precision to justify its inclusion in addition to existing statutory regimes. Of course, it is a legitimate argument to point to the omission of

privacy from the NZBORA as a separate right, however it should not be regarded as conclusive from the perspective of the judiciary in all cases.

286. The judgment of Gault P and Blanchard J had this to say about developing technologies and whether the law of tort was an appropriate method by which to recognise legal restrictions on emerging technologies.

> [3] The law governing liability for causing harm to others necessarily must move to accommodate developments in technology and changes in attitudes, practices and values in society. These are drawn into the law in the main by legislation, often these days to conform with obligations assumed under international treaties and conventions. Such developments, introduced by legislation, emerge from processes which employ extensive consultation and procedures designed to take into account all affected interests.
>
> [4] From time to time, however, there arise in the Courts particular fact situations calling for determination in circumstances in which the current law does not point clearly to an answer. Then the Courts attempt to do justice between the parties in the particular case. In doing so the law may be developed to a degree. It is because the legislative process is inapt to anticipate or respond to every different circumstance that some developments in the law result from such case-by-case decisions. That is the traditional process of the common law.
>
> [5] The Courts are at pains to ensure that any decision extending the law to address a particular case is consistent with general legal principle and with public policy and represents a step that it is appropriate for the Courts to take. In the last respect there are matters that involve significant policy issues that are considered best left for the legislature.

287. The development of new technologies that outstrip legislative protections was also a reason given in *Holland* for development of the common law.[130] We note that the interaction between technological development, legislation and the common law is also a topic of academic discussion.[131]

*288.* *Holland* also drew attention to the role of the right against unreasonable search and seizure as affirming a commitment to privacy, citing judicial decisions in support:

> [25] In public law and criminal contexts, the concept of privacy has been dealt with extensively in the application of the New Zealand Bill of Rights Act 1990. This provides further guidance on the value attached to freedom from intrusion into privacy that might be properly employed in civil legal discourse. While the Bill of Rights Act does not incorporate a general right to privacy,132 s 21 confers a right to be secure against unreasonable search and seizure. Judicial application of s 21 reveals the form, content and weight given to privacy as a legal value. The leading judgment in my view on the concept of intrusion related privacy remains R v Williams, in which the Court of Appeal stated:133
>
> > [48] A touchstone of s 21 of the Bill of Rights is the protection of reasonable expectations of privacy (see R v Fraser [1997] 2 NZLR 442 (CA) at p 449). It is thus only where a person's privacy interest has been breached that his or her rights under s 21 of the Bill of Rights have been breached and a personal remedy is available. The issue therefore is in what circumstances an individual's privacy interest arises.…

[130] Noting at [83] that this was a point made by the Law Commission, and stating it again at [86].

[131] Scholz (above), cites this as follows: "Bruce P. Keller, Condemned to Repeat the Past: The Reemergence of Misappropriation and Other Common Law Theories of Protection for Intellectual Property, 11 HARV. J.L. & TECH. 401, 428 (1998) ("It stands to reason that the faster a technology develops, the more rapidly it will surpass preexisting law, and the more prominent common law theories may become. It is not surprising, therefore, that as the Internet geometrically expands its speed, accessibility, and versatility- thereby vastly increasing the opportunities for economic free-riders to take, copy, and repackage information and information systems for profit-intellectual property owners again must consider the common law as a source of protection at the end of this century, much as it was at the beginning.")."

[132] *Lange v Atkinson* [2000] 3 NZLR 385 (CA) at 396.

[133] Citing *R v Williams* [2007] NZCA 52, [2007] 3 NZLR 207.

*289.* Both decisions also drew on the Privacy Act, the Broadcasting Act and the Residential Tenancies Act. *Holland* drew on s 216H of the Crimes Act, where the defendant had already been convicted.

*290.* The Court in *Hosking* drew on the Broadcasting Act as an indication of the extent to which Parliament has left the development of privacy protections to a specialist body on a case-by-case basis:

> [85] Also relevant in the New Zealand context is the growing body of decisions of the Broadcasting Standards Authority (the BSA). Without creating a civil cause of action, s 4(1)(c) of the Broadcasting Act 1989 provides that broadcasters are responsible for maintaining standards consistent with, inter alia, the privacy of the individual. The BSA is obliged by s 21 of the Act to ensure that broadcasters comply with s 4. To this end the BSA has adopted privacy principles which will be referred to. Eichelbaum CJ accepted in TV3 Network Services Ltd v Broadcasting Standards Authority [1995] 2 NZLR 720 (HC) that the BSA was entitled to draw on United States case law in developing the privacy principles, particularly given the relative paucity of experience in this field of the New Zealand judiciary. As a result, the BSA jurisprudence is derived from the same foundation as the existing High Court authorities on breach of privacy.
>
> [86] The BSA decisions demonstrate that privacy interests do not exist in a vacuum. The facts and context of each case have determined its outcome. These decisions show that protection of private information is workable. An expert authority, experienced in media issues, must be taken as giving useful guidance. Indeed in Britain the Human Rights Act requires professional codes to be taken into account. The BSA has dealt in the New Zealand context with numerous issues likely to come before the Courts whether as matters of privacy or confidence. For example, in Re McAllister [1990] NZAR 324 the BSA commented that on a public street or in any other public place, the plaintiff has no legal right to be let alone, and it is no invasion of privacy to follow him about and watch him there, nor to take a photograph of him. Such an action amounts to nothing more than making a record not essentially different from a full written description of a public site which anyone would be free to see.

## A right to privacy can be open-ended and flexible

291. We think it is important to note the Court's relative comfort with the idea that restrictions on freedom of speech through the use of open-ended torts and statute law are permissible. We think this is important in response to the idea that any intervention whatsoever in the use of synthetic media technologies is an impermissible slippery slope towards absolute censorship. It is, however, predicated on application by the judiciary as a specialist arm of the state, which is fundamentally different to, and in many ways more desirable than the censorship processes adopted by private dissemination platforms. As put in *Hosking*, and relevant to the interpretation of similar standards in the Privacy Act:

> [250] Nor do I think that when the concepts are carefully examined, there is much force in the criticism that the new tort is so uncertain that it should never be born. The plaintiff must show first an expectation of privacy and, more importantly in most cases, that such expectation is a reasonable one. The latter dimension of reasonableness, familiar in many fields of law, controls the subjective expectation of the individual. It introduces an objective element upon which, as with all questions of reasonableness, in the end the Court has to make a value judgment. It is a very familiar exercise and cannot, in my view, validly be criticised on the basis of uncertainty. The concept is clear. The fact that its application in a marginal case may be difficult is not a valid reason to regard the concept as possessing objectionable uncertainty. Expectations of privacy are really no more uncertain or elusive than expectations of confidence; or the expectation that reasonable care will be taken not to damage the interests of others. The parameters of any general duty are constantly being worked out and refined by the Courts. An underpinning jurisprudence can be allowed to develop for privacy as it has for confidence and negligence. What expectations of privacy are reasonable will be a reflection of contemporary societal values and the content of the law will in this respect be capable of accommodating changes in those values.

292. Gault P and Blanchard J in *Hosking* agreed in substance:

> [118] No Court can prescribe all the boundaries of a cause of action in a single decision, nor would such an approach be desirable. The cause of action will evolve through future decisions as Courts assess the nature and impact of particular circumstances. ...

293. The Court in *Holland* made the point forcefully as well:

> [88] Functionally also, the role assumed by Parliament in protecting privacy interests has focused on controlling the collection and dissemination of private information, or at the other end of the spectrum, criminal culpability and the control of state power, including most recently surveillance powers. The reticence of Parliament to wade into the realm of civil claims in the years since Hosking is a matter of conjecture, though the Law Commission report provides several reasons why that might be so including the potential breadth of such a statutory tort. But it is the function of the Courts to hear and determine claims by litigants seeking to vindicate alleged rights or correct alleged wrongs. … this is a case crying out for an answer, and given the value attached to privacy, providing an answer is in my view concordant with the historic function of this Court.

294. It is important to be conscious that the Court in these cases was discussing a tort which would in all cases fall for determination by a member of the judiciary. There may be other factors to take into account when it comes to other constitutional actors (Crown Entities as executive actors) or private actors enforcing similar standards. There is ample force in the points raised by the dissenting judges in *Hosking*: the law can be flexible but not so vague as to make it difficult for conscientious actors to know how to act. Our answer to this is relatively simple: the standards we propose should be adopted with regard to synthetic media are those that are already on the statute books. They have Parliamentary input through democratic processes and are already applied by dedicated institutions with natural justice and rule of law processes to limit their potentially harmful effects.

**Reasonable expectation of privacy, highly offensive to reasonable and ordinary person**

295. A common thread to both torts is the inclusion of elements that limit the scope of the tort of privacy to facts in which there is a reasonable expectation of privacy, and where publication or intrusion would be highly offensive to the reasonable and ordinary person. The Court in *Hosking* found on the facts that there could be no such reasonable expectation and that publication (in *Hosking*) could not be highly offensive:

> [164] The inclusion of the photographs ... would not publicise any fact in respect of which there could be a reasonable expectation of privacy. The photographs taken by the first respondent do not disclose anything more than could have been observed by any member of the public in Newmarket on that particular day. They do not show where the children live, or disclose any information that might be useful to someone with ill intent. The existence of the twins, their age and the fact that their parents are separated are already matters of public record. There is a considerable line of cases in the United States establishing that generally there is no right to privacy when a person is photographed on a public street. Cases such as Peck and perhaps Campbell qualify this to some extent, so that in exceptional cases a person might be entitled to restrain additional publicity being given to the fact that they were present on the street in particular circumstances. That is not, however, this case.
>
> [165] We are not convinced a person of ordinary sensibilities would find the publication of these photographs highly offensive or objectionable even bearing in mind that young children are involved. One of the photographs depicts a relatively detailed image of the twins' faces. However, it is not sufficient that the circumstances of the photography were considered intrusive by the subject (even if that were the case, which it is not here because Mrs Hosking was not even aware the photographs had been taken). The real issue is whether publicising the content of the photographs (or the "fact" that is being given publicity) would be offensive to the ordinary person. We cannot see any real harm in it.

296. These paragraphs illustrate the relevance of the framework, noting a distinction between harms of creation, content and dissemination.

297. The Court in *Hosking* expressed concerns about the breadth of a right to privacy in a public place in relation to a case from Quebec. The Court's analysis noted the fundamentally different framing and consequences of grounding a right to privacy in a civil code legal system rather than a common law

legal system like that of New Zealand, and illustrates the issues undermining our conclusion about the usefulness of personality and publicity rights in New Zealand: [bold emphasis added]

> [62] Quebec has gone further than the federal government towards protecting privacy, enacting s 5 of the Quebec Charter of Human Rights and Freedoms which guarantees every person "a right to respect for his private life". In *Les Éditions Vice-Versa Inc v Aubry* (1998) 157 DLR (4th) 577, a photographer took a picture of the respondent without her knowledge as she sat on a Montreal street. The photograph was subsequently published in an artistic magazine. An award of damages for breach of s 5 was upheld by the majority in the Supreme Court, who considered at p 594 that the purpose of s 5 is to protect a sphere of individual autonomy. To that end, the right to one's image must be included in the right to respect for one's private life, since it relates to the ability of a person to control his or her identity. The right to respect for private life is infringed as soon as an image is published without consent, provided the person is identified. It is irrelevant to the question of breach whether the image is in any way reprehensible, or has injured the person's reputation.
>
> [63] The Court in *Aubry* recognised, however, that expectations of privacy may be less in certain circumstances. This will often be the case if a plaintiff is engaged in a public activity where the public interest in receiving the information should take priority. The right to a private life may also be less significant where the plaintiff appeared only incidentally in a photograph of a public place, or as part of a group of persons.
>
> [64] The *Aubry* case is based on a specific provision of the Quebec charter. Quebec is a civil law jurisdiction with close ties to the law of France (where a right to privacy has long been included in the civil code). Supreme Court decisions on appeal from Quebec have no binding effect on the common law provinces. In *Hung v Gardiner* [2002] BCSC 1234 the Supreme Court of British Columbia declined to follow *Aubry*, on the grounds that it was a decision from Quebec. **The charter provision creates, in effect, a right of property in one's image. It cannot provide the foundation for such a right in New Zealand.**
>
> [Emphasis added]

298. We have considered how the "reasonable expectation" and "highly offensive" tests should be approached in the context of SMT (also noting that digital photography itself is embraced by our framework as such a technology because of the inherent role of digital manipulation technologies).

### Reasonable expectation of privacy and privacy in public

299. The reasonable expectation of privacy is founded in part in the notion of public and private zones. The notion of a public / private divide is being reconsidered by some scholars as untenable and a poor basis for public policy, one which should be abandoned in favour of an intentional and value-oriented policy process about what information we do or do not wish to assert control over. We prefer such a framing, that centres on individual autonomy to control one's presentation to the world at large rather than relying on the notion of a spatial or technological private zone and we note the transition in New Zealand law towards the right against search and seizure as applying to personal expectations of privacy as much as entry on to private property. Once privacy ceases to be seen exclusively in a spatial sense about true facts, as we think is the case in *Hosking* and *Holland* and the Privacy Act (particularly principles 4(b) and 11(b) in their reference to reasonableness and fairness), these tests should take a lesser importance to wider policy debates. In particular, we echo the work of, Woodrow Hartzog, who concludes by writing:[134]

> The "no privacy in public" argument has, thus far, put the cart before the horse. Before lawmakers and society can answer the question of whether privacy can exist in public, we must first understand what the concept of "public" means. As I have demonstrated in this Article, "public" can be conceptualized several different ways, from descriptive to designated. These conceptualizations are at best under-theorized and at worst tautological. This means that the term must be given a more articulated meaning to be useful in law and policy. Most importantly, law and society must recognize that to label something as "public" is both consequential and value-laden. We must reject a neutral, empirical notion

[134] Woodrow Hartzog "the Public Information Fallacy" (2019) 99 Boston University Law Review 459.

> of "public" that is separate from legal and social construction. There is no such thing. How we define public information sets the rules for surveillance and data practices, so we should proceed intentionally and with caution. We should be more critical of claims like "data is public" to justify surveillance and data practices. To move forward, we should focus on the values we want to serve, the relationships and outcomes we want to foster, and the problems we want to avoid.

300. When it comes to synthetic media technologies, we think that the question of a "reasonable expectation of privacy" will need to be determined by very close reference to condition 1 of our framework: the extent to which, a capture technology is deployed and the relationship between that capture and the SMA produced. There is a significant difference to be drawn between a photo of someone in public using capture technologies, and a heavily manipulated SMA that is essentially a novel artefact. Although there is a significant difference, it is difficult to articulate a boundary standard to explain that difference for all future circumstances. The use of a capture technology in a public place leading to a relatively reliable image of what that person looked like in public is one thing (although there are situations where that too will be impermissible – as in the case of children, for example). That must be contrasted with the creation of a photo of a recognisable person doing something that simply never took place, where there is little relationship at all between the SMA and any capture process per Condition 1.

301. This leads to two points to be made about synthetic media where the role of Category 1 capture technologies is limited, specifically deepfakes and synthetic media that allows the generation of representations of events that never took place.

302. The first point is that an SMA should initially be assessed by the law at face value in terms of the law of privacy. The capture and creation process may be relevant to assessing the truth or provenance of the SMA. But ultimately, it will be consumed by the ordinary observer on its face. We do not think there is any merit to going behind the process of creating the image unless there is an explicit requirement to do so. For example, if a celebrity's face is transferred into a pornographic video, there should be no suggestion that the person's face, or the source data for the person's face, is "public", and therefore they can have no reasonable expectation of privacy against it being synthesised into pornography. Setting aside pornographic scenarios, what if anything is to be done about SMA that simply shows somebody walking down the street, in the same way discussed in *Hosking v Runting*? For that we turn to our second point.

303. When taking the SMA at face value, and without reference to its generative techniques, the scene depicted never took place even though some veridical artefacts may have been used in its creation. The scenario is fictional even though it appears real. That is so even where a representation of a person's face is highly persuasive. Accordingly, an SMA is entirely different from the factual scenarios discussed in *Hosking v Runting*. It is in no way a record of a set of events that occurred at a particular place and time. In that sense, it can never have been publicly available. It is not public in any sense, and there can be no suggestion that a person has no reasonable expectation of privacy in it because of its public nature.

### Highly offensive to a reasonable and ordinary person

304. There is another factual question that sits at the heart of the privacy tort: whether publication of the facts would be highly offensive to a reasonable and ordinary person. Again, we think it is wrong to separate the synthetic media artefact into its constituent pieces when considering this question. It is no answer to say that because the data of one's face and voice have been used to produce the SMA, and because one's voice and face are public, a person therefore has no basis to object to the publication of synthetic representations of their face or voice in a public setting. The Court is not assessing the disclosure of the sound or appearance of someone's face or voice. The Court is assessing the nature of the representation disclosed publicly. We suggest that, given the significant body of law that distinguishes between harmful and harmless deception and misrepresentation, it is highly likely that the persuasive and realistic representation of someone saying or doing something that never took place would be highly offensive to a reasonable and ordinary person.

This will be a question of fact and degree in the circumstances, including by examining the particular harms alleged to have been caused. It will take account not only the content of the SMA itself, but contextual factors.

**Conclusion on *Hosking* and *Holland*, reasonable expectation, offensiveness to reasonable and ordinary persons**

305. Synthetic media technology is unparalleled in its ability to produce highly persuasive photorealistic representations of real people. Those representations are, for practical purposes, almost infinitely malleable when it comes for capacity to cause harms that are already recognised throughout New Zealand law. Their potential impacts can therefore be distinguished from other forms of simple capture technology – such as cameras or microphones – on the basis that their capacity for malicious or reckless use is effectively limitless.

306. Any audiovisual representation can also be presented in an artificial context so as to generate harm, including by mis-describing it, or similar "shallow fake" approaches. But that is necessarily limited by what is recognisable within the audiovisual artefact itself: no amount of persuasive captioning can convince a reasonable person that a photo of something recognisable is in fact a photo of something similarly recognisable, within an acknowledged degree of visual ambiguity. Synthetic media takes this capacity significantly further by allowing an individual to entirely synthesise the events shown in that audiovisual representation, and to such a high degree of photo or audio realism that a reasonable observer would believe that the virtual representation must evidence some event taking place in the real world that was "captured" on camera or microphone.

307. On that basis, we think an argument can be made that there is no such thing as an unreasonable expectation of privacy when it comes to the generation of a capacity to produce synthetic media artefacts of a real person. In short, people will always have a reasonable expectation that no person captures or develops, without their consent, the capacity to generate their audiovisual profile through synthetic media technologies in a way that impermissibly undermines their dignity and autonomy. This is a matter for public discussion and debate, or a fact finder on the facts of a given case.

308. We also note that:

    a. there are a range of statutes protecting privacy interests in New Zealand, including the right against unreasonable search and seizure in the NZBORA.

    b. the law of privacy in tort and the Privacy Act are open ended and allow for technological development.

    c. the law of privacy requires a balancing of factors and interests, and these can be done pursuant to a sufficiently rigorous process by judicial officers, pursuant to the common law or legislation.

    d. we doubt the value of personality or property rights to add to the law of privacy in New Zealand.

    e. the idea of a reasonable expectation of privacy and whether publication or intrusion is highly offensive to a reasonable and ordinary person needs to be considered by close examination to our framework, particularly where the creation of an SMA is not reliant on a veridical capture process in terms of category 1 and condition 1 of the framework.

    f. the notion of public or private zones is of lesser value to an overall focus on the autonomy and dignity of an individual pursuant to the right to respect for privacy.

## Conclusion on SMT and Privacy in New Zealand

309. We think that privacy law is the appropriate framing for many of the harmful uses of SMT in New Zealand that deal in appropriation of an identifiable individual's likeness.

310. The idea that a person's face or voice is fundamentally public carries very little weight when it comes to the dealing in SMA themselves as digital artefacts. This is especially the case where those artefacts are generative, such that they can create new personal information that is highly persuasive without recourse at all to the subject of the SMA.

311. There is limited value to a property framing when it comes to an individual's ability to assert control over their audiovisual profile. To the extent there are any gaps in the law of privacy, we think these are dealt with through other regimes, particularly the Fair Trading Act, discussed later in this report.

312. Generally speaking, there is no need to examine the technique behind a Category 3 SMA and consider how it has been captured or synthesised – unless the relevant legal regime calls for attention to be given to that as a harmful process. We think that privacy law adequately captures the various technologies defined by Categories 1 to 3 of our framework and that privacy law in particular is well equipped to consider how harms can arise from Conditions 2 and 3.

313. Privacy is well-equipped to deal with the truth, accuracy or veridical properties of information per Condition 1, even though this is not its core focus. It prioritises an individual's ability to autonomously assert control over their presentation to the community and the way that information about them is collected, used and disclosed, often regardless of its accuracy. In this sense, the accuracy of that information is immaterial unless it is being relied upon to make decisions about the individual or misrepresent them. In that case, the Privacy Act in particular takes a strong focus on the right to control that information (as in privacy principle 8) or assess the impact of breaches of privacy on that individual's distress and dignity. In any event, the right to assert control over that information is given special focus under privacy principles 6 and 7.

# Restrictions on Freedom of Expression

## New Zealand Bill of Rights Act 1990

### Freedom of expression

314. Section 14 of the New Zealand Bill of Rights Act 1990 states:

> 14 Freedom of expression
> Everyone has the right to freedom of expression, including the freedom to seek, receive, and impart information and opinions of any kind in any form.

315. Audio-visual information can be thought of as an expression of fact because of assumptions about the category 1 capture process, explained through condition 1 of the framework. Instead, synthetic media must instead be seen as a piece of evidence. Audio-visual information does not prove the truth of its contents. It merely shows what appears to have been captured by a sensor at a particular point in time.

316. When coupled with tools such as editing or statements about context and content, audio-visual information can also be a statement of opinion. The capacity to use audio-visual information as a statement of opinion about the world is about to drastically increase. Now, highly photo- and phono-realistic audio-visual information should be seen as akin to political cartoons: a highly accessible and effective means of communicating complex ideas to a wider audience at a faster pace than could be done by the written word alone.

317. We refer to the work of Tom Sainsbury, a comedian in New Zealand who uses a face-swapping application (a Category 1-3 technology) and dissemination technologies – snapchat and Facebook – to articulate satirical content about prominent New Zealand politicians, as well as fictional everyday New Zealanders. There is a real risk that content such as his could be restricted or chilled by the suggestion that legislation should be passed to limit the use of synthetic media technologies.

318. Fundamentally, SMA and SMT should be seen as forms of human expression. Much of the discussion around the malicious use of deepfakes has characterised SMT as tools of civil harm whereby a bad actor harms a victim. It would be wrong to limit analysis of synthetic media to this axis, although it is important.

319. When people call for legislation to limit individuals' ability to use synthetic media technologies to communicate, they are calling for legislation by the State that would enable State actors to limit individual freedom of expression. We think this has been drastically under-examined in the context of synthetic media. People calling for regulation of deepfakes are calling for censorship, which must always be carefully examined and conducted according to law in a transparent way.

320. We note that this is a significant concern that has been acknowledged in the HDCA.

321. We have significant concerns about the notion that large social media platforms who specialise in technologies of dissemination should be censoring content. We note an emerging area of legal research into the extent to which the power of social media platforms conforms to rule of law principles.

322. We strongly support the work of WITNESS, and the approach taken by Blackbird AI to misinformation. Blackbird articulate a distinction between two censorship approaches:

    a. The first absolutely limits the communication of and individual exposure to misinformation. Approaches such as these idealise a perfect technological solution. It assumes technology will allow us to identify synthetic media artefacts that were altered and distinguish permissible from impermissible alterations. It also assumes we could agree on that standard of "permissibility"

pursuant to a democratic process. Even if that were possible, we are not confident that it should be deployed to absolutely prevent access to such content. The potential for abuse of such technology is significant. The Films, Videos and Publications Classification Act performs this function and it sets a very high bar based on impermissible content, not the degree of alteration employed.

b. The second approach emphasises the provision of greater contextual information to consumers. This is the approach we understand to be taken by Blackbird. It relies on a higher level of critical engagement by individuals and communities but aims to facilitate decision making by deferring to individual autonomy, rather than censorship.

323. Another helpful distinction was articulated to us by Synthesia, a company focussed on the way that synthetic media technologies can be used to enhance human connection across language barriers. We understand Synthesia to articulate a distinction between two approaches.

a. "Forensics": all audiovisual content is assumed to be "real", and through various techniques, "false" audiovisual content can be identified. Synthesia have expressed confidence that technological solutions can perform this function through the use of machine learning techniques. They refer to Face Forensics ++ as an example of this. We also note that the Rochester Institute of Technology was awarded funding through the AI and the News Open Challenge to pursue similar technological forensic approaches..

b. "Verification": all audiovisual content is assumed to be false or unreliable. Only certain audiovisual artefacts are taken to be reliable based on the inclusion of a watermark or indicator of reliability, including blockchain and cryptographic solutions. Solutions of this nature are seen to be more difficult to achieve: for example, the uploading of a video to a social media platform fundamentally changes its digital character such that it would be recorded as having been "manipulated" even where the content of it may not be deceptive in terms of condition 1. We also think that these will have an exclusionary effect.

324. We find it difficult to see how to justify an approach articulated by some that live-streaming, for example, should be restricted in an absolute sense. We struggle to see how a meaningful difference between live-streaming and, for example, video-conferencing can be maintained.

325. We also note the unforeseeable capacity for transparency and accountability in the public use of power that can be achieved through the use of live-streaming technology. A New Zealand artist, Luke Willis Thompson, worked with the family of Philando Castile, an unarmed victim of a police shooting, on a work which was subsequently nominated for the Turner Prize. He is the second New Zealander ever to be nominated. As described by Metro Magazine:

> In July 2016, Reynolds was travelling with her partner Philando Castile and her daughter in their car in St Paul, Minnesota. They were pulled over, and Castile, in the driver's seat, was shot several times by a police officer. Reynolds, armed with her phone, live-streamed the immediate aftermath on Facebook: a video that has now been viewed online more than nine million times. "What I saw in that video was a performative brilliance that works on a jurisprudence level," Thompson says of Reynolds' decision to stream the events. "It's changed the way we think about witnessing and image production."

326. Like WITNESS, we think it is fundamental to consider how responses to synthetic media technologies will affect both human rights and people's ability to document human rights abuses. There is also the potential for an exclusionary effect: if technological solutions are incorporated in ways that are only accessible to certain sectors of society, then that will disempower and delegitimise the voices of others without access to those technologies.

327. Further, we think that the ability to live-stream synthetic media is one of the simplest ways to limit the probability that it has been deceptively altered. Synthetic media technologies can work deceptively in real time, however the likelihood that someone has been able to deploy these technologies in such a short space of time reduces the probability that sophisticated manipulations

have been deployed. That risk is reduced even further where multiple live stream can corroborate the existence of a sequence of events from different perspectives in real time. For this reason, we also think it is vital that consumers be fully informed about the kinds of manipulations that are being deployed in live-streaming or live audio-visual synthetic media technologies. For example, we should be fully informed about the extent to which the use of category 2 technologies in video-conferencing may be affecting the relationship between the light and sound that is captured by Category 1 technologies is being manipulated, per Condition 1.

### Justifiable limitations on freedom of expression

328. We also accept it would be wrong to treat the right of freedom of expression as absolute.

329. We note the discussion in *Hosking v Runting*, and the work of Petra Butler,[135] in articulating how and why the right to privacy is, is not, or should be reflected in the NZBORA. The right against unreasonable search and seizure is commonly linked to a right of personal privacy, not just spatial privacy, and we include it here for that purpose:

> 21 Unreasonable search and seizure
> Everyone has the right to be secure against unreasonable search or seizure, whether of the person, property, or correspondence or otherwise.

330. The application of the NZBORA in this context is a subject of expert research and analysis beyond the scope of this report, but we think it would be incomplete without noting the extent to which freedom of expression can conflict with the right to privacy as drawn from various sources (including the NZBORA) by the Court in *Hosking v Runting*. We cannot articulate it any better than the Court itself, and include the following passages at paras 333-334 of the report in order to illustrate the way in which these rights and values can conflict.

331. These passages are one example of the kind of sophisticated balancing exercise required when giving effect to two kinds of freedoms under the NZBORA: freedom of expression and the right to respect for privacy (as drawn from the right against unreasonable search and seizure). Freedom of expression is seen as particularly important by the Court on the facts of *Hosking*, which effectively relates to an individual's ability to suppress publication of information about themselves on the basis of respect for their privacy.

332. The remainder of part 3 of the report draws attention to statutes where Parliament has seen fit to limit freedom of expression in particular circumstances, for example in cases of Criminal deception or objectionable material that is injurious to the public good.

333. In *Hosking v Runting*, Anderson J noted the emotive way in which "invasion" of privacy can be seen and we think that point applies with equal force to the way that the products of emerging audiovisual technologies, and deepfakes in particular, have been treated in public discussion. His Honour's reference to the various ways that the law protects privacy values through specific prohibitions is also an approach which we respect, and the basis for much of our analysis of criminal statutes as providing specific answers to many of the problematic uses of deepfakes or SMT, in opposition to the development of a 'Deepfake Act' of some nature. We also include this reference to draw attention to the way that the extent to which privacy as a value can or should be given weight over other competing values is a matter of debate, even among New Zealand's senior judiciary:[136]

> [263] In my respectful view, the emergence of an "invasion of privacy" tort has gained impetus from semantic imprecision and questionable analysis of the relationship between rights and values. The term "invasion" has broad, emotional connotations which can tend to obscure the true nature of the question being examined in this case. It may be an appropriate term for those encroachments on

[135] See, for example, above n 110.
[136] Above n 105 at [263] to [268].

personal autonomy which involve trespass and eavesdropping, but this case is not about invasion even in a metaphorical sense. It is about publication.

[264] What is meant by "privacy" and what is the nature of a right to it? In a strict sense "privacy" is a state of personal exclusion from involvement with or the attention of others. More important than its definition is the natural human desire to maintain privacy. Only a hermit or an eccentric wishes to be utterly separated from human society. The ordinary person wishes to exercise choice in respect of the incidence and degree of social isolation or interaction. Because the existence of such a choice is a fundamental human aspiration it is recognised as a human value. The issue raised in this case is the extent to which the law does, and the common law may, give effect to that aspiration.

[265] The extent to which that human value is also a right is described by the multitude of legal, equitable and administrative remedies and responses for derogation of the value which Keith J has identified in his judgment. The small residue of the concepts with which cases such as the present are concerned has not been a right at all but an aspect of a value. An analysis which treats that value as if it were a right and the s 14 of the NZBORA right as if it were a value, or treats both as if they were only values when one is more than that is, I think, erroneous.

[266] Thus, cases such as the present are not about invasion but publication; and they are not about competing values, but whether an affirmed right is to be limited by a particular manifestation of a value.

[267] Having regard to s 5 of the NZBORA there should be no extension of civil liability for publication of true information unless such a liability is a reasonable limitation which is demonstrably justified in a free and democratic society. Freedom of expression is the first and last trench in the protection of liberty. All of the rights affirmed by the NZBORA are protected by that particular right. Just as truth is the first casualty of war, so suppression of truth is the first objective of the despot. In my view, the development of modern communications media, including for example the worldwide web, has given historically unprecedented exposure of and accountability for injustices, undemocratic practices and the despoliation of human rights. A new limitation on freedom of expression requires, in my respectful view, greater justification than that a reasonable person would be wounded in their feelings by the publication of true information of a personal nature which does not have the quality of legally recognised confidentiality.

[268] Nor is there any demonstrable need for an extension of civil liability. Peeping, peering, eavesdropping, trespassing, defaming, breaking or exploiting confidences, publishing matters unfairly, are already covered by the legislative array. What is left to justify the breach of the right to freedom of expression?

334. As an indication of the opposing view, we include the analysis of Tipping J, which articulates the way that rights are not absolute and can be balanced, even in ways that are iterative and allowed to develop over time:

[224] In the privacy field, as in many other fields of law, the Courts are engaged in reconciling competing values. First, there is the value to society of the right to freedom of expression which is expressly recognised by s 14 of the Bill of Rights. But the Courts should also recognise and give appropriate effect to the values involved in the broad concept of privacy. Those values are also important in our society and hence are recognised in our international commitments. They are recognised less directly, but no less significantly, in provisions such as s 21 of the Bill of Rights, namely the right to be free from unreasonable search and seizure. That right is not very far from an entitlement to be free from unreasonable intrusions into personal privacy. Indeed s 21 speaks of unreasonable search or seizure, whether of the person, property, correspondence "or otherwise". Those last two words signal the breadth of reach which s 21 was intended to have.

[225] Rishworth, Huscroft, Optican and Mahoney, in their recent 2003 publication New Zealand Bill of Rights (to which I will refer simply as "Rishworth"), discuss at pp 419 – 420 the strong privacy rationale of s 21 of the Bill of Rights. They refer to *R v Jefferies* [1994] 1 NZLR 290 (CA), and regard the decision of the Court in that case as having focused on the importance of s 21 for defending "those values or interests which make up the concept of privacy", as Thomas J put it at p 319. Rishworth then states that in so holding, the Court of Appeal followed the lead of what the authors describe as the revolutionary judgment of the United States Supreme Court in *Katz v United States* 389 US 347 (1967). In that case the Court held that the right to be free from unreasonable search or seizure contained in the Fourth Amendment to the US Constitution protected people not places. Rishworth states that this

simple observation modified centuries of common law thinking and established privacy, not property, as the core value guarded by the US constitutional requirement of reasonable search. Indeed Harlan J, who wrote a concurring judgment in *Katz*, indicated that the search and seizure jurisprudence should be triggered whenever the activity in question invaded a "reasonable expectation of privacy".

[226] It is not necessary for present purposes to discuss the scope of the concepts of search and seizure in the light of the privacy rationale. At least at first blush it would seem very strained to view photographs as a form of seizure, or indeed search; and, in any event, seizing the image of a person who is in a public place could hardly be regarded as unreasonable, unless there was some very unusual dimension in the case. My present point is that the values that underpin s 21 and which are reinforced by New Zealand's international obligations can, by reasonable analogy, be extended to unreasonable intrusions into personal privacy which may not strictly amount to search or seizure. The lack of any express recognition of a right to privacy in the Bill of Rights should not, in my view, inhibit common law developments found to be appropriate. Society has developed rapidly in the period of nearly 15 years since the enactment of the Bill of Rights in 1990. Issues and problems which have arisen, or come into sharper focus, as a result of this development should, as always, be addressed by the traditional common law method in the absence of any precluding legislation.

[227] The same can be said of the fact that in 1993 Parliament enacted the Privacy Act. I do not regard the ground as having been entirely captured by that enactment so as to preclude common law developments. Indeed it might well seem very strange to those who see the Privacy Act as preventing the supply of information about whether a friend is in hospital or on a particular flight, for the common law to be powerless to remedy much more serious invasions of privacy than these would be. In the absence of any express statement that the Privacy Act was designed to cover the whole field, Parliament can hardly have meant to stifle the ordinary function of the common law, which is to respond to issues presented to the Court in what is considered to be the most appropriate way and by developing or modifying the law if and to the extent necessary.

[229] The Bill of Rights is designed to operate as between citizen and state. Nevertheless it will often be appropriate for the values which are recognised in that context to inform the development of the common law in its function of regulating relationships between citizen and citizen. The judicial branch of government must give appropriate weight to the rights affirmed in the Bill of Rights when undertaking that exercise.

[230] Freedom of expression must accommodate other values which society regards as important. That accommodation must be carefully worked out, as it has been over many years in the law of defamation which protects personal reputation, a value which is also not expressly recognised in the Bill of Rights. When deciding whether, and if so how, to develop or mould the common law to achieve such an accommodation, the Courts must do their best to strike the right balance between the competing values. In fields like the present this necessarily includes considering whether the limit on a right affirmed by the Bill of Rights such as freedom of expression, which the proposed common law development would create, is both reasonable and demonstrably justified in a free and democratic society.

[231] It is not, however, enough for those who are asked to accept some limit on freedom of expression simply to rely on s 14 of the Bill of Rights as if it were some universal social panacea which must be seen as trumping other rights and values in most, if not all circumstances. ... It would not be in society's interests to allow freedom of expression to become a licence irresponsibly to ignore or discount other rights and values.

[232] But against that, the importance to society of the values enshrined in the right to freedom of expression suggests that the Courts should allow those who invoke that right appropriate latitude in what they say and publish. It is not for the Courts to apply controls which are too exacting in their reach or content. In short, all limitations on freedom of expression must be reasonable and demonstrably justified. They must also, of course be "prescribed by law", a matter to which I will return.

...

[236] In the end someone has to make a judgment on behalf of society as to where the balance falls. The question may often be whether individual harm outweighs public good. The responsibility for striking the right balance is vested in the Courts. In discharging that responsibility it is perfectly appropriate for the judicial branch of government to determine, after hearing argument on all sides,

that an appropriately formulated free-standing tort of privacy should exist; but subject to a defence designed to protect freedom of expression values when the privacy values which the tort is designed to protect fail to outweigh them.

[237] The weight one gives to privacy values in concrete terms is no doubt a matter of assessment in the individual case. But I do not consider there can be any room to doubt that, on appropriately defined occasions, privacy values can outweigh the right to freedom of expression. There is obviously room for differences of view as to how these occasions should be defined but that is a different matter. When privacy values are found to outweigh the right to freedom of expression, and the law recognises that by placing a limitation on freedom of expression, that limitation will, in terms of s 5 of the Bill of Rights, be a limit prescribed by law. It will also be a limit which is reasonable and demonstrably justified in a free and democratic society.

...

[253] I immediately accept that a principle or rule which is enunciated in a wholly uncertain manner could well be a principle or rule which is not sufficiently prescribed by law for the purposes of s 5. What I cannot accept is that incremental common law or equitable developments, or reshapings of the law; or principles which are stated at a higher level of generality than may be the European method, should be regarded in New Zealand as not sufficiently prescribed by law. It is inevitable of course that questions of degree will arise. But I do not consider the phrase "prescribed by law" in s 5 was intended or should be construed so as to stultify traditional common law methodology and prevent Courts from implementing legal developments which they regard as appropriate and necessary, on the premise that the obvious and unavoidable uncertainty that often exists at the margins in some fact situations should prevent an otherwise appropriate development.

## Summary of approach to NZBORA

335. These legal analyses by Justices Tipping and Anderson appear to stake out the ground in policy terms about the conflict between the right to freedom of expression and the right to privacy in New Zealand. As conflicting rights and values, there is a need for a wider public discussion about where the balance should sit. In the meantime, while that discussion takes place, we think much of the debate has been accounted for within the drafting of existing legislation.

336. We note that any restriction on access to or ability to share content should be conducted according to law, with rights to natural justice processes and appeal as appropriate in the circumstances. We think that a case-by-case approach is consistent with this and that is why we have referred to regimes such as the Privacy Act 1993.

337. To the extent that any immediate or urgent response is perceived as being required, we think it preferable to absorb SMTs and SMAs into existing legislative processes, which have already been through a democratic process whereby the ability of the Courts to weigh and balance competing policy factors is appropriately restricted and where a body of precedent and case law can develop.

338. There are some key statutory regimes where the right to freedom of expression is curtailed in a way that is justifiable in a free and democratic society. We deal with these next in our analysis to give examples of the kinds of factors that are taken into account by the legislature when it comes to the restriction on freedom of expression.

339. We note that the legal regime in question infrequently does refer to the NZBORA in an express form. The NZBORA is inserted into any situation where s 3 of that Act applies: namely, "to acts done —(a) by the legislative, executive, or judicial branches of the Government of New Zealand; or (b) by any person or body in the performance of any public function, power, or duty conferred or imposed on that person or body by or pursuant to law." Accordingly, where any power of that nature is exercised under the enactments we identify, the NZBORA's provisions at ss 4, 5 and 6 will apply.

## Broadcasting Act 1989

340. The Broadcasting Act is of limited assistance in dealing with SMA and technologies because it is limited to radio and television broadcasters in its application. Increasingly, large volumes of highly persuasive audio-visual media are consumed outside of these platforms. It is important to take more traditional broadcast media into account, however, because of the effect that institutionalisation has on the credibility of broadcasts made. For example, Youtube recently adopted a "news mode" that promotes more reliable journalistic sources once an event becomes newsworthy, so as to avoid the spread of misinformation and conspiracy theories arising around significant events. The fact that a broadcast is made through a medium that is regulated lends credibility to what it broadcasts. This is a phenomena we also emphasise with respect to the Media Council of New Zealand.

341. Despite the limited formal application of the Broadcasting Act 1989,[137] the text and application of it as recorded in the Broadcasting Standards, Codes, and BSA decisions are invaluable as a starting point for the assessment of how law should apply to the dissemination of audio, visual, and audiovisual information. Like the Films, Videos and Publications Classification Act, there is no need to start again afresh when existing legal analogues address the harms that we are seeking to analyse and avoid.

342. We think that any regulatory response to synthetic media must take account of the way that, over time, the Broadcasting Standards Authority has had to formulate predictable and rational ways of assessing broadcast content.

343. We also note that the Court of Appeal in *Hosking v Runting* drew on submissions it invited the BSA to make to it, and drew to an extent on the way that the BSA had come to define the right to privacy in a broadcast context. We include the Court of Appeal's summary of them here for completeness because of the way they articulate distinctions between public and private information, refer to the articulation of privacy torts that include "misappropriation of image", and the need to maintain the ability to publish information of high public interest:[138]

> [104] The Broadcasting Act does not provide any guidelines for what constitutes a breach of privacy of the individual. In 1992 the BSA enunciated five relevant privacy principles in an advisory opinion. They were the principles that it had been applying in respect of complaints alleging a breach of s 4(1)(c) of the Act. The principles are drawn from American case law and are essentially restatements of Prosser's principles. Two additional principles were added in 1996 and 1999 to address factual situations not covered by the existing principles, but which the BSA considered clearly showed a breach of s 4(1)(c). The possibility of developments like this was foreshadowed in the 1992 advisory opinion, which made the following points:
>
> – These principles are not necessarily the only privacy principles that the [BSA] will apply;
>
> – The principles may well require elaboration and refinement when applied to a complaint;
>
> – The specific facts of each complaint are especially important when privacy is an issue.
>
> Such comments are clearly relevant to any considerations of privacy, whether under statute or in tort, and highlight again the wide-ranging and fact-specific nature of privacy complaints.

344. We again note the iterative case-by-case approach adopted when it comes to privacy (as well as its endorsement by the Court of Appeal) and commend that as a sensible approach to synthetic media given its broad array of applications and outcomes.

345. We think the Broadcasting Act includes definitions which may be useful. We note them here in support of our conclusion that the broadcast of SMAs will be caught by the Act.

[137] Broadcasting Standards Act 1989.
[138] Above n 105 at [104].

346. The definition of "programme" as excluding "visual images … combined with sounds that consist predominantly of alphanumeric text" could equally apply to our definition of category 3 technologies.

> programme—
> (a) means sounds or visual images, or a combination of sounds and visual images, intended—
> (i) to inform, enlighten, or entertain; or
> (ii) to promote the interests of any person; or
> (iii) to promote any product or service; but
> (b) does not include visual images, whether or not combined with sounds, that consist predominantly of alphanumeric text

347. Section 4 of the Act includes explicit reference to the privacy of individuals. It also refers to the role of the Films, Videos, and Publications Classification Act 1993, illustrating the interconnecting nature of legal regimes touching upon SMA.

348. The Broadcasting Act sets out eleven areas from which broadcasting standards have been developed. These provide a useful starting point for any public or private actor who is seeking to understand how to assess the meaning and acceptability of audio, visual, or audio-visual artefacts. The eleven areas Parliament has indicated are legitimate areas of concern are:

a. good taste and decency,
b. programme information,
c. children's interests,
d. violence,
e. law and order,
f. discrimination and denigration,
g. alcohol,
h. balance,
i. accuracy,
j. privacy, and
k. fairness.

349. There is an extent to which the starting point for the Codes varies that may affect the way they are translated across for other uses: for example, the codes distinguish between paid and free-to-air television based on distinctions that may be difficult to maintain when it comes to digital media that is freely available.

350. We note the inclusion of "doorstepping" as the filming or recording of an interview or attempted interview with someone, without any prior warning" and the way that this regulates the use of category 1 technologies in certain contexts.

351. The BSA's Codebook commentary on the standards includes,[139] at p 15 for example, the acknowledgement that "Context is crucial in assessing the programme's likely practical effect."

352. The BSA's discussion of freedom of expression is particularly informative:[140]

> The importance of freedom of expression is such that, at some times, the exercise of it will cause offence to be taken by some or will result in harm being felt by some. Ultimately there is a sensible balance to be struck … We may only uphold complaints where the limitation on the right is reasonable, prescribed by law and demonstrably justified in a free and democratic society … The level of public interest in a broadcast is particularly important … If it deals seriously with political issues or other topics that help us govern ourselves and hold our leaders accountable it will carry a high level of public interest. … Conversely, broadcasting standards exist to ensure that broadcasters do not (for example) misinform us about important things, or unfairly harm the dignity or reputation of the people they

[139] Broadcasting Standards Codebook 2016 at p 15.
[140] Ibid at p 6.

> feature, or leave out significant viewpoints when telling us about issues that matter to us. … Ultimately this is a balancing process.

353. We note the commentary on two standards: "Privacy" and "Fairness".

354. Privacy is "the only broadcasting standard for which compensation may be awarded" and is therefore an area that "Parliament has identified … as particularly important".[141] The commentary on the privacy standard itself heavily emphasises exclusionary or solitude conceptions of privacy, but goes on to state that:[142]

> The privacy standard aims to respect, where reasonable, people's wishes not to have themselves or their affairs broadcast to the public. It seeks to protect their dignity, autonomy, mental wellbeing and reputation, and their ability to develop relationships, opinions and creativity away from the glare of publicity. But it also allows broadcasters to gather record and broadcast material where this is in the public interest. Our expectations of privacy vary with time, culture and technology, which creates some difficult boundaries …

355. We think that this approach to privacy more than adequately accounts for the prospect that the broadcast of wrong personal information about someone could infringe privacy principles.

356. We couple that analysis with the Standard about Fairness. The Codebook states that "The purpose of this standard is to protect the dignity and reputation of those features in programmes." It states that considering "fairness" will "generally take into account the following":[143]

    a. whether the audience would have been left with an unduly negative impression of an individual or organisation

    b. whether an individual or organisation taking part or referred to in a programme was adequately informed of the nature of their participation

    c. whether informed consent was required and/or obtained

    d. whether the individual or organisation was given a reasonable opportunity to comment, and whether their comments were adequately presented in the programme

    e. the nature of the individual, for example, a public figure or organisation familiar with dealing with the media, as opposed to an ordinary person with little or no media experience

    f. whether any critical comments were aimed at the participant in their business or professional life, or their personal life

    g. the public significance of the broadcast and its value in terms of free speech.

357. We note that the standard on fairness deals extensively in privacy considerations and explicitly refers to the guidance on privacy in the codebook.

358. In relation to the Guidance on Privacy we note the following aspects of the Codebook:

    a. per (1.1) and (2), privacy is framed in relation to identifiable individuals in the same manner as the Privacy Act and so our conclusions equally apply; that synthetic media should be considered on face in a factual manner as to whether it is "about an identifiable individual". There is no need to go behind the manner of creation of the SMA unless examining the truth or otherwise of the content of the broadcast.

[141] Ibid at p 21.
[142] Ibid.
[143] Ibid.

b. the discussion of whether there can be a reasonable expectation of privacy in a public place, or in relation to matters of public record or information in the public domain. Again, we think that emphasis should be placed on the broadcast itself and the features of the person it purports to show. The analysis should not be determined by whether someone's face, voice or appearance is generally public. The emphasis should be on what the broadcast shows. We also emphasise our views, in terms of the analysis of *Hosking v Runting*, that: it is difficult to see how a person could have an unreasonable expectation of privacy in relation to generated media that has never been public, and shows them doing something they never said or did; and that it is highly likely that a clip appropriating someone's visual or auditory identity will be highly offensive to the reasonable and ordinary person in most cases. We think that attention should be paid to the way in which privacy aims to protect human autonomy and dignity and the extent to which appropriation of someone's identity, and exclusive control about deceptive facts about them, undermines that dignity and autonomy. We note that the privacy standards really anticipate the use of category 1 capture technologies: we think there is a fundamental distinction between these and SMT that have a distinctly generative and open-ended capacity.

c. per (4) that public figures and people who seek publicity have a lower reasonable expectation of privacy in relation to matters pertaining to their public roles.

d. We note that (6.1) could be taken to refer to the means by which synthetic media artefacts are generated. "The means by which private material is gathered affects the offensiveness of the intrusion or disclosure. For example, it may be highly offensive to broadcast private material gathered by surreptitious, deceptive or dishonest means."

e. We think that number 7 is also relevant in relation to informed consent. Where a person is identifiable in a broadcast, they must be aware they are contributing to the broadcast and freely agree to contribute. Broadcasters will have to take care per 7.5 not to infer consent as being "obvious from the circumstances" or recorded if an SMA is highly manipulated to achieve this effect.

f. We also note the guidance given about the use of hidden cameras and covert filming at item 9 and its impact on the use of capture technologies.

359. We note the guidance given on assessing accuracy and distinguishing fact and analysis, comment or opinion. While it contains useful factors, none are particularly relevant to SMT other than the way that audio-visual evidence may need to be treated more sceptically as a source of evidence about how factual events occurred. We echo the BSA's statement that "none of [the factors referred to] is conclusive. Every case must be assessed on its merits."

360. If necessary, one option for the BSA to deal with synthetic media is to issue, pursuant to s 21(1)(d) "to any or all broadcasters [an] advisory opinion relation to broadcasting standards and ethical conduct in broadcasting". We note that, if such guidance was included in amendment to the broadcasting code of practice, pursuant to s 21((1)(e)-(g), then it would be obliged to consult with the Privacy Commissioner pursuant to sub (4).

361. The Broadcasting Act and associated instruments give guidance on the kind of limitations imposed on dissemination technologies where harms may arise from the content and capture processes.

## Electoral Act 1993

362. The Electoral Act 1993 ("**EA**") offers some limited assistance in preventing the disruption of elections by means of SMT and SMA. We note this limitation for several reasons. Not only is the EA stringently framed (so as to protect and not impinge upon important democratic protocols), but its application is narrowed to very particular circumstances. Moreover, practical realities of the way information is published and consumer throughout the global internet limit the effectiveness of domestic electoral law. Ultimately, the EA may succeed in deterring and punishing the most egregious examples of SMA misuse, such as releasing a politically explosive deepfake on the eve of an election. However, in the broader context of "fake news" and political misinformation, it offers a limited tool by which to protect against the deceptive and disruptive potential of synthetic media. Nevertheless, we remain sceptical that any new law could improve upon the current regime, which is framed as it is to reflect a range of factors, pressures and limitations that any electoral law must reasonably appreciate. Any legislative expansion towards this end must proceed with extreme caution.

363. Much of the concern surrounding emerging audiovisual technologies has centred on its potential to disrupt democratic processes. Many technological demonstrations represent politicians and other public figures as the targets of deliberately fake videos and audio. Figures like politicians and heads of state tend to be photographed and recorded very often, and as a result, large volumes of high-quality digital data of these individuals tends to be publicly accessible. As a consequence, high quality synthetic video and audio representations can be produced when targeting these individuals. This coincides with the fact that misrepresentation of these individuals has the potential for unique and far-reaching disruptive impacts.

364. A typical hypothetical scenario for democratically disruptive synthetic media goes something like this: the day before the general national election, a high quality deepfake video is published and spreads across social media platforms where it is consumed by many thousands of potential voters. In the video, the leader of the opposition appears to have been visually or aurally captured in a scenario that demeans him or her in the eyes of the public. With little time left before the polls open, investigative authorities are left with insufficient time to prove the falsity of the video through forensic techniques, effectively communicate the fact of its falsity to electors, and identify the videos source. This could equally apply to any kind of synthetic media artefact.

365. In a scenario like this, would the publication of the video be prohibited under New Zealand law? When applying the EA, the answer is yes, providing that a range of elements are satisfied. Of course, there also are a number of 'reality checks' on the effectiveness of law in a given scenario where synthetic media may be disrupting an election. For instance, there is the ongoing reality of cross-jurisdictional information exchange via the internet. While New Zealand publishers are obliged to obey the EA, the same may not be said for overseas publishers. And yet, because New Zealand citizens have relatively free access to overseas internet-based materials, they may still be exposed to synthetic media from abroad, even where that media may influence their political decision, and even where the information would otherwise be prohibited by the EA. The pragmatic difficulties of scenarios like this were brought into sharp focus recently in New Zealand via media coverage of the investigation and trial for the murder of Grace Millane. In this case, New Zealand law, and New Zealand law by enforcement authorities by extension, found it difficult to prevent overseas publication of details of the accused, even where the details being revealed posed a serious risk of undermining the integrity of the legal proceedings.[144]

366. With these pragmatic limitations in mind, the two provisions of greatest relevance are s 199A Publishing false statements to influence voters and s 197 Interfering with or influencing voters. Both are likely to overlap in many fact patterns.

367. Section 199A of the EA deals with false statements distributed to influence voters:

[144] RNZ "Grace Millane case: Suppression breaches could endanger trial" (13 December 2018) <www.rnz.co.nz>.

> 199A Publishing false statements to influence voters
> (1) A person is guilty of a corrupt practice if the person, with the intention of influencing the vote of an elector,—
> (a) first publishes or republishes a statement, during the specified period, that the person knows is false in a material particular; or
> (b) arranges for the first publication or republication of a statement, during the specified period, that the person knows is false in a material particular.
> (2) Subsection (1) does not apply if—
> (a) the statement was first published before the specified period and remains available or accessible within all or part of the specified period; but
> (b) the person did not, during the specified period, by any means,—
> (i) advertise or draw attention to the statement; or
> (ii) promote or encourage any person to access the statement.

368. We believe this posits seven elements that must be established: existence of a statement, falsity of that statement, the fact of that falsity being material, knowledge of the falsity, publication of the false statement, the coincidence of these five elements within a specified time period, and finally the intention to influence the vote of an elector.

369. First there is the question of whether the video in the hypothetical qualifies as a 'statement' in accordance with the language of s 3 EA. The EA defines "statement" through an extending definition in a manner that is particularly broad by its inclusion of any "methods of signifying meaning".

> **statement** includes not only words but also pictures, visual images, gestures, and other methods of signifying meaning

370. The EA does not contain a purpose provision and its long title does not assist. However, with regards to the intention behind s 199A at the time of enactment, in the case of *Peters v The Electoral Commission* [2016] NZHC 394,[145] the judgment of Mallon J notes that it was the view of the Electoral Commission that:[146]

> … s 199A was inserted to address a specific concern about parties or candidates publishing late statements on the eve of an election which meant that other candidates and parties had insufficient time to correct the statement.

371. This intention corresponds to the original s 199A provisions, later repealed and replaced by the Electoral Amendment Act 2017. Originally, s 199A read:

> 199A Publishing false statements to influence voters
> Every person is guilty of a corrupt practice who, with the intention of influencing the vote of any elector, at any time on polling day before the close of the poll, or at any time on any of the 2 days immediately preceding polling day, publishes, distributes, broadcasts, or exhibits, or causes to be published, distributed, broadcast, or exhibited, in or in view of any public place a statement of fact that the person knows is false in a material particular.

372. Justice Mallon accepted the assertion of the Electoral Commission regarding legislative intention based on a comprehensive analysis of the legislative history of the EA, including Parliamentary speeches. Her Honour summarised:[147]

> [72] Having reviewed these materials I accept that the intention was to capture false statements made shortly before an election. The reason for the provision was to address the problem of statements which could influence a voter when there would not be a sufficient opportunity to correct them. This was seen as a justified limit on freedom of expression in contrast with the defamation provision which was viewed as an unjustified limit and excluded from the Act.

[145] *Peters v The Electoral Commission* [2016] NZHC 394.
[146] Ibid at [11].
[147] Ibid at [72].

373. The changes introduced by the Electoral Amendment Act 2017 clarified one of the central issues in *Peters*. Nevertheless, record of the Electoral Commission's opinion of the intention for s 199A remains a useful interpretative guide. Based on both the meaning of "statement" provided by s 3 and the purpose behind s 199A, it seems almost certain that many forms of SMA – including deepfake videos or synthetic voice audio clips – will qualify as 'statements' for the purposes s 199A.

374. Next is the element of falsity. We have gone to some lengths to illuminate the issues around describing the "falsity" of SMA: namely, that all digital media is to some degree false by virtue of creation and manipulation technologies, and that this media can be heavily manipulated but otherwise benign or beneficial and still a reliable record of events.

375. The statement must also be false in a material particular. This is a question of fact, the possible permutations of which are numerous. Ultimately it is a task for the judgment of the judiciary on a case by case basis. In the case of a malicious electoral deepfake, however, it seems unlikely that the statement would false in a way that is immaterial.

376. The first of two *mens rea* considerations arising out of s 199A is knowledge of falsity. Proving this element is an evidential matter for the fact-finder based on the particular facts of each case and takes on its own peculiarities in relation to synthetic media artefacts. It is highly unlikely that any person who created a synthetic media artefact could ever be unaware of its falsity. The only rare situation where this may be plausible is where an unsupervised automation process has played a large role in producing the artefact without any person's direct oversight, so much so that the creator of the SMA may be to some greater or lesser degree unaware of the falsity of that which they have created. Such a scenario is plausible through use technologies like GANs, but remains an unlikely occurrence.

377. More difficult to establish is the matter of knowledge of falsity when the statement is published or re-published by someone other than the creator of the statement. For example, a person might truthfully be unaware of the falsity of a deepfake video if the video is of such photorealistic quality that it appears to have been produced via a capture process. A person could come across this video in the course of using the internet, re-publish it by sharing it on social media platforms, and the crucial element of knowledge could well be absent in their actions.

378. Next is the element of publication, which is almost certainly present with regards to our hypothetical deepfake. The statement must be published or re-published, the definition for which is established by ss 199A(3)(a) and (b). For synthetic media technologies, ss 199EA(a)(i), (ii), (viii), (ix), (x), (xi) are particularly relevant:

> (3) In this section,—
> **publish**, in relation to a statement, means to bring to the notice of a person in any manner,—
> (a) including by—
> (i) displaying on any medium:
> (ii) distributing by any means:
> ...
> (vii) broadcasting by any means:
> (viii) disseminating by means of the Internet or any other electronic medium:
> (ix) storing electronically in a way that is accessible to the public:
> (x) incorporating in a device for use with a computer:
> (xi) inserting in a film or video; but
> (b) excluding addressing 1 or more persons face to face

379. Sixth is the coincidence of all previous elements within the specified time period as established by s 199A(3):

> **specified period** means the period—
> (a) beginning 2 days immediately before polling day; and
> (b) ending with the close of the poll.

380. Finally comes the overarching element of intention to influence the vote of an elector, second of two mens rea considerations. Again, this is a matter to be assessed on the facts of each case, but in many cases, intention is likely to be inferred simply by:

a. the number of steps and amount of effort required to create and distribute an SMA;

b. the contents of that SMA, particularly whether it represents something that would be significant to electors; and

c. to create it to a degree of realism that gives rise to condition 1 of our framework; and lastly,

d. the context in which the SMA is presented.

381. Section 197 of the Act provides a range of circumstances where interference with or influence of voters is prohibited, the most relevant of which are included below:

> 197 Interfering with or influencing voters
> (1) Every person commits an offence and shall be liable on conviction to a fine not exceeding $20,000 who at an election—
> ...
> (c) at any time on polling day before the close of the poll makes any statement having direct or indirect reference to the poll by means of any loudspeaker or public address apparatus or cinematograph or television apparatus:
> provided that this paragraph shall not restrict the publication by radio or television broadcast made by a broadcaster within the meaning of section 2 of the Broadcasting Act 1989 of—
> (i) any advertisement placed by the Electoral Commission or a Returning Officer; or
> (ii) any non-partisan advertisement broadcast, as a community service, by a broadcaster within the meaning of section 2 of the Broadcasting Act 1989; or
> (iii) any news in relation to an election:
> ...
> (g) at any time on polling day before the close of the poll exhibits in or in view of any public place, or publishes, or distributes, or broadcasts,—
> (i) any statement advising or intended or likely to influence any elector as to the candidate or party for whom the elector should or should not vote; or
> (ii) any statement advising or intended or likely to influence any elector to abstain from voting; ...

382. With regards to s 197(1)(g), again there is no reason why various SMA will not be captured by the definition of "statement" provided by s 3. Therefore, any person exhibiting a deepfake, synthetic voice audio clip, or other in view of any public place, or publishing, or distributing, or broadcasting that information with intention to influence any elector in their vote will be liable to legal sanction. Where s 197(1) goes further than s 199A is in extending liability for acts merely "likely" to influence an elector. The "likelihood" of influence in any given case is a question of fact.

383. With regards to s 197(1)(c), the same reasoning with regards to "statement" applies. As a result, displaying a deepfake video via any of the specified technologies is prohibited in the stipulated circumstances.

## Possible amendment to s 199A

384. Recklessness towards the possibility that a statement is false is not an element of the s 199A offence, and therefore without demonstrable knowledge that a statement is false, s 199A cannot apply. It is feasible that an individual may be unaware that a an SMA is false, especially where the SMA is highly realistic in terms of condition 1. Their sharing of the artefact may be based on genuine belief in the veracity and veridicality of its contents. Alternatively, an individual may claim that they did not know for certain even when they did, or when they at least ought to have been suspicious. It therefore is worth considering whether some lower threshold than 'knowing' ought also to be

culpable where SMA are published or re-published, particularly in close proximity to elections. This would of course need to be weighed against the right to freedom of expression established by s 14 of the NZBORA and the chilling effect it could have on political speech, or even a tendency to share audiovisual information that is both reliable and significant to voters' decisions.

385. Standards for recklessness would likely shift as SMA become more common, and as the citizens and consumers become more widely alert to the phenomenon of synthetic media and its verisimilar potential. For now, many people are still unaware of the existence of things like deepfakes, and thus a determination of recklessness in any decision to pass on such audio or video likely would be unjust. One can scarcely be reckless towards a risk that few people are yet aware of. However, as deepfake-style information becomes more frequent in the public sphere, and as consumers become accustomed to its presence and its risks, the bar for reckless behaviour ought to lower. One may not 'know' for certain that a video is synthesised and therefore represents events which never actually occurred. Nevertheless, one ought to be aware of the possibility when making a decision to publish or re-publish for the purposes of influencing the vote of an elector.

386. The adequacy of the 2 day specified period also requires serious consideration. As far back as the second reading of the original Bill, Richard Worth (National) questioned its effectiveness:[148]

> It is an unusual provision in many respects, and I question its workability. … I would like the Minister … to explain the justification for that 2-day period. I suggest that particularly in rural electorates throughout New Zealand, if such defamatory material was published to influence voters, then 2 days would be nowhere near sufficient time to correct what might be highly objectionable, highly offensive, and possibly criminally libellous material. In the context of the Defamation Act, for example, 2 days will not permit resort to the range of remedies available in the legislation.

387. Furthermore, Mallon J in *Peters* added the following:

> [73] The materials do not particularly assist with why the two day time frame was selected. The Committee noted the four day time frame which, at that time, applied to complaints under the Broadcasting Act. There would therefore be insufficient time for a false statement made three days before polling day to be deal with under the broadcasting Act procedures.
>
> [74] The conclusion I draw is that the two day time frame was selected to ensure the offence was not too widely cast. If it was too widely cast it could have a chilling effect on legitimate campaigning and thereby potentially impinge on the right to freedom of expression beyond that which was justified. False statements three days out would not be criminal because a candidate of a party would at least have three days before polling day to respond in some way (not necessarily through the BSA processes). Two days out from polling was regarded as insufficient time to respond.

388. It is important to note that, as well as the assumption of Parliamentary intent when first introduced in the EA 1993, the 2-day stipulation also survived the substantial changes introduced by the Electoral Amendment Act 2017. The presumption must therefore be that it represents careful thought and purposive drafting by Parliament. Nevertheless, the scope of technological change since then may oblige us to consider whether there is anything about synthetic media that might instigate an extension of this period.

389. With this in mind we ought to consider whether emerging audiovisual technologies introduce any new capacity justifying revision of the 2-day time period. The increase in advance voting in New Zealand in recent years may also be relevant to a 2-day time period. In terms of dissemination, emerging synthetic media artefacts are not substantially different from any other digital technology. In terms of their content, there is some room to argue that things like deepfakes are more persuasive than traditional "Photoshop" artefacts or so-called "shallow fakes", and moreover, that electors still wrongly believe in the reliability of video and audio. But it is not clear that these differences are so great as to behove Parliament to extend the 2-day time period, all things considered. One of the

[148] (15 November 2001) 596 NZPD 13167.

benefits of our framework is that it allows this compare and contrast process to take place pursuant to a consistent process.

390. Moreover, if two days were deemed insufficient as a result of technological change, it remains unclear what sort of time period would be more appropriate. It could take weeks to determine the falsity of a particular persuasive SMA, either by traditional investigative techniques or digital forensics. Some SMA may simply be unfalsifiable, even with the most advanced technological detective tools. As such, the length of the time period - if deemed too short - may be better amended to reflect the time it takes for electors to 'move on' from a given political phenomenon than by reference to the time it takes to conduct an effective investigation into the truth or falsehood of a given SMA. In the context of an election, democratic values like the right to free expression take on particular importance. Too heavy-handed a limitation on speech in the lead-in to an election could generate as many harms as it seeks to prevent, albeit less sensational ones than those arising from targeted use of SMA.

### Political interference outside of these prohibited circumstances

391. The combination of ss 197 and 199A deter against the use of synthetic media technologies to influence elections, but their application only touches narrow circumstances: either published false statements in the 2 days preceding polling, or all published statements on the day of polling. Moreover, these generally require that the publisher have knowledge of the falsity - which may be undermined by high-quality SMT – as well as an intention to influence an elector.

392. This leaves open the issue of political interference that may occur outside of these periods but which nonetheless has the potential to interfere with an election or influence an election result. In all likelihood, the majority of politically disruptive deepfakes are not likely to qualify as "election advertisements" unless they are issued by an individual who receives payment in respect of the deepfake as a result of s 3A(2)(e):

> 3A Meaning of election advertisement
> …
> (2) None of the following are election advertisements:
> …
> (e) any publication on the Internet, or other electronic medium, of personal political views by an individual who does not make or receive a payment in respect of the publication of those views.

393. We conclude that the threat of politically disruptive deepfake images, videos, and audio remains a matter for further policy debate, in light of the other potential statutory limitations on these artefacts, and with due deference to factors like freedom of expression under the NZBORA. It is essential to note that synthetic media is just one means by which disinformation may be accidentally or intentionally spread. Any policy response which restricts this media more than others would need to be justifiable. On balance, existing law seems positioned to do a relatively good job in narrow circumstances, and a relatively poor one outside of those. The latter is true for many scenarios involving internet-based digital media, and is by no means distinct to synthetic media. It is also important to consider the possibility that intense restriction in narrow circumstances against a background of relative freedom is important for free elections.

## Films, Videos, and Publications Classification Act 1993

394. The Films, Videos, and Publications Classification Act 1993 ("**FVPCA**"),[149] according to its long title, consolidates and amends "the law relating to the censoring of films, videos, books, and other publications". It repealed "the Indecent Publications Act 1963, the Films Act 1983, and the Video Recordings Act 1987". It is applied by the Classification Office led by the Chief Censor.

395. It explicitly deals in censorship of audiovisual content based on perceptions of harm and is therefore central to assessing the extent to which New Zealand law touches upon harmful synthetic media artefacts.

396. The Act primarily focuses on the harms flowing from the content of a publication (or SMA) and its dissemination. It generally does not require a decision-maker to go behind the category 3 artefact to assess the impact of category 1 and 2 technologies. It is therefore less concerned with deception and more concerned with harmful content.

397. Clearly, in the wake of the live-streamed terrorism in Christchurch,[150] and in the context of child sexual exploitation material, sufficiently harmful content on the internet can be censored and dealing in it can lead to criminal sanction. It is important to note, however, that there are careful procedural and substantive limitations on the Classification Office's ability to do so. This is consistent with the high value placed on freedom of expression in a democratic political system.

398. When it comes to considering the extent to which harmful SMA should be banned based on their content, we think careful attention should be paid to the drafting of the FVPCA. The drafting also reflects the distinctions we note in our framework.

### Relevant definitions

399. As with the Copyright Act, there are a number of useful definitions in the FVPCA. We argue that SMA fall within the ambit of the FVPCA when understood in terms of Categories 1-3 of our framework:[151]

> publication means—
> (a) any film, book, sound recording, picture, newspaper, photograph, photographic negative, photographic plate, or photographic slide:
> (b) any print or writing:
> (c) a paper or other thing that has printed or impressed upon it, or otherwise shown upon it, 1 or more (or a combination of 1 or more) images, representations, signs, statements, or words:
> (d) a thing (including, but not limited to, a disc, or an electronic or computer file) on which is recorded or stored information that, by the use of a computer or other electronic device, is capable of being reproduced or shown as 1 or more (or a combination of 1 or more) images, representations, signs, statements, or words

400. Paras (c) and (d) are broad enough to encompass all SMA we have encountered, particularly the definition of a publication as a "thing" with further extensions. The reference to reproduction again corresponds to our definition of category 3 of the framework.

401. The terms in para (a) are also defined in ways that would include SMAs, although notably there is no notion of a category 1 capture technology being involved in the way the definitions are drafted, contrary to the Copyright Act (dealt with later in Part 3):

> **film** means a cinematograph film, a video recording, and any other material record of visual moving images that is capable of being used for the subsequent display of those images; and includes any part of any film, and any copy or part of a copy of the whole or any part of a film

[149] Films, Videos, and Publications Classification Act 1993.
[150] Consecutive shootings a the Al Noor Mosque and Linwood Islamic Centre on 15 March 2019.
[151] Ibid s 2.

**video recording** means any disc, magnetic tape, or solid state recording device containing information by the use of which 1 or more series of visual images may be produced electronically and shown as a moving picture

**video game** means any video recording that is designed for use wholly or principally as a game

402. The Act is relatively limited because it only requires films to be labelled when they will be supplied or exhibited to the public.

> 6 Films to be labelled
> (1) Subject to sections 7 and 8, a film must not be supplied to the public or offered for supply to the public unless—
> (a) a label has been issued in respect of that film; and
> (b) the requirements of this Act and of any regulations made under this Act with respect to the display of that label are complied with.
> (2) Subject to sections 7 and 8, a film must not be exhibited to the public unless—
> (a) a label has been issued in respect of that film; and
> (b) the requirements of this Act and of any regulations made under this Act with respect to the display and advertising of the contents of that label are complied with.

403. Supply is given a relatively narrow commercial meaning that excludes a large number of video services:

> **supply** means to sell, or deliver by way of hire, or offer for sale or hire
>
> **supply** to the public, in relation to a film,—
> (a) means supply by way of sale, hire, exchange, or loan, in the course of any business; and includes ...
>
> **exhibit**, in relation to a sound recording, means to play that sound recording
>
> **exhibit to the public**, in relation to a film,—
> (a) means to screen or arrange or organise the screening of, or to assist any other person to screen or arrange or organise the screening of, the film—
> (i) to the public, or any section of the public; or
> (ii) to any group or class of persons otherwise than in a private residence,—
> whether or not a charge is made for admission to the premises in which the exhibition is held; but
> (b) does not include the broadcasting of the film;—
> and public exhibition has a corresponding meaning

404. Many kinds of films are exempt from labelling requirements, although they can be submitted for labelling at the requirement of the Chief Censor if exhibited or supplied to the public pursuant to sub (2) .

> 8 Films exempt from labelling requirements
> (1) Subject to subsections (2) and (3), section 6 does not apply in respect of any of the following films:
> ...
> (g) any film of news and current affairs, any documentary, and any historical account containing a unity of subject matter:
> (k) any film that is wholly or mainly a commercial advertisement relating to the advertiser's or sponsor's activities:
> (l) any film directly related to the curriculum of pre-school, primary, secondary, or tertiary educational institutions:
> (m) any film wholly or mainly of a religious nature:
> (n) any film depicting wholly or mainly travel:
> (o) any film depicting wholly or mainly cultural activities:
> (p) any film intended for supply or exhibition solely to ethnic organisations:
> (q) any video game.

(2) The Chief Censor may, at any time, require any person who proposes to exhibit to the public or supply to the public any film of a class mentioned in subsection (1), or who has exhibited to the public or supplied to the public any such film, to make an application under section 9 for the issue of a label in respect of that film.

(3) Nothing in subsection (1) exempts any film from the requirements of section 6 if—

(a) the film is a restricted publication; or

(b) the Chief Censor has required the film to be submitted to the labelling body under subsection (2).

### Objectionable publications

405. The Act regulates "objectionable" publications. Per s 3(1), "For the purposes of this Act, a publication is objectionable if it describes, depicts, expresses, or otherwise deals with matters such as sex, horror, crime, cruelty, or violence in such a manner that the availability of the publication is likely to be injurious to the public good."[152]

406. Within that wider context, it deals with sexual publications about children or young persons.

407. In all cases, the Act deems publications to be objectionable if they deal in certain matters. Notably, there is no reason to believe that the definition of "promotes or supports, or tends to promote or support" would exclude material that has not made significant use of capture technologies. Therefore, there is little need to go behind the Category 3 artefact itself. Notably, "promotion or support" goes beyond mere depiction.

(2) A publication shall be deemed to be objectionable for the purposes of this Act if the publication promotes or supports, or tends to promote or support,—

(a) the exploitation of children, or young persons, or both, for sexual purposes; or

(b) the use of violence or coercion to compel any person to participate in, or submit to, sexual conduct; or

(c) sexual conduct with or upon the body of a dead person; or

(d) the use of urine or excrement in association with degrading or dehumanising conduct or sexual conduct; or

(e) bestiality; or

(f) acts of torture or the infliction of extreme violence or extreme cruelty.

408. Section 3 further creates a list of factors that must be given "particular weight". Again, the drafting is that a publication "describes, depicts, or otherwise deals with" certain matters, a lesser standard than deployed in sub (2). The other operative language in paras (b)-(e) does not appear to require that the extent of category 1 or 2 technologies have been utilised:

(3) In determining, for the purposes of this Act, whether or not any publication (other than a publication to which subsection (2) applies) is objectionable … particular weight shall be given to the extent and degree to which, and the manner in which, the publication—

(a) describes, depicts, or otherwise deals with—

(i) acts of torture, the infliction of serious physical harm, or acts of significant cruelty:

(ii) sexual violence or sexual coercion, or violence or coercion in association with sexual conduct:

(iii) other sexual or physical conduct of a degrading or dehumanising or demeaning nature:

(iv) sexual conduct with or by children, or young persons, or both:

(v) physical conduct in which sexual satisfaction is derived from inflicting or suffering cruelty or pain:

(b) exploits the nudity of children, or young persons, or both:

(c) degrades or dehumanises or demeans any person:

(d) promotes or encourages criminal acts or acts of terrorism:

(e) represents (whether directly or by implication) that members of any particular class of the public are inherently inferior to other members of the public by reason of any characteristic

[152] Ibid s 3(1).

> of members of that class, being a characteristic that is a prohibited ground of discrimination specified in section 21(1) of the Human Rights Act 1993.

409. Further, it creates a list of other factors that must also be considered, and these relate to the way that the publication will be consumed in context and who will be exposed to it. We think this acknowledges the way that different harms arise from SMA based on the way it is disseminated (condition 3) and the way that context and assumptions can contribute to the way it is consumed (including for example whether it is deceptive per condition 1, although "deception" is not a relevant characteristic in s 3):

> (4) In determining, for the purposes of this Act, whether or not any publication (other than a publication to which subsection (2) applies) is objectionable … the following matters shall also be considered:
> (a) the dominant effect of the publication as a whole:
> (b) the impact of the medium in which the publication is presented:
> (c) the character of the publication, including any merit, value, or importance that the publication has in relation to literary, artistic, social, cultural, educational, scientific, or other matters:
> (d) the persons, classes of persons, or age groups of the persons to whom the publication is intended or is likely to be made available:
> (e) the purpose for which the publication is intended to be used:
> (f) any other relevant circumstances relating to the intended or likely use of the publication.

410. The Classification Office is required to make classification decisions per s 23 "as soon as practicable after a publication has been submitted or referred" to it. Section 23(3) specifically allows a publication "that would otherwise be classified as objectionable may be classified as a restricted publication in order that the publication may be made available to particular persons or classes of persons for educational, professional, scientific, literary, artistic, or technical purposes."

### How does an SMA come to be classified under the Act

411. Publications can be submitted pursuant to s 13 by a list of identified Government officers or "any other person" who has the leave of the Chief Censor pursuant to a process at s 15. This requires a notice of submission to be lodged in a particular form. Where the Chief Censor declines to grant leave, they must give reasons. The Chief Censor can issue guidelines on whether leave should be given to submit the publication for classification.[153]

412. Where a person submits a publication under s 13, "[t]he Chief Censor must immediately determine the notice of the submission that is to be given to any person (other than the submitter) who the Chief Censor reasonably believes should be given notice of the submission by reason of that person's interest in the publication", but significantly, those interests are limited to "an interest as owner, maker, distributor, or publisher of the publication".

413. Where a publication is submitted under s 13, the person who submitted the publication can make written submissions in respect of the classification to be made. Per s 20(1)(d), the right to make submissions can also be extended to "such other persons who satisfy the Chief Censor that they are likely to be affected by the classification of the publication."

414. We note that a labelling body can be approved by the Minister pursuant to s 72 and submissions made to the Chief Censor by that labelling body too per s 12. The relevance of these labelling bodies will depend on the extent to which SMA that are films available for supply pursuant to the definitions in the Act will lead to harms under the framework.

[153] Ibid s 16.

415. The Act is limited to some extent in the way that the classification process applies or does not apply to a large volume of the SMA consumed by individuals in New Zealand.

### Technical assistance

416. We note pursuant to s 21 that the Office may seek the assistance of "any person whom [it] considers may be able to assist the Office in forming an opinion … on which to base the decision". Further, it can invite such persons as it thinks fit to make written submissions and "obtain information from such persons, and make such inquiries, as it thinks fit."

417. Section 88 of the Act creates an "Information Unit". This unit could be used to assess or disseminate information about SMT and SMT. Its function is to provide "such research services as may be necessary to enable the Classification Office to perform its functions effectively" and:

> (b) to disseminate to the public information about—
> (i) the functions and powers of the Classification Office; and
> (ii) the procedures for the classification of publications:
> (c) to receive inquiries and complaints concerning the operation of the classification system established under this Act.

418. Section 47 of the Act creates a list of people who can seek review by a separate Board if they are dissatisfied with a classification decision. The list at (2) does not appear to specifically anticipate that an identifiable individual who is the subject of a publication could seek a review, although "any other person" may seek review "with the leave of the Secretary". Where a review of a classification decision is sought, the Board can make interim restriction orders in relation to the publication, which could be used to limit someone's ability to deal in a publication while review is sought.

419. A decision by the Board can be appealed on a question of law to the High Court pursuant to s 58 demonstrating the importance of judicial oversight over censorship decisions.

### Enforcement powers

420. The Act confers powers of seizure in relation to publications.[154]

421. It also creates criminal offences for distribution of restricted and objectionable publications. It creates strict liability offences in relation to making, copying, importing, supplying, distributing, possessing, and displaying an objectionable publication. The gravity of penalty in relation to these offences increases if it can be shown a person had knowledge that the publication was classified as objectionable. Some of these offences can lead to imprisonment.

### Could individual privacy or deception be added as a criteria under the FVPCA?

422. The FVPCA is important because it shows that:

a. where the content of an SMA reaches a particular threshold of harm, the law will intervene;

b. there is a distinction between harms arising from content itself and the harms from disseminating that content, although these will often be strongly linked;

c. in appropriate circumstances the law will intervene in ways that are particularly intrusive, including by criminalising possession of information and enabling seizure by authorised individuals and criminal penalties;

[154] Ibid ss 107, 108.

d. the law will intervene in harms that are disparate and injurious to the public good without necessarily identifying any particular victim;

e. the kinds of harms of content that are perceived to be so serious as to justify this intervention, which enables us to consider whether SMAs and SMT are of a similar level of harm such that similar interventions are justified.

423. The FVPCA already deals in content which can be harmful per se or based on its context. Does this make it a good home for wider concerns about fake or harmful media? One question we have considered is whether it would be possible for an SMA that does not otherwise depict the matters described in s 3 to be so deceptive or so harmful in terms of the absence of consent or the privacy of the subject that it could be said to be injurious to the public good.

424. A subsequent question is whether such considerations (absence of consent, capacity for deception) are relevant under the Act as currently drafted, or whether they would need to be added to the list of criteria at s 3 such that publications of a sufficiently harmful deceptive or non-consensual nature could be objectionable.

425. We identify several issues with this approach:

a. First, the volume of potential publications that are being produced on a regular basis are likely to overwhelm the Office of Film and Literature Classification.

b. Second, the classification office would require access to digital forensic services dealing with capture and manipulation. As it stands, it primarily deals with harms of content and dissemination that means it can deal with category 3 artefacts on face. This would be a substantial increase in scope for the Office, who already faces challenges arising from the significant volume of audiovisual content in the modern world.

c. Third, the Act requires submission by somebody. This requires their knowledge of it. If that is the case, why focus on this enactment rather than the Privacy Act or HDCA? Further, the Act does not create civil remedies for individuals, it is drafted in terms of a relationship between individual and State.

d. Fourth, it would thereby lead to the Chief Censor taking on a role that requires it to consider concepts of privacy and truth. Privacy is already the expertise of the Privacy Commissioner. We have made it clear that a standard of truth or deception can already be incorporated into a privacy regime, both in terms of correction of personal information and ability to limit consent to deal with personal information in the form of data by manipulating it in certain ways. Further, there is no easy standard of "truth" that can be applied to synthetic media's content to enable the Classification Office to make predictable decisions. Finally, the HDCA also deals in digital communications that are harmful on the basis of falsity or confidentiality.

426. There is also a substantial risk in increasing the Chief Censor's remit to be an arbiter of truth. That would require extensive public consultation. We have considered whether, for example, misleading disinformation about disparate harms such as "fake news" or to the political process would be "injurious to the public good" in such a way that the Chief Censor's office is a logical home for such concerns. Firstly, this is already a task given in part to the electoral commission under the Electoral Act. Secondly, this would politicise the Chief Censor in a way that could be concerning.

427. Another requirement for this to be effective would be that the Chief Censor would adopt responsibility for effectively any audio or visual recording. This is unlikely to be sustainable or politically palatable. The Act itself evidences a clear restriction of scope that acknowledges the potential risks of government censorship.

428. We also add that the potential volume of SMA which may or may not be prohibitable material under the Act could easily overwhelm the limited capacity of the authority to respond unless it were to receive greater resources. This is an issue arising across most or all of the applicable legal regimes, to some extent evidencing that some of the main impediments to effective law are practical limitations, rather than defects in the legislation.

## Defamation

429. The law of defamation is intricate, vast and detailed beyond the scope of this report. We note it here in support of the following high-level conclusions which merit further scholarly investigation and may assist any person who believes they have suffered actionable harm from the use of synthetic media artefacts.

430. Defamation is a tort governed in part by the Defamation Act 1992.[155] It relates to civil wrongs between two parties.

431. Defamation requires that there has been a publication beyond the plaintiff and the respondent. In this sense, the Harmful Digital Communications Act potentially goes where defamation cannot, being applicable even to private communications between two persons, irrespective of whether those communications ever extend to others. Defamation therefore is largely to do with reputation, targeting the harms to reputation which can result from dissemination of a synthetic media artefact per Condition 3. Dissemination may be wide (to many parties) or narrow (to one other party than the plaintiff). Potential for defamation is axiomatic in synthetic media, where highly persuasive fake audio and video may cause extensive reputational damage, either to the individual who is represented in the artefact, or to some other person.

432. The law of defamation is a crucial part of the network of regulation surrounding synthetic media. It has an existing body of law that weighs and balances public interest in freedom of speech with a person's right not to have their reputation defamed unjustifiably.

433. Defamation deals heavily in the publication of wrong facts and can be contrasted with the way that people generally think about privacy law, which is the publication of true facts about which someone has a reasonable expectation of privacy (although we refer to our conclusion that wrong information about an identifiable individual can and should be personal information for the purposes of the Privacy Act).

434. A range of defences to defamation are likely to be unavailable to the creator of and SMA where defamation is alleged to have occurred in the form of an SMA. Defences like truth,[156] and honest opinion,[157] are likely made unavailable as a result of the fact that SMA must be constructed or generated. However, there is potential for scenarios to arise where persons view an SMA and form opinions based on its contents. The opinions may be genuine and honest,[158] and therefore repeating them more defensible under the DA. This problem is unlikely to be novel to defamation. Rather, it may simply be a factor to consider as more realistic, persuasive SMA become common on the informational market.

435. We think that Defamation will apply to emerging SMA. As a tort, it would require no obvious statutory amendments to do so. We note that cartoons have been held to be capable of carrying a defamatory meaning,[159] even though they are also widely regarded as being expressions of opinion that tend to be protected by the defence of honest opinion.[160]

436. The access to justice barriers involved in civil litigation are significant. This means that defamation will only be available as a remedy in cases where the use of synthetic media has generated significant harms to well-resourced litigants.

[155] Defamation Act 1992.
[156] Ibid s 8.
[157] Ibid s 9.
[158] Ibid s 10.
[159] Stephen Todd; Ursula Cheer; Cynthia Hawes; W. R. Atkin *The law of torts in New Zealand* (7th edition). Wellington: Thomson Reuters 2016 at 16.3.03(4).
[160] Ibid.

437. We note that the Harmful Digital Communications Act includes what we have heard referred to as a "mini-defamation" regime. Principle 6 states that "A digital communication should not make a false allegation."[161]

438. The law of defamation requires a plaintiff to plead that a particular statement has a defamatory meaning. We see no reason to believe that defamation would exclude the prospect of a piece of audio-visual material as being capable of carrying a defamatory meaning.

439. It will be important for the Court to assess the extent to which a defamatory meaning arises from the SMA itself and how far it comes from surrounding context and affirmative statements. This distinction is captured by conditions 1 and 3 of our framework. A video may be highly realistic such that any dissemination of it is taken to be a statement that its contents are veridical. This could be enhanced by statements about the video or other contextual indicators. This is something Courts already do in defamation law and there is a sound basis for conducting this assessment already.

[161] Harmful Digital Communications Act 2015, s 6, principle 6.

## The Media Council of New Zealand

440. The Media Council of New Zealand describes itself as "an industry self-regulatory body" set up in 1972. Its "main objective … is to provide the public with an independent forum for resolving complaints involving the newspapers, magazines and the websites of such publications and other digital media. The Council is also concerned with promoting media freedom and maintaining the press in accordance with the highest professional standards."

441. The Council's FAQs state that: "The Press Council does not deal with legal issues. These must be taken up with a lawyer. The Press Council's adjudications are based on ethical Principles. It does not recover debts or seek monetary recompense for complainants."

442. The Media Council states: "Editors have the ultimate responsibility for what appears in their publications, and for adherence to the standards of ethical journalism which the Council upholds. In dealing with complaints, the Council seeks the co-operation of editors and publishers. News bloggers and digital media are similarly required to participate responsibly." Accordingly, it does not have any coercive powers or lawful authority.

443. The Council applies a range of principles which touch upon the use of technologies of capture, manipulation and display, including dissemination and warranties of accuracy. It is possible that these principles would influence any assessment of the extent to which a media organisation belonging to the Council would be judged as having complied with relevant professional standards to the extent a piece of harmful synthetic media was created or published.

444. An interesting case study of this kind occurred recently in New Zealand when a National Party Member of Parliament, Jami-Lee Ross, provided what he said were recordings of telephone conversations between him and now Leader of the Opposition, Simon Bridges. In those recordings, Mr Bridges made comments about the ethnicity of MPs and the performance of a fellow MP. There was no indication from either party to the conversation that the phone call recordings were manipulated, however this is an option we believe could have been considered by media seeking to verify the accuracy of the recordings before publishing them.

445. Media organisations themselves will be best placed to assess developing industry practices with regard to synthetic media, however we note the following:

    a. The Wall Street Journal,[162] and Reuters,[163] have adopted deepfake detection procedures and training and practices.

    b. The work of the New York Times' Visual Investigations Unit and the techniques deployed by them in verifying the accuracy and reliability of video evidence.[164]

    c. The work of Bellingcat, an open source intelligence organisation that uses various tools to verify the reliability of audiovisual material.[165]

    d. The work of WITNESS, a human rights advocacy organisation focussed on using audio and video to document human rights abuses.[166]

[162] Francesco Marconi, Till Daldrup "How the Wall Street Journal is preparing its journalists to detect deepfakes" (15 November 2018) NiemanLab <https://www.niemanlab.org/2018/11/how-the-wall-street-journal-is-preparing-its-journalists-to-detect-deepfakes/>.

[163] Lucinda Southern "How Reuters is training reporters to spot 'deepfakes'" (26 March 2019) Digiday <https://digiday.com/media/reuters-created-a-deepfake-video-to-train-its-journalists-against-fake-news/>.

[164] See: Visual Investigations, The New York Times <https://www.nytimes.com/interactive/2018/world/visual-investigations.html>,

[165] See: Bellingcat <https://www.bellingcat.com/>.

[166] See: WITNESS <https://witness.org/> ; Sam Gregory "Deepfakes and Synthetic Media: What should we fear? What can we do?" WITNESS Blog <https://blog.witness.org/2018/07/deepfakes/> ; Sam Gregory "Deepfakes and

e. The availability of verifiable capture technologies such as TruePic,[167] and ObscuraCam,[168] that attach metadata to photos in order to enhance their verifiability and track the way that they have been augmented since the time of capture.

### Application to Council principles

446. We think the following principles from the Council's website are relevant to synthetic media artefacts and technologies, and again we emphasise that each case will be a matter of judgement in light of the individual circumstances involved:

> Principle 1: "Accuracy, Fairness and Balance. Publications should be bound at all times by accuracy, fairness and balance, and should not deliberately mislead or misinform readers by commission or omission. In articles of controversy or disagreement, a fair voice must be given to the opposition view. …"

447. The practice of seeking comment from the subject of a video will often be an important step in acknowledging and verifying the limitations of a synthetic media artefact. We think that standards of accuracy, fairness and balance may include an obligation to be conscious of the way that emerging technologies enable the manipulation of audio, visual, and audio-visual material through synthetic media technologies.

> Principle 2: "Privacy. Everyone is normally entitled to privacy of person, space and personal information, and these rights should be respected by publications. Nevertheless the right of privacy should not interfere with publication of significant matters of public record or public interest. …"

448. We have concluded that the definition of personal information under the Privacy Act will apply to synthetic media artefacts even where they are not veridical. The entitlement to privacy of person and personal information should be seen as extending to the use of a person's likeness and publication of their image within the public interest considerations referred to within the principle itself. The principle's reference to privacy of space clearly anticipates limitations on the use of Category 1 capture technologies, as well as the display of Category 3 SMAs through dissemination technologies.

> Principle 4: "Comment and Fact. A clear distinction should be drawn between factual information and comment or opinion. An article that is essentially comment or opinion should be clearly presented as such. Material facts on which an opinion is based should be accurate."

449. Synthetic media technologies raise the prospect that audiovisual material may as much be an expression of opinion as an expression of fact, to the extent they can be altered. Media organisations should be aware of the extent to which synthetic media technologies enable deceptive manipulation of a video beyond standard editing and capture techniques with which they will already be familiar, such as inclusion and exclusion of information from a scene or removal of parts of an audio or visual record using editing software. This also presents the possibility that audio-visual material may not be as reliable as previously thought as an indication of the accuracy of material facts. Editors will already be conscious of the manipulation of static images through synthetic media technologies, however this should also be applied within reason to video and audio technologies.

> Principle 5: "Columns, Blogs, Opinion and Letters. Opinion, whether newspaper column or internet blog, must be clearly identified as such unless a column, blog or other expression of opinion is widely understood to consist largely of the writer's own opinions. Though requirements for a foundation of

Synthetic Media: Survey of Solutions against Malicious Usages" WITNESS Blog <https://blog.witness.org/2018/07/deepfakes-and-solutions/>.

[167] See: Truepic <https://blog.witness.org/2018/07/deepfakes-and-solutions/>.

[168] See: <https://guardianproject.info/apps/obscuracam/>.

> fact pertain, with comment and opinion balance is not essential. Cartoons are understood to be opinion. …"

450. There is a question about the extent to which a highly photorealistic artefact may cease to be seen as a cartoon. There is a real prospect that political leaders' faces and voices can be emulated through synthetic media technologies for satirical effect. As above, we believe there is also the prospect that synthetic media products of a certain kind should be viewed as evidence of opinion rather than fact, depending on the extent to which there is no real relationship between the artefact produced and the scene that it purports to have captured (see Condition 1 of the framework).

> Principle 6: "Headlines and Captions. Headlines, sub-headings, and captions should accurately and fairly convey the substance or a key element of the report they are designed to cover."

451. It will be important to acknowledge any doubt about the reliability of synthetic media artefacts with significant news value in headlines and captions. Deepfakes in particular are frequently eye-grabbing and have significant "click-bait" appeal.

> Principle 8: "Confidentiality. Publications have a strong obligation to protect against disclosure of the identity of confidential sources. They also have a duty to take reasonable steps to satisfy themselves that such sources are well informed and that the information they provide is reliable. Care should be taken to ensure both source and publication agrees over what has been meant by "off-the-record"."

452. There is a prospect that confidential sources could produce synthetic media artefacts in order to persuade a publication that they are well-informed and the information they provide is reliable. There will be a question of proportionality here, but in cases of significant news value and social consequences, it is worth considering the extent to which persuasive media artefacts may have been influenced by SMTs. The question of whether "reasonable steps" have been taken may be influenced by increasing awareness among the media about the capacities of SMTs.

> Principle 9: "Subterfuge. Information or news obtained by subterfuge, misrepresentation or dishonest means is not permitted unless there is an overriding public interest and the news or information cannot be obtained by any other means.

453. Synthetic media technologies could be produced in a way that facilitates subterfuge, misrepresentation or dishonest means in obtaining information or news. For example, an audio recording could be manufactured of someone ostensibly giving a source authority to divulge certain information. Again, we do not wish to speculate too far given the likelihood that hypothetical situations become unrealistic, however the deceptive capacity of synthetic media artefacts is an important factor to consider.

> Principle 11: "Photographs and Graphics. Editors should take care in photographic and image selection and treatment. Any technical manipulation that could mislead readers should be noted and explained. Photographs showing distressing or shocking situations should be handled with special consideration for those affected."

454. This principle explicitly references the capacities of Category 2 manipulation technologies and the need to explicitly acknowledge the extent to which Category 3 display may have been manipulated. It is not clear how far the reference to photography or images should also be taken to refer to compilations of photographs or images constituting a video. This should be clarified.

455. We think this principle also implies an obligation on media organisations to consider the extent to which images they have selected may have been able to be manipulated by other parties.

456. Further, we note the way that media organisations elected to distribute excerpts or links to the terrorist shooter's video material from the Christchurch shootings in March 2019. We think this shows a willingness among some media organisations to weigh news value and public interest over the need to have special consideration for those affected by distressing or shocking audio-visual material. Clearly, the Christchurch shooter was aware that media outlets would take this course of

action,[169] demonstrating that bad actors are aware of institutional influences on media organisations and incentives to publish audio-visual material in a competitive global news environment. Media organisations will need to proceed with extra caution in light of increasing awareness about the capacity for bad actors to produce extremely newsworthy audio-visual material that may be partially or completely inaccurate or deceptive.

> Principle 12: "Corrections. A publication's willingness to correct errors enhances its credibility and, often, defuses complaint. Significant errors should be promptly corrected with fair prominence. In some circumstances it will be appropriate to offer an apology and a right of reply to an affected person or persons."

457. The ability to effectively correct a misleading publication of synthetic media artefacts, or the publication of misleading synthetic media artefacts will be severely challenged in light of the rapid and widespread nature of dissemination technologies like the internet and social media. We think this should increase media organisations' obligations to exercise special caution before publishing a SMA that may have been manipulated.

[169] Katie Kenny "Q+A: Troll hunter Ginger Gorman on the Christchurch mosque shootings and cyberhate" Stuff.co.nz (3 April 2019) <https://www.stuff.co.nz/national/christchurch-shooting/111743226/qa-troll-hunter-ginger-gorman-on-the-christchurch-mosque-shootings-and-cyberhate>: "He also employed a technique called 'media f.....', which is a tactic where [terrorists] essentially co-opt the media into proliferating their messages. He certainly succeeded in that. I know The Daily Mail published his manifesto in full. The document is full of media bait. Through it, [the gunman] is signalling to his white supremacy community."

## Human Rights Act 1993

458. A piece of synthetic media could be used to inflict harm against a group or community. New Zealand is accelerating a review of its hate speech laws in light of the Christchurch shootings. Clearly, our discussion of this enactment could be covered under criminal harms in the next section of the report at Part 3. Nothing should be inferred in terms of substantive commentary from its location in this part of the report.

459. These provisions were recently examined by both the Human Rights Review Tribunal and the High Court in *Wall v Fairfax* [2017] NZHRRT 17 and *Wall v Fairfax* [2018] NZHC 104 in relation to a cartoon published by the defendant. The shortcomings of the provision and the history of their interpretation, domestically and internationally, are extensively canvassed in those decisions.

460. We do not examine the provisions in detail other than noting that we think it is possible for a piece of synthetic media to be caught by these provisions depending on the facts of the case.

461. Unlike a cartoon, a highly photorealistic SMA may not be obviously fabricated, and therefore it may appear as a reliable record of factual events rather than an expression of opinion. As such, it may incite hostility in a way anticipated by the Human Rights Act in the hands of the creator, but only appear as factual material in the hands of subsequent disseminators. This may also make it difficult to establish intent where someone can reasonably rely on the apparent accuracy of the SMA.

462. The offence of inciting racial disharmony creates a definition of "written matter" referring to "words", although these words and written matter can be broadcast by means of radio or television. An argument could be made that this limits the application of the section to the use of actual words within a piece of synthetic media, however the situation where something is capable of "inciting racial disharmony" but includes no "words" is likely to be rare. Again, we avoid speculating and note that a Court will regularly ascertain the meaning of a synthetic media artefact in its context and the circumstances.

> 131 Inciting racial disharmony
> (1) Every person commits an offence and is liable on conviction to imprisonment for a term not exceeding 3 months or to a fine not exceeding $7,000 who, with intent to excite hostility or ill-will against, or bring into contempt or ridicule, any group of persons in New Zealand on the ground of the colour, race, or ethnic or national origins of that group of persons,—
> (a) publishes or distributes written matter which is threatening, abusive, or insulting, or broadcasts by means of radio or television words which are threatening, abusive, or insulting; or
> ...
> being matter or words likely to excite hostility or ill-will against, or bring into contempt or ridicule, any such group of persons in New Zealand on the ground of the colour, race, or ethnic or national origins of that group of persons.
> (2) For the purposes of this section, publishes or distributes and written matter have the meaning given to them in section 61.

463. Per s 131(2), many of the defined terms in s 131 are drawn from s 61:

> 61 Racial disharmony
> (1) It shall be unlawful for any person—
> (a) to publish or distribute written matter which is threatening, abusive, or insulting, or to broadcast by means of radio or television or other electronic communication words which are threatening, abusive, or insulting; or
> ...
> (c) to use in any place words which are threatening, abusive, or insulting if the person using the words knew or ought to have known that the words were reasonably likely to be published in a newspaper, magazine, or periodical or broadcast by means of radio or television,—

being matter or words likely to excite hostility against or bring into contempt any group of persons in or who may be coming to New Zealand on the ground of the colour, race, or ethnic or national origins of that group of persons.

(2) It shall not be a breach of subsection (1) to publish in a newspaper, magazine, or periodical or broadcast by means of radio or television or other electronic communication a report relating to the publication or distribution of matter by any person or the broadcast or use of words by any person, if the report of the matter or words accurately conveys the intention of the person who published or distributed the matter or broadcast or used the words.

(3) For the purposes of this section,—

**newspaper** means a paper containing public news or observations on public news, or consisting wholly or mainly of advertisements, being a newspaper that is published periodically at intervals not exceeding 3 months

**publishes or distributes** means publishes or distributes to the public at large or to any member or members of the public

**written matter** includes any writing, sign, visible representation, or sound recording.

464. We think the definition of written matter is broad enough to anticipate emerging audiovisual technologies under "visible representation" and "sound recording", but note it is an extending definition that is likely to be interpreted broadly rather than narrowly given the purpose of the Act. We note the reference to "electronic communication" in s 61.

465. Section 63 includes explicit reference to a person's use of language or visual material in relation to racial harassment. Subsection 2(k) was added by the Harmful Digital Communications Act and is taken to be a reference to internet fora. Notably, s 63 does not follow the same drafting around "written matter" as ss 61 and 131 and instead adopts an explicit focus on spoken language and visual material.

63 Racial harassment

(1) It shall be unlawful for any person to use language (whether written or spoken), or visual material, or physical behaviour that—

(a) expresses hostility against, or brings into contempt or ridicule, any other person on the ground of the colour, race, or ethnic or national origins of that person; and

(b) is hurtful or offensive to that other person (whether or not that is conveyed to the first-mentioned person); and

(c) is either repeated, or of such a significant nature, that it has a detrimental effect on that other person in respect of any of the areas to which this subsection is applied by subsection (2).

(2) The areas to which subsection (1) applies are—

...

(k) participation in fora for the exchange of ideas and information.

466. These definitions effectively focus on harms of content as well as harms of dissemination. The reference to repeated or significant behaviour in s 63(1)(c) reflects a similar focus in the Harassment Act and Harmful Digital Communications Act. The addition of ss (2)(k) also indicates a legislative recognition about the interrelationship between the Human Rights and Harmful Digital Communications Acts.

467. We think that condition 2 is likely to be difficult for assigning intent, depending on the content of the video. In some cases, synthetic media will be harmful enough that intent can be inferred. Difficult cases are likely to arise where an SMA is highly deceptive in terms of Condition 1 meaning that it can be shared without the requisite intent, but the intention of the creator may closely match the wording of the section and therefore be unlawful.

# Interpersonal harms that are Criminal or approaching criminal

## Crimes Act 1961

468. We identify certain key offences which we think will be relevant to the main uses of synthetic media for malicious purposes.

469. We conclude that the use of synthetic media for otherwise criminal purposes is likely to be covered by a range of offences under the Crimes Act 1961. Further, not all offences will require evidence that a SMA is "fake" or does not show what it purports to show, thereby avoiding one of the key concerns about the rise of "deepfake" SMAs.

### Inducement or threats

470. Synthetic media technologies or artefacts could generate harmful impacts by threats to create or disclose it with harmful content, or to make representations about the circumstances in which the SMA was created or captured. We think offences of inducement are well covered by the Act and we do not see any need for reform of the law itself.

471. We think that the language of s 237 of the Act in relation to blackmail will not exclude any attempt to use a SMA. We do not think that the truthfulness or otherwise of the "something" that is disclosed is material. Clearly the making of a false allegation that is difficult to disprove to cause someone to act in accordance with someone's will is just as likely to lead to harm.

> 237 Blackmail
> (1) Every one commits blackmail who threatens, expressly or by implication, to make any accusation against any person (whether living or dead), to disclose something about any person (whether living or dead), or to cause serious damage to property or endanger the safety of any person with intent—
> (a) to cause the person to whom the threat is made to act in accordance with the will of the person making the threat; and
> (b) to obtain any benefit or to cause loss to any other person.
> (2) Every one who acts in the manner described in subsection (1) is guilty of blackmail, even though that person believes that he or she is entitled to the benefit or to cause the loss, unless the making of the threat is, in the circumstances, a reasonable and proper means for effecting his or her purpose.
> (3) In this section and in section 239, benefit means any benefit, pecuniary advantage, privilege, property, service, or valuable consideration.

472. We have also examined the Crimes Act to identify offences that involve "inducement". We note that s 80 of the Crimes Act makes it an offence to attempt to induce or compel someone to take an oath or engagement to commit an offence. Similarly, s 98 deals in the inducement of slavery, s 98AA deals in inducement of a person under the age of 18 years for sexual exploitation, removal of body parts, or engagement in forced labour.

473. Similarly, s 129A of the Act makes it an offence liable to imprisonment for 14 years if they induce consent to sexual conduct by making a threat to make an accusation or disclosure (whether true or false) about misconduct that is likely to damage the reputation of any person.

### Incitement

474. Offences directed to incitement can cover the use of SMAs where the consequence is that an offence is committed or attempted. This can sound extreme, however there are increasing media reports of social media being suspended by governments in response to concerns about incitement

to violence.[170] The SMA may also require affirmative statements beyond the content itself (for example a call to action), however this will be a question to be examined on the facts at hand examining the Category 3 artefact and the statements attached to its dissemination.

475. Section 179 of the Crimes Act makes it an offence to incite, counsel, or procure any person to commit suicide regardless of the means involved, "if that person … attempts to commit suicide in consequence thereof". They are still able to be imprisoned if their actions do not lead that person committing or attempting suicide, although imprisonment cannot exceed 3 years. Section 174 creates a similar regime in respect of incitement to commit murder, even if the murder is not committed.

476. Section 73 makes treason an offence if synthetic media is used to incite a force to invade New Zealand. This is perhaps extreme, but not implausible in today's international security environment, particularly where world leaders or political dissidents could be depicted or mimicked aurally in a way that incites international disputes.

477. Section 70 of the Act anticipates a situation where synthetic media is used to incite or procure an offence even where the offence is committed in a different manner, or is a different offence.

478. We also note that incitement to commit an offence can lead to a person being a party to that offence pursuant to s 66 of the Crimes Act.

## Deception

479. We think that perjury at s 108 of the Act merits close treatment:

> 108 Perjury defined
> (1) Perjury is an assertion as to a matter of fact, opinion, belief, or knowledge made by a witness in a judicial proceeding as part of his or her evidence on oath, whether the evidence is given in open court or by affidavit or otherwise, that assertion being known to the witness to be false and being intended by him or her to mislead the tribunal holding the proceeding.
> ...
> (3) Every person is a witness within the meaning of this section who actually gives evidence, whether he or she is competent to be a witness or not, and whether his or her evidence is admissible or not.
> (4) Every proceeding is judicial within the meaning of this section if it is held before any of the following tribunals, namely:
> (a) any court of justice:
> (b) the House of Representatives or any Committee of that House:
> (c) any arbitrator or umpire, or any person or body of persons authorised by law to make an inquiry and take evidence therein upon oath:
> (d) any legal tribunal by which any legal right or liability can be established:
> (e) any person acting as a court or tribunal having power to hold a judicial proceeding:
> ...

480. This is an important protection against the knowing use of synthetic media to mislead in a judicial proceeding, including the House of representatives or any other actor listed at s 108(4).

481. There are very few situations where audio-visual information will be admitted as evidence without a witness attesting to its reliability, except in situations where both parties agree to its admission. Witnesses should be asked to explicitly confirm their knowledge as to whether synthetic media has been manipulated and to what extent, including in any ways that might affect its reliability as a Category 3 display of an SMA produced via a Category 1 capture technology.

482. We also note that s 113 of the Crimes Act makes it an offence punishable by 7 years imprisonment to intentionally mislead any tribunal holding a judicial proceeding to which s 108 applies, by

[170] For instance, Sri Lanka, Papua New Guinea, India, Myanmar.

fabricating evidence by any other means than perjury. This would account for any situation where no assertion is made as required by s 108 in addition to the fabricated evidence caught by s 113.

483. Section 249 of the Crimes Act has broad drafting that we think makes it an offence to use a Category 2 technology to dishonestly obtain property or cause loss. It is not necessary that the computer system is accessed through dishonest means – for example by hacking.

> 249 Accessing computer system for dishonest purpose
> (1) Every one is liable to imprisonment for a term not exceeding 7 years who, directly or indirectly, accesses any computer system and thereby, dishonestly or by deception, and without claim of right,—
> (a) obtains any property, privilege, service, pecuniary advantage, benefit, or valuable consideration; or
> (b) causes loss to any other person.
> (2) Every one is liable to imprisonment for a term not exceeding 5 years who, directly or indirectly, accesses any computer system with intent, dishonestly or by deception, and without claim of right,—
> (a) to obtain any property, privilege, service, pecuniary advantage, benefit, or valuable consideration; or
> (b) to cause loss to any other person.
> (3) In this section, deception has the same meaning as in section 240(2).

484. The section effectively criminalises use of a computer that consequently leads to obtaining by deception. This could include the manipulation of images – even with consent and lawful authority – so long as the product is used to obtain by deception.

485. We think there is no reason to believe that common offences involving causing loss or obtaining by deception would not apply to synthetic media.

486. Section 217 of the Act defines several key terms, including "dishonestly", "document", and "obtain":

> **dishonestly**, in relation to an act or omission, means done or omitted without a belief that there was express or implied consent to, or authority for, the act or omission from a person entitled to give such consent or authority
> **document** means a document, or part of a document, in any form; and includes, without limitation,—
> (a) any paper or other material used for writing or printing that is marked with matter capable of being read; or
> (b) any photograph, or any photographic negative, plate, slide, film, or microfilm, or any photostatic negative; or
> (c) any disc, tape, wire, sound track, card, or other material or device in or on which information, sounds, or other data are recorded, stored (whether temporarily or permanently), or embodied so as to be capable, with or without the aid of some other equipment, of being reproduced; or
> (d) any material by means of which information is supplied, whether directly or by means of any equipment, to any device used for recording or storing or processing information; or
> (e) any material derived, whether directly or by means of any equipment, from information recorded or stored or processed by any device used for recording or storing or processing information
> **obtain**, in relation to any person, means obtain or retain for himself or herself or for any other person.

487. We think the framing of "dishonestly" in relation to consent is notable and engages questions around implied or express authority to manipulate data or disclose what we have elsewhere concluded is personal information governed by the Privacy Act.

488. It seems clear that a piece of synthetic media can be a document for the purposes of the Crimes Act. We think paras (b) and (e) of the definition of document describe a piece of synthetic media without much gloss needing to be applied, if any.

489. We set out s 240 of the Crimes Act to enable close analysis:

240 Obtaining by deception or causing loss by deception

(1) Every one is guilty of obtaining by deception or causing loss by deception who, by any deception and without claim of right,—
  (a) obtains ownership or possession of, or control over, any property, or any privilege, service, pecuniary advantage, benefit, or valuable consideration, directly or indirectly; or
  (b) in incurring any debt or liability, obtains credit; or
  (c) induces or causes any other person to deliver over, execute, make, accept, endorse, destroy, or alter any document or thing capable of being used to derive a pecuniary advantage; or
  (d) causes loss to any other person.

(1A) Every person is liable to imprisonment for a term not exceeding 3 years who, without reasonable excuse, sells, transfers, or otherwise makes available any document or thing capable of being used to derive a pecuniary advantage knowing that, by deception and without claim of right, the document or thing was, or was caused to be, delivered, executed, made, accepted, endorsed, or altered.

(2) In this section, deception means—
  (a) a false representation, whether oral, documentary, or by conduct, where the person making the representation intends to deceive any other person and—
    (i) knows that it is false in a material particular; or
    (ii) is reckless as to whether it is false in a material particular; or
  (b) an omission to disclose a material particular, with intent to deceive any person, in circumstances where there is a duty to disclose it; or
  (c) a fraudulent device, trick, or stratagem used with intent to deceive any person.

490. We think s 240(1A) also criminalises the act of providing SMT services in order to manipulate SMA for dishonest purposes. For example, if I were to sell my services to manipulate a video for a client knowing that the video was to be used to gain a favourable impression from someone dishonestly, then I may fall foul of (1A). We note that s 251 of the Crimes Act would also criminalise the provision of software intended to be used in the commission of an offence (including by deception or inducement elsewhere referred to in this report).

491. The definition of "deception" at s 240(2) is also of interest because of the way that it corresponds to our articulation of Condition 1 of our framework. A SMA can be misleading: (a) by explicit false representation or statement with intention to deceive or being reckless as to deception; (b) by failing to correct a mistaken assumption in a situation where that is likely to arise; or (c) by the use of a "fraudulent device, trick or stratagem". We doubt that recourse to (c) is necessary, however we note that the use of many SMT would be indistinguishable to a naïve consumer from something from a magician's toolbox.

492. In particular, in relation to s 240(2)(a)(ii) and (2)(b), we note that deception can include "an omission to disclose a material particular, with intent to deceive any person, in circumstances where there is a duty to disclose it". We think that this is likely to put an obligation on people in certain circumstances, where they know an SMA is highly deceptive in terms of Condition 1, to disclose the extent of the manipulation that has occurred.

493. Section 241 of the Crimes Act creates gradations of prison terms depending on the financial value of the loss or gain resulting of up to seven years or as little as three months.

494. It is an offence pursuant to s 242 of the Act for a person to make or publish a false statement about an incorporated or unincorporated body with intent to induce any person to acquire or not acquire financial product within the meaning of the Financial Markets Conduct Act 2013, or to cause loss or deceive any person, or to induce any person to entrust or advance property to any other person. For the purposes of the section, recklessness as to the falsity of the statement in a material particular is sufficient.

495. Section 258 of the Act is an offence which we think is closely oriented to the use of Category 2 technologies. It makes it an offence to alter or reproduce a document with intent to deceive. Given our conclusion that synthetic media will be a document within the definition of the Crimes Act at s

217, this offence would be engaged against manipulation of synthetic media artefacts with intent to obtain by deception or cause loss.

> 258 Altering, concealing, destroying, or reproducing documents with intent to deceive
> (1) Every one is liable to imprisonment for a term not exceeding 10 years who, with intent to obtain by deception any property, privilege, service, pecuniary advantage, benefit, or valuable consideration, or to cause loss to any other person,—
> (a) alters, conceals, or destroys any document, or causes any document to be altered, concealed, or destroyed; or
> (b) makes a docume nt or causes a document to be made that is, in whole or in part, a reproduction of any other document.
> (2) An offence against subsection (1) is complete as soon as the alteration or document is made with the intent referred to in that subsection, although the offender may not have intended that any particular person should—
> (a) use or act upon the document altered or made; or
> (b) act on the basis of the absence of the document concealed or destroyed; or
> (c) be induced to do or refrain from doing anything.
> (3) Every person is liable to imprisonment for a term not exceeding 3 years who, without reasonable excuse, sells, transfers, or otherwise makes available any document knowing that—
> (a) the document was altered, concealed, or made, in whole or in part, as a reproduction of another document; and
> (b) the document was dealt with in the manner specified in paragraph (a) with intent to—
> (i) obtain any property, privilege, service, pecuniary advantage, benefit, or valuable consideration; or
> (ii) cause loss to any other person.

496. We note that s 258(2) makes an offence against sub (1) complete "as soon as the alteration or document is made with the intent referred to in that subsection, although the offender may not have intended that any particular person should—(a) use or act upon the document altered or made; or (b) act on the basis of the absence of the document concealed or destroyed; or (c) be induced to do or refrain from doing anything."

497. The offence at s 258(3) further criminalises the sale, transfer or making available of a synthetic media artefact without reasonable excuse with the knowledge that it was altered with intent to obtain advantage or cause loss.

498. Section 259 makes the use of an altered document with intent to deceive an imprisonable offence, even if the document was altered outside New Zealand.

> 259 Using altered or reproduced document with intent to deceive
> (1) Every one is liable to imprisonment for a term not exceeding 10 years who, knowing any document to have been made or altered in the manner and with the intent referred to in section 258, with intent to obtain by deception any property, privilege, service, pecuniary advantage, benefit, or valuable consideration, or to cause loss to any other person,—
> (a) uses, or deals with, or acts upon, the document; or
> (b) causes any person to use or deal with, or act upon, the document.
> (2) For the purposes of this section, it does not matter that the document was altered or made outside New Zealand.

499. Lastly we turn to extreme cases: s 160 of the Crimes Act defines culpable homicide as including killing of any person "by causing that person by threats or fear of violence, or by deception, to do an act which causes his or her death". Culpable homicide is either murder or manslaughter per s 160(3). Where, in terms of condition 1 of the framework, someone is deceived in a way that causes them to do an act which causes their death, they may be charged with murder of manslaughter.

## Intimate visual recordings and non-consensual pornography

500. The use of SMT for pornographic purposes without the consent of the subject is one of the more significant public policy challenges raised in public discussion. We have noted how the "deepfake"

moniker arose from the name of a Reddit user who was distributing non-consensual pornography depicting prominent actresses.

501. We understand that child sexual exploitation material was made online using the faces of recognisable victims using technologies such as photoshop before the advent of deepfake technology and the use of neural networks has merely accelerated the digital manipulation process.

502. There is a significant volume of policy and extra-legal material that led to the inclusion of offences dealing with intimate visual recordings. We have not found any indication that the offence was originally intended to deal with deepfake pornography, i.e. where the role of capture technologies is limited in its production.

503. There are two approaches to whether the offence can be applied to SMAs. We are conscious that our definition of "SMA" is broad and so here we are referring to SMT such as "deepfakes", where the digital data collected by capture technologies is merged so as to create a new Category 3 product which is non-veridical – it shows something that never happened – but without any other correction would lead someone to believe that it was an authentic product of a single instance of capture technology.

504. The first argument goes that the policy history of the provision should be of limited relevance when considering the plain and ordinary meaning of it. The key thing is that someone could look at the offence as drafted and have doubts about the lawfulness of synthesising pornography. The Interpretation Act 1999 dictates that text and statutory purpose be the guide to interpretation. We conclude that a prosecutor would be justified in using this section to prosecute the creation of intimate visual recordings using SMTs.

505. Section 216G defines an intimate visual recording and is drafted as follows. We note the reference in s 216G(1)(a) to terminology similar to the "reasonable expectation of privacy" concept discussed in relation to the Privacy Act and privacy torts. We also note that, in relation to other enactments, we have broadly concluded that a "recording" in most cases does not explicitly require the use of a Category 1 capture technology. The exception to this is the Copyright Act.

> 216G Intimate visual recording defined
> (1) In sections 216H to 216N, intimate visual recording means a visual recording (for example, a photograph, videotape, or digital image) that is made in any medium using any device without the knowledge or consent of the person who is the subject of the recording, and the recording is of—
> (a) a person who is in a place which, in the circumstances, would reasonably be expected to provide privacy, and that person is—
> (i) naked or has his or her genitals, pubic area, buttocks, or female breasts exposed, partially exposed, or clad solely in undergarments; or
> (ii) engaged in an intimate sexual activity; or
> (iii) engaged in showering, toileting, or other personal bodily activity that involves dressing or undressing; or
> (b) a person's naked or undergarment-clad genitals, pubic area, buttocks, or female breasts which is made—
> (i) from beneath or under a person's clothing; or
> (ii) through a person's outer clothing in circumstances where it is unreasonable to do so.
> (2) In section 216H, intimate visual recording includes an intimate visual recording that is made and transmitted in real time without retention or storage in—
> (a) a physical form; or
> (b) an electronic form from which the recording is capable of being reproduced with or without the aid of any device or thing.

506. One view is that the wording "made in any medium using any device" could encompass the use of digital manipulation technologies, including to synthesise the impression that a Category 1 capture technology was deployed and that the final Category 3 display generated has a high degree of

similarity between light captured and light displayed. If "recording" is not interpreted as requiring the capture of light or sound energy, then this would also support this interpretation. The recording is described as being "of" someone in a situation with a reasonable expectation of privacy, which could merely mean depicting someone in that setting, without further reference to Category 1 technologies.

507. Further, if we take the stance – as we do – that the difference between a deepfake and a digital photograph or video is one of degree rather than kind, then there is little basis for distinguishing between the synthetic processes involved in "making" a digital video without intentional or impermissible manipulation of data, and the synthetic processes involved in intentionally manipulating that data in a way that may be impermissible.

508. In terms of the other provisions we have analysed, the offence is an interesting one because it does explicitly call attention to the process by which the SMA is created – or "made" – unlike other enactments such as the Privacy Act and Harmful Digital Communications Act. However, it uses the terminology of a "recording", which is used in a range of other enactments, although particularly in the Copyright Act 1994, which acknowledges that copyright can arise in synthetic media products such as films with heavy digital effects aspects.

509. Section 216G(1)(a) complicates interpretation further. In particular, in deepfake pornography – as one example – one person's face is digitally inserted onto another person's body. In that respect, two people are depicted, and only one of those people have their genitals displayed. On one reading, this would exclude deepfake pornography from the definition.

510. A counter argument is to note the transition in drafting between s 216G(1) and paras (a) and (b): sub (1) refers to "the person who is the subject of the recording" (using the definite article "the", indicating reference to a specific person); paras (a) and (b) then transition to reference to "a person" (indefinite article, meaning any person). As a result, the section could be taken to be referring to two different people on close analysis. In that sense, it would perfectly anticipate the way that more than one person is depicted in a synthetic media artefact.

511. We think this section provides an excellent use case for the application of our framework. The section is situated in Part 9A of the Act, being "Crimes against Personal Privacy". As referred to in *Hosking v Runting*, and *C v Holland*, and their analysis of the NZBORA above, privacy can be both spatial and informational, related to identity. The reference to the International Covenant on Civil and Political Rights in clause 3 of the Privacy Bill also supports this approach.

512. The question, at a policy level, is whether the section is intended to criminalise the use of:

    a. Category 1 capture technologies, that capture light or sound in an intimate setting without the knowledge or consent of the subject;

    b. Category 2 manipulation technologies, that deal in manipulations of personal data about an individual without their consent;

    c. Category 3 technologies of display directed at the misrepresentations of authenticity pursuant to conditions 1 and 3.

    d. It also illustrates the issues caused by condition 2: the non-linear process by which SMAs are created. It is entirely possible that an intimate visual recording was captured and shared by consent, and only subsequently used in an impermissible or non-consensual manner, perhaps by an entirely different actor.

513. Are these harms the same or different? Which harms does the section intend to mitigate? We think this should be clarified, either by action in the Courts or by Parliament.

514. We do not think any answer is supplied by the offences at ss 216H-216N.

    a. Section 216H prohibits the "making" of an IVR.

    b. Section 216I prohibits possession of an IVR. Significantly it applies criteria either of actual knowledge, or recklessness, when there is an intention to publish, export or sell it. It does not criminalise possession with recklessness, mere possession requires actual knowledge.

    c. Section 216J creates explicit prohibitions on publishing, importing, exporting or selling IVR, although this does not indicate whether entirely synthesised IVR are intended to be caught. We note that s 216J defines "publishes" to include display by any means, and distribution, including conveying or storage electronically.

> **publishes** means any of the following:
> (a) displays by any means:
> (b) sends to any person by any means:
> (c) distributes by any means:
> (d) conveys by electronic medium:
> (e) stores electronically in a way that is accessible by any other person or persons

515. We do not think that these sections assist matters in terms of the extent to which highly manipulated or generative media is caught by the definition at s 216G.

516. We briefly note ss 216A-216F and the way that they regulate the use of interception devices and disclosure of information obtained by their use, as well as the sale of them. We think this indicates perhaps a normative basis for saying that dealing in highly synthesised private material, which would otherwise be an offence to capture, may also justify legal intervention, however we think this is simply a similar debate to be had with respect to the intimate visual recording offences.

### Conclusion on Crimes Act

517. The Crimes Act is broadly drafted in a media neutral way that grants significant latitude to a criminal court to find that new forms of emerging audiovisual media can be used to commit much older forms of criminal activity. This broad latitude is countered by procedural restraints, including the need for charge to be laid, a criminal trial and to the higher criminal standard of proof (beyond a reasonable doubt). Policymakers can gain some comfort from the drafting of the Crimes Act.

518. Section 216G of the Act requires revision to assess the extent to which it is intended to criminalise the use of category 2 or category 1 technologies. There is an argument to be made that the harms of capture are distinct from the harms of content or dissemination, but this argument would benefit from the experience and expertise of people working in the area of image-based abuse. Netsafe pays close attention to developments in this area. To an extent, victims could also have recourse to the Privacy Act for a civil remedy in appropriate cases, to the extent s 56(2) of the Privacy Act removes the exemption for domestic activities.

519. Another point to note about the Crimes Act is the way that it allows for charging of attempted crimes, as well as parties to crimes. As a result, even if someone uses an unpersuasive piece of synthetic media to attempt a crime of deception, that can be subject to criminal sanction.

## Harmful Digital Communication Act 2015

520. The creation of the Harmful Digital Communications Act 2015 (HDCA) occurred in the context of vast increases in the frequency of virtual interactions as well as increasing democratization of the tools for producing digital information like video recordings. The HDCA creates legal powers and deterrent effects in relation to intimate visual recordings and revenge pornography, including more generic harmful interactions through digital technologies. This is reflected in the purpose of the Act:

> 3 Purpose
> The purpose of this Act is to—
> (a) deter, prevent, and mitigate harm caused to individuals by digital communications; and
> (b) provide victims of harmful digital communications with a quick and efficient means of redress.

521. This dual purpose reflects the particular characteristics of digital communications, which differ from analogue communications in the speed and 'distance' with which they may spread, as well as their permanence. In short spaces of time and with little effort, digital communications can be published widely and often anonymously, shared by disparate actors across multiple virtual domains and legal jurisdictions. This increases the potential harms of digital communications while generating questions of which actors are accountable and to what extent.

522. Within this context, it is unremarkable to proceed from the position that the HDCA is likely to be operative when synthetic media is communicated. It is primarily directed to harms of dissemination as may arise in relation to Condition 3, and to a lesser extent content of an SMA displayed per Category 3. First, it is axiomatic that all synthetic media is digital according to our framework, and second, its dissemination often will occur through digital means. It is also reasonable to conclude that while artefacts like "deepfakes" were not expressly contemplated in the formulation of the Act they nonetheless fit within Parliament's general intention to address a wide range of harms which may result through a digital medium but are experienced in an offline environment.

523. The s 4 interpretation corroborates the Act's application to synthetic media, including but not limited to deepfakes. It includes the following definition, which includes terms also used in the Copyright Act and others canvassed in this report.

> digital communication—
> (a) means any form of electronic communication; and
> (b) includes any text message, writing, photograph, picture, recording, or other matter that is communicated electronically

524. In effect, this definition covers any conceivable present or future synthetic media in a scenario where that media it is communicated through the use of digital technologies, like the internet or multimedia messaging services.

525. Furthermore, the s 4 definition of "posts a digital communication" is equally broadly drafted:

> posts a digital communication—
> (a) means transfers, sends, posts, publishes, disseminates, or otherwise communicates by means of a digital communication—
> (i) any information, whether truthful or untruthful, about the victim; or
> (ii) an intimate visual recording of another individual; and
> (b) includes an attempt to do anything referred to in paragraph (a)
>
> [underline emphasis added]

526. It is beyond reasonable dispute that, in most fact patterns, an act in which synthetic media is communicated will qualify as a posting a digital communication.

527. The next element in any assessment under the HDCA turns to the element of harm and its threshold. The definition offered by s 4 is "harm means serious emotional distress."

528. A range of relevant factors for considering whether a post would cause harm are set out at s 22(2). Amongst these are factors including whether the digital communication was repeated, and whether is true or false. Both of these factors are noteworthy for the prospect of synthetic media. They anticipate the possibility that artefacts like deepfakes could cause harm without being innately offensive content. For instance, posting a new deepfake every day depicting the same person presumably may reach a threshold of nuisance so as to be harmful to the individual depicted, even if the content of each video is otherwise benign.

529. The Act also sets out at s 6 a set of ten Communication Principles which must be taken into account in the course of any determination by a fact-finder. The principles are listed at s 6 and could easily apply to the use of generative or non-veridical synthetic media:

| | |
|---|---|
| Principle 1 | A digital communication should not disclose sensitive personal facts about an individual. |
| Principle 2 | A digital communication should not be threatening, intimidating, or menacing. |
| Principle 3 | A digital communication should not be grossly offensive to a reasonable person in the position of the affected individual. |
| Principle 4 | A digital communication should not be indecent or obscene. |
| Principle 5 | A digital communication should not be used to harass an individual. |
| Principle 6 | A digital communication should not make a false allegation. |
| Principle 7 | A digital communication should not contain a matter that is published in breach of confidence. |
| Principle 8 | A digital communication should not incite or encourage anyone to send a message to an individual for the purpose of causing harm to the individual. |
| Principle 9 | A digital communication should not incite or encourage an individual to commit suicide. |
| Principle 10 | A digital communication should not denigrate an individual by reason of his or her colour, race, ethnic or national origins, religion, gender, sexual orientation, or disability. |

530. Between them, the principles canvass the majority of imaginable digital communications with harmful characteristics. It is unnecessary for us to speculate on the range of fact scenarios where an artefact like a deepfake might be subject to a claim of causing emotional harm. What is important, and what can be said in advance, is that the HDCA is likely to apply in any scenario where synthetic media is communicated. Whether or not the harmful effects of that communication meet the test for "serious emotional distress" is an evidential question to be determined on the facts of the case.

531. The HDCA will be engaged only in situations where a communication has occurred. It does not assist with fact scenarios where the mere existence or private consumption of synthetic media is at issue. For instance, the HDCA is not engaged where a person generates a digital asset of another person which may be animated for any purpose, including purposes that would be offensive to the reasonable and ordinary person. Without communication, the mere existence of the digital asset that resembles another individual does not attract the attention of the HDCA. Even if the artefact in question was objectively harmful, for instance depicting an individual in a highly compromising or offensive position, the "harms" the HDCA is interested in arise from communication (dissemination in our framework), and not from the artefact itself or its private consumption by the user or creator. This does not preclude the possibility that some other statutory or tortious mechanism might be pursued in such a scenario, it simply precludes does not engage the HDCA.

532. Despite the emphasis on communication and harms of dissemination under the Act, it is possible to imagine fact patterns in which the synthetic media artefact itself need not be communicated for a harmful digital communication to take place. Case law from prosecutions under the HDCA

frequently deals with situations in which allusion is made to the existence of an audiovisual artefact, though the artefact itself is withheld. For instance, a claim that one person possesses compromising photographs about another person. Assuming it occurs digitally, this communication itself may be harmful, even without posting the compromising photographs to any other person. In fact, the photographs in question need not actually exist so long as the effect of the communication meets the "serious emotional distress" test. It is not difficult to imagine scenarios in which the existence of synthetic media is alluded to, and this allusion alone causes harm. For example, serious emotional distress could be caused by one person claiming they have created a synthetic video of the other person that represents the individual in an offensive manner in terms of its content, or that they possess a digital asset of the person's face and voice which they can animate in any way they wish – including to generates representations of torture or pornography. The communication surrounding the synthetic video may satisfy the HDCA tests without ever sending the actual SMA to any other person, and perhaps without even possessing such a video at all. This will be particularly true where an extortionate or coercive threat accompanies the communication, implicitly or explicitly, in turn inviting the attention of criminal provisions, like blackmail.

533. With this in mind, the overarching effect of the New Zealand Bill of Rights Act 1990 is explicitly invoked by s 6(2)(b), which provides that:

> (2) In performing functions or exercising powers under this Act, the Approved Agency and courts must—
> (a) take account of the communication principles; and
> (b) act consistently with the rights and freedoms contained in the New Zealand Bill of Rights Act 1990.

534. The most important effect of this is to establish that the s 14 right to freedom of expression features heavily in any assessment of digital communications under the HDCA. By extension, s 14 also protects the creation and posting of synthetic media artefacts, including deepfakes, subject to usual limitations drawn from other statutes we have identified. The explicit direction to "act consistently" with the NZBORA must be taken as significant given the way that Netsafe would generally be caught by the NZBORA at s 3 as an agency exercising a statutory power or performing a statutory function, indicating something more was intended by the drafter.

535. Returning to the s 6 Communication Principles, Principle 6 is of general relevance to synthetic media: "A digital communication should not make a false allegation." The elements of "false" and "allegation" are notable here in the sense that it is feasible that both may be satisfied through the mere existence of a given synthetic media artefact, by virtue of the nature of that artefact. We refer again to the terms of Condition 1 of the framework: an SMA may be deceptive because it gives the impression it was created by the use of a capture technology and that the use of any Category 2 technologies has not materially undermined the reliability of the Category 3 product as a record of that. In general, some synthetic media artefacts may be so realistic that a reasonable person believes them to be true even when the creator or publisher makes no claim to truth. At the same time, they are categorically non-veridical, being the product of combining multiple digital data to produce novel audiovisual information that does not correspond to anything that actually took place in the real world. Therein arises potential for both implied falseness and implied allegation, even where the artefact is not accompanied by any explicit claim. For example, a synthesised video might depict a man, Mr Doe, in the act of committing a crime, and this video may be so realistic that any reasonable observer would believe the video must be the product of a capture process which has recorded an actual scene as it unfolded, with Mr Doe present and engaged in the activity. Moreover, this video might also be seen as constituting an allegation, even in the absence of any verbal claim like "Here is Mr Doe committing a robbery". Is the video itself both false and allegation? Ultimately the court or Approved Agency will need to account for the facts of each case in its context. Nevertheless, it is another example of how synthetic media will raise questions as to how far we should perceive audiovisual information as opinion, as opposed to fact.

536. Truth or falsehood remains a factor in decisions with regards to orders under the HDCA. Per s 19(5):

> (5) In deciding whether or not to make an order, and the form of an order, the court must take into account the following:
>
> …
>
> (f) the truth or falsity of the statement

537. As outlined in parts 1 and 2 of our report, assessing truth or falsity in a synthetic media artefact must be done carefully. If "false" is understood as "manipulated", then all synthetic media is false because it involves a digital manipulation process. The question of whether its content is false should be understood by condition 1 of our framework in a narrow sense. This will have to be weighed by the fact-finder in light of the many other elements and factors relevant to the HDCA, including s 19(5)(b), which accounts for the purpose or intention of the communicator, in particular whether the communication was intended to cause harm. It would be difficult to infer an intention or purpose to cause harm where manipulation was only incidental, or no reasonable observer would be able to tell that the SMA is a "false statement".

538. The sorts of order that a court may make in response to a successful application under the HDCA are located at s 19 of the Act, and are helpful remedies directed toward harms of dissemination:

> 19 Orders that may be made by court
>
> (1) The District Court may, on an application, make 1 or more of the following orders against a defendant:
>
> (a) an order to take down or disable material:
>
> (b) an order that the defendant cease or refrain from the conduct concerned:
>
> (c) an order that the defendant not encourage any other persons to engage in similar communications towards the affected individual:
>
> (d) an order that a correction be published:
>
> (e) an order that a right of reply be given to the affected individual:
>
> (f) an order that an apology be published.

539. In theory, all of these are useful tools by which to redress harms caused and prevent further spread of harmful synthetic media, however these tools may have limited effect where viral media spreads rapidly. Section 19(5)(j) anticipates this by establishing that a court must take into account, "the technical and operational practicalities, and the costs, of an order." This foreshadows the practical difficulties in any scenario in which communication has occurred over the internet or digital communication technologies. It may be practically impossible to undo the effects of a deepfake by the time it has travelled halfway around the world and across multiple legal jurisdictions and virtual domains. Moreover, regardless of the intention of an order, any apology or right of reply is unlikely to ride upon the same wave as the original video did, thus rendering its effect negligible. By contrast, other legal regimes may become more important, for example the law of defamation, restrictions on broadcasters, resort to social media community guidelines, or copyright takedown claims where appropriate.

540. In summary, the HDCA is a statute focussed on harms of dissemination, and the reason for including condition 3 in our framework. There is a lesser focus on harms of the content of those communications. To the extent that it regulates the truth or falsity of statement, it also calls attention to the capture and creation process, including where it discloses sensitive personal facts.

541. We can say with confidence that the HDCA will play a role in responding to synthetic media like deepfakes in certain circumstances, but that these are limited by a focus on communication, and the practical limitations on court orders. These limitations are a result of an intentional focus on a narrow area given potential risks to NZBORA rights and freedoms.

542. The HDCA is an important tool in a wider range of legislative remedies. This can be good or bad for public policy, but the need to resort to a wide range of remedies does create risks that people will "fall through the cracks" from an access to justice perspective, or that agencies may insist on referring people to other agencies before examining an individual's complaint.

## Harassment Act 1997

543. The Harassment Act is useful because it penalises the kinds of acts that might be caught by the HDCA but do not necessarily meet the definition of "posting a digital communication". For example, harassment could still feasibly occur through Category 1 or Category 2 technologies without ever sending them digitally. The situations where this is the case may be slim, given that mere allusion to an SMA without sending it could still be caught by the HDCA.

544. We think a pattern of harassment using SMT or SMA would allow someone to seek a restraining order against a perpetrator. Section 3(1) states: "For the purposes of this Act, a person harasses another person if he or she engages in a pattern of behaviour that is directed against that other person, being a pattern of behaviour that includes doing any specified act to the other person on at least 2 separate occasions within a period of 12 months."

545. Section 6 of the Act states the object (understood as objective) of the Act:

> 6 Object
> (1) The object of this Act is to provide greater protection to victims of harassment by—
> (a) recognising that behaviour that may appear innocent or trivial when viewed in isolation may amount to harassment when viewed in context; and
> (b) ensuring that there is adequate legal protection for all victims of harassment.
> (2) This Act aims to achieve its object by—
> (a) making the most serious types of harassment criminal offences:
> (b) empowering the court to make orders to protect victims of harassment who are not covered by domestic violence legislation:
> (c) providing effective sanctions for breaches of the criminal and civil law relating to harassment.
> (3) Any court which, or any person who, exercises any power conferred by or under this Act must be guided in the exercise of that power by the object specified in subsection (1).

546. It also creates an offence of criminal harassment:

> 8 Criminal harassment
> (1) Every person commits an offence who harasses another person in any case where—
> (a) the first-mentioned person intends that harassment to cause that other person to fear for—
> (i) that other person's safety; or
> (ii) the safety of any person with whom that other person is in a family relationship; or
> (b) the first-mentioned person knows that the harassment is likely to cause the other person, given his or her particular circumstances, to reasonably fear for—
> (i) that other person's safety; or
> (ii) the safety of any person with whom that other person is in a family relationship.
> (2) Every person who commits an offence against this section is liable, on conviction, to imprisonment for a term not exceeding 2 years.

547. At section 2, "safety" is defined to include a person's mental well-being, and also extends to concerns about the safety of anyone with whom they have a family relationship. Accordingly, it seems well suited to the harms that might be caused by SMT. Section 8(1)(b) requires that the particular circumstances of the victim are taken into account. The focus on "safety" can also be contrasted with the focus on "serious emotional harm" under the HDCA.

548. Interestingly, the harassment Act also enables a court to impose restrictions on a perpetrator's associates. We wonder whether this could be used in situations where a malicious user of a SMA could be inciting others to distribute the SMA.

549. Specified act is defined at s 2 and s 4(1), but we think is broad enough to include capture, manipulation, creation, and dissemination of audiovisual information. The definition of "specified acts" does not explicitly include the use of capture technologies, although such acts do include watching, and people are entitled generally to capture what they watch. Accordingly, we think if

watching is restricted then capture in electronic media will be too, particularly in light of the object of the Act at s 6(1)(a).

550. We refer to the discussion of what can be a specified act in *NR v District Court at Auckland* [2016] NZCA 429 (12 September 2016) at [34]-[39]. There is no requirement that the specified act itself be harmful or culpable: the Harassment Act is intended to acknowledge that, per s 6(1), an innocent or trivial act in isolation may, if a pattern of behaviour amount to harassment in context. It is therefore highly flexible and examines the context of the case. The section 6 object makes it very clear that the Act is intended to catch all kinds of situations without limitation.

551. Accordingly, it could provide a remedy for repeated use of capture technologies against an individual, or repeated dissemination or manipulation of images if the requisite elements are met.

552. Notably, s 9(4) requires persons in a domestic relationship to use the Domestic Violence Act. Accordingly, it may not assist with intimate visual recordings.

553. In *Beadle v Allen* [2000] NZFLR 639: it was noted that harassment sits parallel to defamation as a remedy. There is a "more rigorous test" entailed by s 4(1)(f) of the Act.

554. Section 25 makes it a criminal offence to breach a restraining order without reasonable excuse.

555. Section 3(2) would also account for slight variations in a pattern of conduct, to account for condition 2 of the framework. For example, a pattern of behaviour doing a specified act on at least two separate occasions in twelve months could include, taking photos, then manipulating, then broadcasting and disseminating over a period of time.

> (2) To avoid any doubt,—
> (a) the specified acts required for the purposes of subsection (1) may be the same type of specified act on each separate occasion, or different types of specified acts:
> (b) the specified acts need not be done to the same person on each separate occasion, as long as the pattern of behaviour is directed against the same person.

556. Section 4 subs (2) and (3) seem to indicate a legislative intention that s 4(1)(f) be very broad and without any limitation so long as acting in a way to undermine "safety" as defined by the Act. The Court of Appeal decision in *NR* indicates that specified acts do not have to be unlawful acts and can even be acts with a lawful purpose, despite the apparent drafting of the defence of lawful purpose at s 17.

557. Per s 10, a victim of harassment can also apply to seek direction under s 19 against a person whom the respondent has encouraged to do a specified act to the person. This could include secondary disseminators being encouraged to maximise the impact of a course of action.

558. Amendments were introduced at s 3(4) by the Harmful Digital Communications Act. It introduces a definition of a "continuing act" directed toward effects that have an effect over a protracted period, suggesting the Harassment Act and HDCA were intended to be used in a complementary manner:

> For the purposes of subsection (3), continuing act includes a specified act done on any one occasion that continues to have effect over a protracted period (for example, where offensive material about a person is placed in any electronic media and remains there for a protracted period).

559. Section 19(1A), related to the terms of restraining orders, inserts a requirement that any order in relation to a specified act that is a continuing act includes an obligation to take reasonable steps to prevent the act from continuing.

560. We think that s 16 is sufficiently broad to apply to conduct by someone that involves intentional appropriation of a person's image to cause that person distress. Accordingly, it is a kind of privacy or personality right that can be exercised in specific circumstances. Breach of such a restraining order is an offence punishable by up to two years imprisonment per s 8.

16 Power to make restraining order

(1) Subject to section 17, the court may make a restraining order if it is satisfied that—
  (a) the respondent has harassed, or is harassing, the applicant; and
  (b) the following requirements are met:
    (i) the behaviour in respect of which the application is made causes the applicant distress or threatens to cause the applicant distress; and
    (ii) that behaviour would cause distress, or would threaten to cause distress, to a reasonable person in the applicant's particular circumstances; and
    (iii) in all the circumstances, the degree of distress caused or threatened by that behaviour justifies the making of an order; and
  (c) the making of an order is necessary to protect the applicant from further harassment.

(2) For the purposes of subsection (1)(a), a respondent who encourages another person to do a specified act to the applicant is regarded as having done that specified act personally.

(3) To avoid any doubt, an order may be made under subsection (1) where the need for protection arises from the risk of the respondent doing, or encouraging another person to do, a specified act of a different type from the specified act found to have occurred for the purposes of paragraph (a) of that subsection.

561. It is notable that s 16(2) applies to other people encouraged by the respondent. Similar legislative intention is reflected in the drafting of ss 19(1)(b) and 19(2).

562. A court has power under the Harassment Act to restrict publication of proceedings (s 39), which allows the limitation of harms arising from the dissemination of SMAs.

563. The Court can impose general conditions and specific conditions in a restraining order which can be tailored to the particular circumstances at hand.

564. There may also be an advantage given to a victim of harassment by the lower civil standard of proof that applies to a restraining order, although breach of a restraining order as a criminal offence will require the higher criminal standard to be met.

565. The Act contains references throughout to the circumstances of the people involved and reasonableness tests. This allows a Court to recognise the unique harms that may be caused by the content of the SMA or the vulnerability of the individual, including any pre-existing relationship that may affect perceptions of the veracity of an artefact.

566. The Court has flexible standards for admission of evidence – it is not strictly bound by Evidence Act.

567. We note s 46 savings provision that states "Nothing in this Act limits or affects any right or remedy that exists or is available, apart from this Act, in respect of harassment.

568. Jurisdictional issues will arise when it comes to enforcement of restraining orders or charging of criminal offences, however one benefit of the Harassment Act is that there would appear to be no requirement that specified acts be conducted solely in New Zealand.

569. We think the broad flexibility of the Harassment Act makes it a useful tool for dealing with the range of harms associated with synthetic media, however it will be subject to the common legal issues identified at the start of Part 3 of our report, namely evidential, jurisdictional, and access to justice issues.

## Fair Trading Act 1986 and Advertising Standards

570. Synthetic media poses few new challenges to existing legal regimes for fair trading and advertising. These regimes have long been concerned with the tension of, essentially, 'acceptable deception'. The technological means of conveying messages in trade and advertising have been subject to constant change, and as such, new technological developments like synthetic media are largely anticipated.

571. Synthetic media may generate new possibilities for permissible deception but the Act is well equipped to anticipate these. The Act demonstrates that something can be relatively deceptive, yet still permissible. Similar to the framework approach we have adopted, whether the use of any given synthetic media artefact strays into impermissible deception is a question of fact. It assesses Category 3 artefacts according to Condition 1, including the extent to which Category 2 technologies have made the artefact deceptive.

572. The Act also creates specific prohibitions on misleading representations about endorsement or sponsorship of goods and services, engaging a kind of right of publicity.

### Unfair conduct that is misleading

573. The issues of synthetic media in light of the Fair Trading Act 1986 (FTA) are relatively straightforward in light of the s 1A purpose provisions:

> 1A Purpose
> (1) The purpose of this Act is to contribute to a trading environment in which—
> (a) the interests of consumers are protected; and
> (b) businesses compete effectively; and
> (c) consumers and businesses participate confidently.
> (2) To this end, the Act—
> (a) prohibits certain unfair conduct and practices in relation to trade; and
> (b) promotes fair conduct and practices in relation to trade; and
> (c) provides for the disclosure of consumer information relating to the supply of goods and services; and
> (d) promotes safety in respect of goods and services.

574. As synthetic media may often possess the quality of being realistic but non-veridical, or making it look or sound like something happened when it did not, the deceptive or misleading capacity of such media is axiomatic. In the context of prohibiting unfair trade, protecting consumers from the potential misuses of synthetic media is a normal concern for law. At the same time, wholesale prohibition of the use of SMAs like deepfakes or digital humans in trade would not only result in absurdities, but depart from general norms surrounding advertising and marketing.

575. The FTA prohibits unfair conduct in trade through ss 9, 10, 11, 12, with the indication that "unfair" means conduct that is misleading or deceptive. The set of provisions reads:

> 9 Misleading and deceptive conduct generally
> No person shall, in trade, engage in conduct that is misleading or deceptive or is likely to mislead or deceive.
>
> 10 Misleading conduct in relation to goods
> No person shall, in trade, engage in conduct that is liable to mislead the public as to the nature, manufacturing process, characteristics, suitability for a purpose, or quantity of goods.
>
> 11 Misleading conduct in relation to services
> No person shall, in trade, engage in conduct that is liable to mislead the public as to the nature, characteristics, suitability for a purpose, or quantity of services.

> 12 Misleading conduct in relation to employment
> No person shall, in relation to employment that is, or is to be, or may be offered by that person or any other person, engage in conduct that is misleading or deceptive, or is likely to mislead or deceive, as to the availability, nature, terms or conditions, or any other matter relating to that employment.

576. These provisions are drafted in a manner that addresses unfair, misleading or deceptive conduct generally, whether it occurs face to face, over the telephone, via email, or through any other technological intermediary. As such, unfair conduct in trade by means of SMTs remains prohibited conduct, regardless of any technological novelty. Many misuses of synthetic media conceivably might breach these provisions.

## "Deceptive content" versus "condition 1 deception"

577. There is another point here that relates to the possibility of deception of another kind. Our focus has been on technologies that make it look like something happened when it did not happen in an audiovisual sense. It is also possible that, even where someone is aware that a SMA is deceptive in the sense of Condition 1, the content of the representation is nevertheless unsubstantiated or deceptive:

> 12A Unsubstantiated representations
> (1) A person must not, in trade, make an unsubstantiated representation.
> (2) A representation is unsubstantiated if the person making the representation does not, when the representation is made, have reasonable grounds for the representation, irrespective of whether the representation is false or misleading.
> (3) This section does not apply to a representation that a reasonable person would not expect to be substantiated.
> (4) In this section and sections 12B to 12D, representation means a representation that is made—
> (a) in respect of goods, services, or an interest in land; and
> (b) in connection with—
> (i) the supply or possible supply of the goods or services; or
> (ii) the sale or grant or possible sale or grant of the interest in land; or
> (iii) the promotion by any means of the supply or use of the goods or services or the sale or grant of the interest in land.

578. We also refer to the Advertising Standards Authority "Advertising Standards Code" (the Code), which is a non-binding self-regulatory regime that sets out a range of principles and rules for responsible advertisement and applying to "all advertisements placed in any media." The broad definition of "advertisement" as offered by the Code will inevitably include most synthetic media artefacts that are consumed as display-based products:

> **"Advertising and Advertisement(s)"** means any message, the content of which is controlled directly or indirectly by the advertiser, expressed in any language and communicated in any medium with the intent to influence the choice, opinion or behaviour of those to whom it is addressed.

579. Principle 2 of the Code is particularly relevant, establishing a norm of "Truthful Presentation". Rule 2(a) in relation to "Identification" has practical implications for the way synthetic media can be used in conveying meaning to consumers. One of the potential values of artefacts like synthesised videos or virtual avatars, especially where these are 'learning' systems that can respond to data collected from consumers, is that they may appear more real or natural in the way that they deliver an advertising message. Rule 2(a) would nonetheless require that these identify themselves as advertisements. For example, in the near-future we can expect a proliferation of advanced chatbot-like "digital humans" in marketing. These artefacts use sensors and computer vision paradigms to assess a target person's emotions and facial expressions. They also utilise sophisticated animation methods to produce much better-quality computer-generated representations of 'humans' than has previously been possible. This combines with advanced natural language processing so that the machine can actively listen and respond to what a consumer says. When perceived over virtual

channels, it may be difficult for a consumer to identify that they are interacting with a machine, engaging deception of a kind anticipated by Condition 1 of our framework. But further, even where a consumer is aware they are talking to a machine, there may be a separate question as to whether they are aware that some or all of what the machine tells them is an advertisement.

580. Already there is debate around possible mandatory self-identification when a computer system interacts with a human user. This is an increasingly relevant question, as it becomes more difficult to distinguish between machines and humans in certain settings. For example, Google's "Duplex" AI conducted a telephone call and booked an appointment with a salon. All the while, the human receptionist appeared to be unaware that they were speaking to a computer system. The driving force in favour of mandatory machine identification rests on its deceptive capacities. Existing requirements to identify advertisements are, in this regard, a fascinating parallel – both are an extension of the same perceived need to mitigate potential deception. Feasibly, machine self-identification would fall within the normal ambit of consumer information standards, which may be established by Order in Council pursuant to the recommendation of a Minister as established by s 27 of the FTA.

581. By way of example, consider the following likely future scenario. An advanced "digital human" computer system is used to direct consumers towards certain products and services, or generally help them find solutions to problems in a variety of contexts. This system also possesses a highly realistic animated face and voice, can use language naturally and adaptively, and has its own facial expressions and simulated emotional responses. When engaging in conversation with a consumer, it may or may not be mandatory for the system to identify itself as a machine. Similarly, if the system sometimes engages in advertisements, these may need to be identified, either generally by saying that this system is an advertising system, or as specific recommendations are given to the consumer, like "your problem is interesting, you should consider contacting Company X, and by the way, what I just said was an advertisement paid for by Company X." Consistent with our conclusions above, we think situations of this kind are already covered by advertising standards.

582. Any work in this area should be done by close reference to subject matter experts and a realistic understanding of the technology's capabilities. Nonetheless, it makes sense for both government and business to look forward and anticipate technologies of this nature to avoid the risk of potential harm.

### Category 3 product assessed in context

583. Returning to the FTA, s 12B provides supplement to the s 12A provisions:

> 12B Court must have regard to certain matters
> (1) In proceedings concerning a contravention of section 12A, and in assessing whether a person had reasonable grounds for a representation, a court must have regard to all of the circumstances, including—
> (a) the nature of the goods, services, or interest in land in respect of which the representation was made:
> (b) the nature of the representation (for example, whether it was a representation about quality or quantity):
> (c) any research or other steps taken by or on behalf of the person before the person made the representation:
> (d) the nature and source of any information that the person relied on to make the representation:
> (e) the extent to which the person making the representation complied with the requirements of any standards, codes, or practices relating to the grounds on which such a representation may be made, and the nature of those requirements:
> (f) the actual or potential effects of the representation on any person.

584. The application of these provisions to representations made using synthetic media is not likely to be controversial. Of particular interest is how the general view of the courts might develop in its

perspective on synthetic media as a tool for advertising representations. Ultimately the FTA and its supporting devices are equipped for assessments like these without the need for development of legislation specifically designed to account for more advanced synthetic media products.

585. Without exploring the innumerable potential fact patterns of synthetic media misrepresentations, s 13 sets out some clear examples of how an SMA or SMT might be used deceptively:

> 13 False or misleading representations
> No person shall, in trade, in connection with the supply or possible supply of goods or services or with the promotion by any means of the supply or use of goods or services,—
> (a) make a false or misleading representation that goods are of a particular kind, standard, quality, grade, quantity, composition, style, or model, or have had a particular history or particular previous use; or
> (b) make a false or misleading representation that services are of a particular kind, standard, quality, or quantity, or that they are supplied by any particular person or by any person of a particular trade, qualification, or skill, or by a person who has other particular characteristics; or
> …
> (d) make a false or misleading representation that goods are new, or that they are reconditioned, or that they were manufactured, produced, processed, or reconditioned at a particular time; or …

## Personality and publicity rights

586. Section 13 also anticipates the kinds of harms associated with misappropriation of image, infringement of the right of publicity, or the unauthorised use of someone's identity or likeness anticipated by personality rights:

> (e) make a false or misleading representation that goods or services have any sponsorship, approval, endorsement, performance characteristics, accessories, uses, or benefits; or
> (f) make a false or misleading representation that a person has any sponsorship, approval, endorsement, or affiliation; or
> ...

587. As noted by the authors of Todd on torts, many possible tortious remedies associated with appropriation of someone's likeness also fall to be determined under the Fair Trading Act, and there are certain advantages to pursuing a remedy through each.[171] In any event, the misappropriation of someone's likeness in a commercial setting is anticipated by New Zealand law within particular parameters on appropriate facts. We think the Fair Trading Act will be the preferable regime given other access to justice barriers and uncertainty in pursuing tortious action. Notably, there will be a factual issue similar to that discussed under the Privacy Act about whether, in an evidential sense, a person's identity, image, or likeness has been appropriated on the basis that visual or aural identity is difficult to define in a consistent and objective sense. This is a vexed question in publicity rights jurisdictions.[172]

588. The subsequent effect is that many potential false or misleading representations made through the use of SMAs might already be prohibited. This included the use or over-use of rapid editing tools that manipulate the presentation of goods or services, "deepfake" style videos that represent publicly recognisable individuals endorsing people or products, or which represent publicly recognisable individuals using particular products, and so on. This would be balanced against protection of free expression and the limits of acceptable 'puffery'. The provisions will also work alongside other law which generates limitations on legal representations but protects similar interests.

[171] See Todd, above n 159, 14.2.02, 14.4, and discussions of *Tot Toys Ltd v Mitchell* [1993] 1 NZLR 325 (HC) at 359–366 in the context of the tort of passing off.
[172] Zapparoni, above n 106.

# Copyright and the rights of creators

## The Copyright Act 1994

589. We think that SMT and SMA have significant socioeconomic benefits. In fact, New Zealand companies are leaders in the use of SMT through its film, visual effects and artificial intelligence industries. For this reason, it is important that the rights of creators are secure to strike a balance between innovation and the right to recover financial reward for that creativity.

590. Another benefit of the Copyright Act is that it has been grappling with terminology to describe the various rights and interests in audiovisual material for some time, including issues about the way it is modified or disseminated.

591. In this part we note helpful definitions from the Copyright Act which we think lend support to our framework.

592. We also note the varying property interests granted by copyright to illustrate the diverse ways that SMA can be acted upon.

593. We think that any property framing in relation to SMA should be restricted to copyright. We think that the rights of individuals featured in copyright works must be dealt with through an individual privacy framework as supplemented by the Fair Trading Act. The interaction between these legal regimes is complex and will require further development in order to protect both the dignity and autonomy of individuals and the commercial certainty of creators.

594. There is a wide range of potential uses of SMT in a creative context. To the extent that all SMA involve the use of SMT to capture, manipulate, display and disseminate light and sound energy and digital data, those specific tasks are all anticipated by the Copyright Act because they occur in orthodox film, music and digital effects industries. The Copyright Act is currently under review by the Ministry of Business, Innovation and Employment, and data-mining for use in Artificial Intelligence is a topic of discussion, which is anticipated by Condition 2 (multiplicity).[173]

### Copyright as a framework for synthetic media

595. There are similarities between the definitions adopted within the Act and the three categories we articulate in our framework. To the extent that there is a need to regulate various uses of SMA and SMT, definitions in the Copyright Act may provide a useful starting point.

596. We think SMA of various kinds can be described by the following definitions which we link to elements of our framework.

597. Category 1 technologies that capture light or sound are defined and anticipated by the following and we think reference to light and sound lends significant support to the boundaries drawn in our framework:

> sound recording means—
> (a) a recording of sounds, from which the sounds may be reproduced; or
> (b) a recording of the whole or any part of a literary, dramatic, or musical work, from which sounds reproducing the work or part may be produced,—
> regardless of the medium on which the recording is made or the method by which the sounds are reproduced or produced

[173] Ministry of Business, Innovation & Employment *Issues Paper: Review of the Copyright Act 1994* (November 2018): see paras 132-133, 149-152, and 296-306.

**photograph** means a recording of light or other radiation on any medium on which an image is produced or from which an image may by any means be produced; but does not include a film or part of a film

**film** means a recording on any medium from which a moving image may by any means be produced

598. Notably, the definitions of "sound recording", "photograph" and "film" share the following properties, which lend support to the terms of condition 1 of our framework:

    a. they do not explicitly draw attention to the role of category 2 manipulation technologies;

    b. they treat the relationship between the light or sound captured (Category 1) and the light or sound displayed (Category 3) as being relatively direct. We acknowledge that the word "recording" may be read to include the use of digital effects processes in the course of recording.

599. The Copyright Act anticipates the role of manipulation technologies (Category 2) in generating a work by the use of computers. It also anticipates Condition 2 of the framework, multiplicity, by defining a "compilation" so as to include other works and parts of works, including where there may be "distinct contributions by different authors" or more than one author's work incorporated. This could account for the use of artificial intelligence to generate works from large databases.

**compilation** includes—
(a) a compilation consisting wholly of works or parts of works; and
(b) a compilation consisting partly of works or parts of works; and
(c) a compilation of data other than works or parts of works

**computer-generated**, in relation to a work, means that the work is generated by computer in circumstances such that there is no human author of the work

**collective work** means—
(a) a work of joint authorship; or
(b) a work in which there are distinct contributions by different authors or in which works, or parts of works, of different authors are incorporated

600. A computer program can be a literary work on the basis that it is writing

**literary work** means any work, other than a dramatic or musical work, that is written, spoken, or sung; and includes—
(a) a table or compilation; and
(b) a computer program

**writing** includes any form of notation or code, whether by hand or otherwise and regardless of the method by which, or medium in or on which, it is recorded; and written has a corresponding meaning.

601. The Act also defines what could be seen as category 2 manipulation technologies that reproduce, record or store a work digitally, including by extracting part of a work from it:

**copying**—
(a) means, in relation to any description of work, reproducing, recording, or storing the work in any material form (including any digital format), in any medium and by any means; and
...
(c) includes, in relation to an artistic work, the making of a copy in 3 dimensions of a two-dimensional work and the making of a copy in 2 dimensions of a three-dimensional work; and
(d) includes, in relation to a film or communication work, the making of a photograph of the whole or any substantial part of any image forming part of the film or communication work—
and copy and copies have corresponding meanings

602. The Act also anticipates adaptations of works, which could embrace the use of Category 2 manipulation technologies across multiple authors in the manner anticipated by Condition 2.

Specifically, an adaptation can be made of a computer program in a way that is not incidental to the course of running the program.

> **adaptation**,—
> (a) in relation to a literary or dramatic work, includes—
> (i) a translation of the work from one language to another:
> (ii) a version of a dramatic work in which it is converted into a literary work or, as the case may be, of a literary work in which it is converted into a dramatic work:
> (iii) a version of the work in which the story or action is conveyed wholly or mainly by means of pictures in a form suitable for reproduction in a book, or in a newspaper, magazine, or similar periodical:
> (b) in relation to a literary work that is a computer program, includes a version of the program in which it is converted into or out of a computer language or code or into a different computer language or code, otherwise than incidentally in the course of running the program:
> (c) in relation to a musical work, means an arrangement or transcription of the work

603. We think the transduction processes whereby light or sound is converted to electrical energy is also anticipated by the Act in the definitions of "electronic" and the inclusion of electronic storing of information "in electronic form". Similarly, the definition of a "document" includes "information derived from that information", in the same way as the definition of "document" under the Privacy Act, and appears to anticipate the kinds of adaptation or compilation works cited above.

> **electronic** means actuated by electric, magnetic, electro-magnetic, electro-chemical, or electro-mechanical energy; and **in electronic form** means in a form usable only by electronic means
>
> **document**, for the purposes of Part 6A and sections 144A and 144C to 144E, means—
> (a) any material, whether or not it is signed or otherwise authenticated, that bears symbols (including words and figures), images, or sounds, or from which symbols, images, or sounds can be derived, and includes—
> (i) a label, marking, or other writing that identifies or describes a thing of which it forms part, or to which it is attached:
> (ii) a book, map, plan, graph, or drawing:
> (iii) a photograph, film, or negative; and
> (b) information electronically recorded or stored, and information derived from that information

604. We see Condition 3 of our framework reflected in the following definitions, which distinguish between a work and the way it is disseminated:

> **Internet service provider** means a person who does either or both of the following things:
> (a) offers the transmission, routing, or providing of connections for digital online communications, between or among points specified by a user, of material of the user's choosing:
> (b) hosts material on websites or other electronic retrieval systems that can be accessed by a user

605. The definition of "telecommunications system", "communication work" and "communicate" illustrate the difficult boundary between Category 3 display and Condition 3, whereby display and dissemination may be very similar:

> **telecommunications system** means a system for conveying visual images, sounds, or other information by electronic means
>
> **communicate** means to transmit or make available by means of a communication technology, including by means of a telecommunications system or electronic retrieval system, and communication has a corresponding meaning
>
> **communication work** means a transmission of sounds, visual images, or other information, or a combination of any of those, for reception by members of the public, and includes a broadcast or a cable programme

606. Section 14 of the Act describes the property right conferred by copyright and we think that SMA and SMT can be caught by s 14(1)(b), (c), (d). We note that s 14(2) may undermine copyright in a

compilation work as defined above such as a deepfake, although the inclusion of other works may be difficult to show evidentially in works produced from large databases:

> 14 Copyright in original works
> (1) Copyright is a property right that exists, in accordance with this Act, in original works of the following descriptions:
> (a) literary, dramatic, musical, or artistic works:
> (b) sound recordings:
> (c) films:
> (d) communication works:
> ...
> (2) A work is not original if—
> (a) it is, or to the extent that it is, a copy of another work; or
> (b) it infringes the copyright in, or to the extent that it infringes the copyright in, another work. ...

607. Having established that SMA, including the use of SMT, may be covered by the Copyright Act, we note the various ways that dealing in copyrighted SMAs may be restricted by the author:

> 16 Acts restricted by copyright
> (1) The owner of the copyright in a work has the exclusive right to do, in accordance with sections 30 to 34, the following acts in New Zealand:
> (a) to copy the work:
> (b) to issue copies of the work to the public, whether by sale or otherwise:
> (c) to perform the work in public:
> (d) to play the work in public:
> (e) to show the work in public:
> (f) to communicate the work to the public:
> (g) to make an adaptation of the work:
> (h) to do any of the acts referred to in any of paragraphs (a) to (f) in relation to an adaptation of the work:
> (i) to authorise another person to do any of the acts referred to in any of paragraphs (a) to (h).
> ...

608. The author of a work is defined at s 5, and we particularly note s 5(2)(a) and (b) and their application to SMA produced by SMT through Category 1 and 2 technologies:

> 5 Meaning of author
> (1) For the purposes of this Act, the author of a work is the person who creates it.
> (2) For the purposes of subsection (1), the person who creates a work shall be taken to be,—
> (a) in the case of a literary, dramatic, musical, or artistic work that is computer-generated, the person by whom the arrangements necessary for the creation of the work are undertaken:
> (b) in the case of a sound recording or film, the person by whom the arrangements necessary for the making of the recording or film are undertaken:
> (c) in the case of a communication work, the person who makes the communication work:
> ...
> (3) The author of a work of any of the descriptions referred to in subsection (2) may be a natural person or a body corporate.

609. We conclude that SMAs and the use of SMTs are regulated by the Copyright Act and that the definitions therein are helpful support for our framework. The remainder of the Act sets out an established framework for dealing in copyright works. The exact application of that framework to particular SMA is difficult to predict in an abstract sense, but we are confident that the Act broadly applies.

610. Despite the application of copyright to many kinds of SMT and SMA, we think that the role of copyright in the protection of the subjects of SMA will be limited. That is because the defining feature of copyright is that it primarily protects the interests of creators and artists.

611. Many of these issues are not novel or different when it comes to emerging SMT. The interests of creators are particularly important, however the comprehensive treatment of the interests of creators is beyond the scope of this report. Instead, we conclude:

    a. There is no reason to think that SMAs will not be caught by the Copyright Act at a high level.

    b. The generation of an SMA using SMT will involve a trade-off between a range of actors with copyright in artefacts at various stages of the process.

    c. It contains a range of useful definitions that should be considered if any attempt is to be made to define SMAs. We think most SMAs are caught by these definitions.

    d. We note that many of the definitions of a "recording" process overlook the role of digital manipulation technologies. We think that the review of the Copyright Act should consider the extent to which technologies of capture, manipulation and display may generate different artefacts which merit separate treatment.

    e. We think that any suggestion that property concepts should be used to determine the way that a person's profile, appearance or publicity is regulated should instead defer to an interaction between concepts of privacy and copyright. Privacy should determine the rights of a subject in a copyright work and copyright should determine the property in the work itself. The interaction of these legal doctrines will require careful research and reflection and broad consultation.

    f. There is a broad issue of enforceability of copyright, consistent with wider changes in copyright as influenced by digital media and the internet. The fact that SMAs may draw on a wide range of copyright materials as its source will add to this complicated area.

    g. We note that technological solutions to copyright enforcement are well advanced on social media platforms, but that the ease with which machine learning processes can recognise copyright content is very different to the way that machine learning algorithms would recognise content that has been manipulated in an impermissible way, or generated afresh. The extent to which manipulation is permissible or impermissible relies heavily on context and social norms, which algorithmic detection methods are particularly ill-equipped to assess.

    h. There is a wide exception in the Act for research purposes. The development of many Category 2 technologies may be able to be claimed as research purposes.

    i. Fair dealing is relatively open textured and is a useful concept for developing acceptable boundaries of copyright on a case-by-case-basis, in the same way that the law of privacy takes an iterative approach to uses of personal information about an identifiable individual.

**Indigenous intellectual property.**

612. New Zealand is founded on a partnership in the Treaty of Waitangi. Creators of synthetic media artefacts need to be aware of protections for indigenous intellectual property in New Zealand. This is especially important where there is any consideration that SMA could be used to animate chatbots in education or healthcare settings in ways that facilitate access for Maori populations.

613. Respect must be had for tapu and noa, including Maori responsibilities of stewardship over tāonga.

614. The use of Māori intellectual property in the creation or content of synthetic media should be treated with extreme caution and deference to tino rangatiratanga and other Treaty concepts, with a full understanding of the taonga in question and consultation with Māori groups.

615. This is reflected in guidance by the intellectual property office in New Zealand:[174]

> Māori attribute physical, economic, social, cultural, historic, and/or spiritual significance to certain words, expressions, performances, images, places, and things. There are many cases where it would not be appropriate to copy or use a Māori cultural element, especially a traditional one.

616. Like other prospective uses of synthetic media technologies, we can only make creators aware of the potential harms that may be caused by their products and the ways that the law will intervene.

[174] See <https://www.iponz.govt.nz/about-ip/maori-ip/concepts-to-understand/>.

# Part 4: Conclusions

617. It is difficult to point to clear gaps in New Zealand law. It is much easier to point to gaps in retrospect once a particular factual pattern has been established by evidence, assessed by a decision-maker with reasons, and then compared to a public policy standard which it was or was not intended to achieve. The nature of synthetic media is that it is sufficiently novel that its potential uses are broad and indeterminate. As a result, we have attempted to define what synthetic media is and point to various ways that it will be caught by existing legal and self-regulatory standards.

618. The law in New Zealand is broadly well-equipped to deal with the impact of technologies which make it look like something happened when it didn't happen. Importantly, it demonstrates that there are positive and benign uses of such technologies. We also note that ostensible 'gaps' in the law may sometimes be intentionally constructed, based on tacit acknowledgement of the limits of law itself and the importance of individual freedoms.

619. A common theme in these legal frameworks is that they deploy definitions, standards and principles that are media neutral and open-textured to allow for gradual development of the law to circumstances as they arise. This level of flexibility, though, is supported by rights of process, transparency, and oversight that give the law its legitimacy. A decision-maker also has to apply legal standards that have been articulated in advance. This allows people to consider the law and how it may apply to them, including by seeking legal advice or even guidance from the regulatory agency itself in some situations.

620. One of the more significant gaps in New Zealand law is not so much a gap as a boundary. It is a result of the nature of its jurisdictional limits to its own sovereign borders (in most cases): in particular, its application to overseas actors, whether other internet users, or to large social media platforms. Importantly, this is not an issue unique to New Zealand or to synthetic media technologies.[175] One of the benefits of our framework is that it allows us to isolate the kinds of harms considered in relation to synthetic media and assess how far, for example, they arise from synthetic media itself or the way it is disseminated.

621. For this reason, we briefly comment on the role of social media platform guidelines and terms of service.

## Social media platform guidelines

622. Our framework broadly allows policymakers to consider what specifically is being alleged to cause harm about the creation, use, content and dissemination of SMT. One important result of this is that many legal regimes dealing with, for example, dissemination of audiovisual material or the making of public statements do not appear to require modification to adapt to SMT. Many of the constituent elements of SMT are already recognisable and it is simply a matter of the technologies in Categories 1-3 being used in more innovative ways, as described in Conditions 1-3 of the framework.

623. There is, however, a reverse side to our findings about how far SMT already generate recognisable impacts: that is that existing difficulties faced by the law in dealing with online harms will apply to SMT too. In particular, standards for truth and falsehood and the tendency for digital media to spread rapidly through dissemination platforms will apply to SMT as much as they do to "fake news" and other forms of online harms or harassment.

624. This makes resort to dissemination platforms' community guidelines and terms of service an important feature of responding to SMT.

[175] See a comprehensive recent report on this topic by Marianne Elliott "Digital Threats to Democracy" (The Workshop, 2019) <www.digitaldemocracy.nz> ISBN: 978-0-473-48026-4.

625. The Community Guidelines of Facebook, Twitter and YouTube broadly reflect their origins in Western democracies. They therefore cover similar areas of harmful human conduct that are already proscribed in New Zealand law, while drawing the line at different places at times. They can often be more prescriptive and entail even greater accessibility than domestic legal mechanisms, even if there is widespread dissatisfaction with the way they are applied, including concern about the role of private entities in performing a censorship function.

626. We note that the community guidelines generally acknowledge the same difficulties faced by the law in determining what is acceptable or unacceptable speech in light of various values such as the rights of creators, individual privacy, freedom of expression and intentionally or recklessly harmful conduct.

627. We understand that many issues brought to the attention of Netsafe are frequently resolved through a takedown request using a platform's terms of service. We think it is vital to consider the ways that the flexibility and speed of using non-legal mechanisms of redress can actually provide access to justice benefits. Simply passing new law without a clear understanding of how it is to be applied, what standard it is intended to create, what evidence is required, whether it can be understood without access to a lawyer, or who will enforce it will simply exacerbate many of the issues faced by victims of synthetic media technologies.

628. In the same way as we did with New Zealand's legal system, we conclude that:

    a. to the extent a restriction on SMAs applies to harmful content or dissemination without going behind the process of creating that SMA, there should be little difficulty in applying standards as they are (acknowledging the public discussion about how effectively these restrictions are applied in practice).

    b. to the extent a restriction calls for attention to truth or falsehood, there may be difficulties in providing an evidential foundation for alleging that an artefact is deceptive in terms of condition 1. This difficulty exists in a forensic sense directed toward manipulation techniques, as well as an evidential sense in relation to what the SMA appears to show.

    c. to the extent that a restriction relies on privacy standards, we think that deference to individual dignity and autonomy should dictate any application or development. Evidently, this is particularly difficult and much of the law of Privacy is directed toward enhancing user power over social media platforms and other users.

629. We note that many of the digital forensic indicia relied upon to identify the extent to which videos have been manipulated may be rendered unusable by the way that platforms alter the digital data comprising the image upon uploading it to their services through the use of compression software.

630. We reiterate our conclusions about the caution that must be taken before taking an approach that adopts a position of censorship by default. In our discussions, there has been a distinct variation in the confidence held by different individuals about the ability of technological detection systems to identify harmful content. To build systems that can identify harm, we have to have a good idea of what kind of harm we mean. The development of these systems involves expertise beyond our own and it is important to consider the extent to which legal intervention imposes standards that are technologically impossible to implement. In summary, the assessment of content or behaviour can be difficult without reference to context. In a similar way, to moderate content based on its manipulation or deceptiveness is not a question of identifying whether a piece of content has been generated or manipulated at all, it is a question of identifying whether that manipulation is meaningful in context.

631. A key finding in our research is that often, the law itself cannot do any better at articulating certain and precise standards that distinguish between permissible and impermissible content and deception.

632. We also note an area of research that applies rule of law values to the terms of service of prominent social media platforms[176] and value this way of thinking about the rights of users to access what are pervasive communications systems.

633. The question of whether, as put by some commentators, social media platforms should be effectively deprived of a licence to operate until they can conclusively monitor all content on them is a democratic one. Our suggestion is that any legislative amendment is made with appropriate caution given the relative positive and negative impacts of synthetic media technologies.

## Specific gaps in New Zealand law

634. The law is poorly equipped to deal with disparate harms. Some commentators have expressed concern about a general proliferation of misleading audiovisual content. The short answer to this concern is that the law will not be concerned with this proliferation until an identifiable harm of a legally cognisable kind results to an identifiable actor. One exception to this is the way that the FVPCA deals in concepts of injury to the public good, however this Act illustrates the high bar to be applied before such restrictions are implemented. Further, the classification office's ability to deal with such content is limited and people are only required to submit a film for labelling if it is to be supplied commercially, indicating a focussed and targeted population. This means that generic concerns about the role of synthetic media in "fake news" are not well suited to legal remedies. We have identified general restrictions and guidelines here through the law of defamation, the Broadcasting Act and the Media Council Guidelines.

635. We have noted how, unlike traditional digital media or analogue media, many different actors can be involved in the creation and dissemination of synthetic media. It can be difficult to identify each of them and their respective roles. That can make it difficult to seek a remedy against them for causing a particular kind of harm. It can also make it difficult for creators to ensure their technology or copyright artefacts are not being used harmfully. Even if that actor can be identified, how can it be shown who did what? What was the material action that caused the harm?

636. Our research has demonstrated that there are a wide range of harms that can result from the creation, content and dissemination of synthetic media that are already covered by the law. There is a risk that consumers and victims do not know where to turn to in order to seek access to justice. Therefore we think there is a risk of "falling through the cracks" and a degree of coordination is required to triage complaints.

637. Any legal standard that calls for a decision-maker to assess whether a video is misleading or manipulated in an impermissible way will require an evidential foundation, both for the complaint itself and to justify the exercise of powers to limit freedom of expression by dealing with the content. The more persuasive the content, the greater the risk of harm, and the more difficult it will be to meet this evidential threshold.

638. We also identify an area of conflict between three areas of the law. We have concluded that an individual can or should have privacy rights in material so long as it is about them as an identifiable individual. At the same time, a creator can have copyright in the same SMA. Any public entity intervening to restrict the way in which someone deals in that SMA may also be limiting the right to freedom of expression. On this basis, we think SMT will create a complex interaction between privacy, copyright, and freedom of expression.[177] The Criminal Law may also be invoked if there has been a criminal capture or dissemination process.

[176] See for example N Suzor "Digital Constitutionalism: Using the Rule of Law to Evaluate the Legitimacy of governance by Platforms" (2018) Social Media + Society 1-11.

[177] This topic is already the subject of exploration by some scholars, for example: A Sims "Strange bedfellows: Fair dealing and freedom of expression in New Zealand" European Intellectual Property Review 33(8):490-499 2011

639. We have not been able to reconcile these areas of the law other than noting that contract law is likely to play a strong role in allocating the rights of property and privacy involved. We note there is specific discussion of parody and satire in the review of the Copyright Act review at pages 55, 56, 68, 71-72, and 111. We also note the following sections of the Privacy Act 1993 appear to anticipate this interaction.

> 115 Protection against certain actions
> (1) Where any personal information is made available in good faith pursuant to principle 6,—
> (a) no proceedings, civil or criminal, shall lie against the Crown or any other person in respect of the making available of that information, or for any consequences that follow from the making available of that information; and
> (b) no proceedings, civil or criminal, in respect of any publication involved in, or resulting from, the making available of that information shall lie against the author of the information or any other person by reason of that author or other person having supplied the information to an agency.
> (2) The making available of, or the giving of access to, any personal information in consequence of a request made under principle 6 shall not be taken, for the purposes of the law relating to defamation or breach of confidence or infringement of copyright, to constitute an authorisation or approval of the publication of the document or of its contents by the individual to whom the information is made available or the access is given.
>
> 28 Trade secrets
> (1) Subject to subsection (2), an agency may refuse to disclose any information requested pursuant to principle 6 if the withholding of the information is necessary to protect information where the making available of the information—
> (a) would disclose a trade secret; or
> (b) would be likely unreasonably to prejudice the commercial position of the person who supplied or who is the subject of the information.
> (2) Information may not be withheld under subsection (1) if, in the circumstances of the particular case, the withholding of that information is outweighed by other considerations which render it desirable, in the public interest, to make the information available.

640. We also note s 105 of the Copyright Act 1994 also anticipates privacy concerns in copyrighted material:

> 105 Right to privacy of certain photographs and films
> (1) A person who, for private and domestic purposes, commissions the taking of a photograph or the making of a film has, where copyright exists in the resulting work but is owned by some other person, the right—
> (a) not to have copies of the work issued to the public; and
> (b) not to have the work exhibited or shown in public; and
> (c) not to have the work communicated to the public.
> (2) Subject to subsection (3), the right conferred by subsection (1) is infringed by a person who does an act of the kind described in paragraph (a) or paragraph (b) or paragraph (c) of subsection (1).
> (3) The right conferred by subsection (1) is not infringed by an act that, under any of the following provisions of this Act, would not infringe copyright in the work:
> (a) section 41 (which relates to the incidental copying of a work in an artistic work, film, or communication work):
> (b) section 59 (which relates to parliamentary and judicial proceedings):
> (c) section 60 (which relates to Royal commissions and statutory inquiries):
> (d) section 66 (which relates to acts done under statutory authority):
> (e) section 67 (which relates to acts permitted on assumptions as to expiry of copyright or death of the author in relation to anonymous or pseudonymous works).
> (4) The right conferred by subsection (1) is infringed by a person who does an act described in subsection (2) or who authorises another person to do such an act.

641. The NZBORA applies to any action taken by the judicial branch or to any person or body performing a public function power or duty conferred by law. No law can be held to be ineffective if it is

inconsistent with the NZBORA, and in any case there may be justified limitations on that restriction. Section 6 of the NZBORA also calls for NZBORA-consistent interpretations to be preferred.

642. We think that the interaction between copyright, freedom of expression and privacy (including contractual dealing in the rights conferred by these areas) merits further exploration in the way that it touches upon New Zealanders as creators, citizens and consumers.

643. The use of synthetic media technologies to create non-consensual pornography is one of the most pressing policy issues. We think s 216G of the Crimes Act must be amended to clarify whether intimate visual recordings that do not involve the use of capture technologies by an accused are restricted by the Crimes Act. The scope of the offence could also be clarified by bringing a prosecution in an appropriate case.

# Specific Recommendations

644. Our recommendations flow from our conclusion that law will find it difficult to prevent harm except by signalling deterrent consequences and communicating the impact of the law on certain activities with regard to SMT.

A. There are a wide range of legal and pseudo-legal regimes touching upon the potential harms caused by the creation, content and dissemination of synthetic media. In particular, we have identified regulation dealing with harms through the lens of privacy law, criminal law, electoral law, property and copyright law, and broadcasting law.

B. Synthetic media can be used in a vast number of ways, both positive and negative. As a result, this report can only be a starting point. We encourage closer ongoing investigation into this area by collaboration between legal and technological subject matter experts.

C. We recommend caution in developing any substantial new law without first understanding the complex interaction of existing legal regimes. Before acting, it is essential to continue to develop an understanding of how these regimes apply to factual scenarios as they arise. Where new law is necessary, it is likely to take the form of minor or nuanced amendment to existing regulation. For now, existing legislation should be given the opportunity to deal with harms from synthetic media technologies as they arise.

D. Any new legislation must take the position that synthetic media technologies and artefacts touch upon individual rights of privacy and freedom of expression, deserving careful attention from policymakers and broad public consultation. There are benefits, risks, and trade-offs to be discussed in deciding whether to allocate responsibility for restricting synthetic media technologies to the State or to private actors. Human rights, the rule of law, natural justice, transparency and accountability are essential ingredients in whatever approach is adopted.

E. There is a risk that the issues resulting from synthetic media will be lost among the wide range of statutes and agencies involved. Accordingly, agencies and stakeholders responsible for the legislation covered here should take the following steps.

    a. First, formulate agreement on the conclusions in this report and their respective responsibilities for the use of synthetic media technologies and artefacts, as defined by our framework.

    b. Secondly, collaborate to issue public statements on their respective responsibilities for the harmful uses of synthetic media technologies, with the goal being to:

        i. provide commercial certainty to actors operating in New Zealand generating artefacts through synthetic media technologies; and

ii. facilitate access to justice for victims of harmful uses by educating legal professionals and members of the public about the remedies available.

c. Thirdly, in light of their conclusions above, consider how to best publicise the potential impacts of synthetic media technologies in a way that:

i. does not cause undue scepticism about audiovisual information generally; and also

ii. increases the chance that individuals will exercise appropriate caution before relying on audiovisual information in a way that generates risk of harm.

d. Fourthly, consider their need for and access to a range of digital forensic services in relation to audiovisual information. In doing so, agencies should note whether private entities can also gain access to these services. Complaints volumes can be limited by increasing access to evidential services in a way that avoids unnecessary dispute about the reliability of audiovisual information and therefore facilitates dispute prevention.

F. The New Zealand Government, along with New Zealand's tech and visual effects sectors, should consider the opportunities for New Zealand in building capacity for digital forensics and expert evidential services to international markets, given New Zealand's strength in the innovation and use of synthetic media technologies.

G. Pursuant to its functions at s 13 of the Privacy Act, the Office of the Privacy Commissioner should initiate public discussion on the extent to which someone has a reasonable expectation against the creation of synthetic media artefacts about that person without their consent, and the extent to which the creation of such synthetic media artefacts might be considered offensive to a reasonable and ordinary person.

H. The legislature consider and make amendment to s 216G of the Crimes Act clarifying whether it is an offence against Category 1 capture technologies or Category 2 manipulation technologies. Stakeholders should be given the opportunity to have input because of the criminal penalties being imposed and the potential infringement on the New Zealand Bill of Rights Act from broad drafting.

I. The review of the Copyright Act 1994 should account for Condition 2 of our framework (multiplicity) and the greater use of Category 2 manipulation technologies in the synthesis of audiovisual artefacts.

J. Apart from existing Copyright protections, New Zealand should not adopt a property-based framework for restricting unauthorised use of an individual's audio-visual profile and should instead prefer a policy response based on individual privacy.

K. Further legal and policy research should be done on the interaction between the law of copyright, privacy and freedom of expression in New Zealand when an individual authorises the use of generative synthetic media technologies to create new synthetic media artefacts about them.

L. The New Zealand government should consider how it can use New Zealand's strengths in effective policy and synthetic media technologies to benefit the international community and facilitate positive international relationships with state and non-state actors.

M. Any individual or agency generating or disseminating synthetic media technologies, or synthetic media artefacts that are highly photo- or phono-realistic should exercise extreme caution and consider how to affix statements or contextual indicators that make it clear how far Category 2 manipulation technologies have been deployed and the extent to which a synthetic media artefact is the result of a Category 1 capture process.

# Concluding remarks

645. Our research has been about distinguishing the specific harms and capabilities of synthetic media from wider issues examined in the context of broader issues like "fake news". We have found that New Zealand law already does touch upon the harms that could be caused by the creation, content and dissemination of synthetic media.

646. We have found that synthetic media is an important avenue for freedom of expression. We have also found that New Zealand law recognises many restrictions on freedom of expression, including the way that synthetic media could be used to generate harmful impacts. We think any attempt to articulate more consistent standards to deal with the harms of synthetic media should take the concepts and drafting of existing law as its starting point, because that existing law incorporates a complex set of trade-offs necessary in a free and democratic society.

647. We have noted that questions of enforceability of that law is a separate issue which would benefit from more comprehensive treatment. There is a wider question about the extent to which domestic law can impact upon the conduct of multinational platforms. That is not unique to synthetic media.

648. We note an apparent inconsistency that, as a society, we can be both: extremely concerned about freedom of expression as a value when it comes to respect for news media and the Trump administration's conduct and attitude to factual claims; yet at the same time be considering greater suppressive powers against expressive content, by both government and private actors. That is particularly so where the calls for greater censorship are being directed towards private entities by people who might otherwise say that those private entities cannot be trusted to act in the public interest in a range of other areas. The very reason for assessing whether greater regulation is required over social media platforms is because they are so ubiquitous and important to everyday life. That cuts both ways: it is also an argument for forestalling regulation that could threaten those benefits as well as avoid those harms.

649. One aspect that we think is being overlooked in the regulatory debate is that law can only be applied in retrospect based on evidence and natural justice processes. We acknowledge that law can signal consequences that deter certain kinds of behaviour. Even so, law does not intervene in situations of diffuse or minimal harm. Law frequently requires evidence of harm before it will provide a remedy. Where a criminal sanction is involved, or a fundamental freedom is restricted, the level of harm required and the level of malicious intent expected is commensurately higher before the law will intervene.

650. We think it is important to note that there are benefits to adopting flexible iterative approaches to removal of content. Social media platforms can actually exert a much greater degree of control over a much wider range of content to a greater level of detail on a much wider basis than State-level actors can. As they did in the case of the Christchurch livestream, they can deploy automated mechanisms that trigger take-downs that lead to false positives removing legitimate content. Such conduct by a government agency would be much more serious. We cannot separate domestic law from the fact that it is enforced by State actors, and that inevitably raises human rights concerns.

651. Regulation can have the effect of increasing barriers to innovation and competition. There is talk of Facebook paying a $5b fine. That is impossible for any new company to face. It is the reasoning behind s 230 of the Communications Decency Act in the US that immunises social media platforms from being treated as publishers. Even the need to seek legal advice on compliance with matters detailed in this report entails financial expenditure that can cripple small competitors.

652. Another point to note is that New Zealand, like many other countries, faces a real crisis of access to justice. If social media platforms can provide faster and cheaper remedies to victims of harmful behaviour than domestic law can, this must be taken into account when assessing how far State actors should be intervening in online behaviour.

653. It is simply not the case that the State only acts benevolently: this is the purpose of the rule of law, constitutionalism and human rights. Just because large private companies may also be guilty of abuses of power that might be regulated, that does not mean we should relinquish long-standing limitations that we generally impose on the exercise of State power.

654. When it comes to human rights and fundamental freedoms, it is vital to take a long-term approach. The way in which the United States of America has come to be governed by an Executive with substantially different values from the previous administration indicates the value of maintaining fundamental democratic freedoms, even in times where there is a high degree of trust in government benevolence. Further, it cannot be overlooked that Government is a large entity crossing a range of different functions. The Government may simultaneously inflict and alleviate harms at the same time in the same or different sectors. The ability to call out government breaches of human rights or to express one's opinion about those breaches is absolutely fundamental. There is a risk that regulating certain kinds of technologies will undermine that ability. We point to two specific cases. The first is the way that open-source intelligence organisations such as Bellingcat perform valuable work falsifying misinformation by using the same kinds of information technologies that can be used to spread that misinformation. Their work would not be possible without using the same technologies that facilitate misinformation's spread. The second is the work of WITNESS, a human rights organisation devoted to exposing human rights abuses through audiovisual information. The same live-streaming technology used to livestream murder in Christchurch in a way that generated unacceptable harm was used to livestream the repeated shootings of unarmed civilians in the US, leading to significant social movements that cut across social, political and economic divides. We note that the law as it is in New Zealand has intervened in that Christchurch content.

655. None of this is to say that no action whatsoever should be taken. It is completely legitimate to call for regulatory intervention. But the merits of any course of action cannot be assessed without specifics. What exactly is being proposed? In the case of harmful synthetic media, even if we all agreed we should ban it or regulate it, how could we realistically do that? What exactly are we looking to prevent?

656. We have focussed heavily on assessing the extent to which the kinds of harms anticipated by our legal subject are already subject to regulation, and the extent to which those pieces of regulation are able to actually set reliable and specific standards in advance that allow for effective intervention. We have yet to see any specific proposals for regulation that would do a better job at being specific about the kinds of harms being caused than those already incorporated into New Zealand's legal system. We think that many suggestions about the role of the law as it relates to digital democracy, synthetic media, fake news and disinformation face similar challenges.

# Selected Bibliography


A Sims “Strange bedfellows: Fair dealing and freedom of expression in New Zealand” European Intellectual Property Review 33(8):490-499 2011

Adobe “Remove objects from your videos with the content-aware fill panel” <https://helpx.adobe.com/nz/after-effects/using/content-aware-fill.html>.

Alex Hern “Internet crackdown raises fears for free speech in Britain” (8 April 2019) The Guardian <www.theguardian.com/technology/2019/apr/08/online-laws-threaten-freedom-of-speech-of-millions-of-britons>

Alex Hern “New AI fake text generator may be too dangerous for release, say creators” (14 February 2019) The Guardian <https://www.theguardian.com/technology/2019/feb/14/elon-musk-backed-ai-writes-convincing-news-fiction>.

Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton "ImageNet Classification with Deep Convolutional Neural Networks" (2012) 25 Advances in Nural Information Processing Systems.

Ali Breland “The Bizarre and Terrifying Case of the “Deepfake” Video that Helped Bring and African Nation to the Brink” (15 March 2019) Mother Jones <www.motherjones.com/politics/2019/03/deepfake-gabon-ali-bongo/>.

Andreas Rossler et al "FaceForensics: A Large-scale Video Dataset for Forgery Detection in Human Faces" (2018).

Andrew Gilden "IP, R.I.P." (2017) 95 Washington University Law Review 639.

Aria Thaker “Should India worry about deepfakes affecting the upcoming election?” (26 March 2019) Quartz India <https://qz.com/india/1575860/could-deepfake-videos-spread-fake-news-in-2019-indian-election/>.

Arthur C Clarke “Hazards of Prophecy: The Failure of Imagination” in Profiles of the Future: An Enquiry into the Limits of the Possible (1973) 14 to 36.

Bellingcat <https://www.bellingcat.com/>.

Blackbird <https://www.blackbird.ai/>.

Bobbie Johnson “Deepfakes are solvable - but don’t forget that “shallowfakes” are already pervasive” 25 March 2019) MIT Technology Review <https://www.technologyreview.com/s/613172/deepfakes-shallowfakes-human-rights/>.

Brian Resnick “We’re underestimating the mind-warping potential of fake video” (24 July 2018) Vox <https://www.vox.com/science-and-health/2018/4/20/17109764/deepfake-ai-false-memory-psychology-mandela-effect>.

Broadcasting Standards Act 1989.

Broadcasting Standards Codebook 2016 at p 15.

Bruce P. Keller, Condemned to Repeat the Past: The Reemergence of Misappropriation and Other Common Law Theories of Protection for Intellectual Property, 11 HARV. J.L. & TECH. 401, 428 (1998)

Bryce Edwards “Jacinda Ardern’s ‘Christchurch Call’ might not be so simple” (29 April 2019) New Zealand Herald <https://www.nzherald.co.nz/nz/news/article.cfm?c_id=1&objectid=12226256>.

BuzzFeed “You Won’t Believe What Obama Says In This Video!” (17 April 2018) <www.youtube.com/watch?v=cQ54GDm1eL0&feature=youtu.be>.

C v Holland [2012] NZHC 2155; [2012] 3 NZLR 672 (24 August 2012)

Canadian Charter of Rights and Freedoms, art 8

Caroline Chan “Everybody Dance Now” (22 August 2018) <www.youtube.com/watch?v=PCBTZh41Ris&feature=youtu.be>.

Caroline Chan, et al. "Everybody dance now." (2018) <arXiv preprint arXiv:1808.07371>.

Carolyn Giardina "ILM's New chief Talks Deepfakes, A.I. and the Virtual Reality Used on 'Solo'" (10 August 2018) The Hollywood Reporter <https://www.hollywoodreporter.com/news/deepfakes-take-hollywood-says-ilms-new-chief-1132560>.

Casey Chin "This Broswer Extension is Like an Antivirus for Fake Photos" (20 August 2018) Wired <https://www.wired.com/story/surfsafe-browser-extension-save-you-from-fake-photos/>.

CBN v McKenzie Associates [2004] NZHRRT 48 (30 September 2004) at [41].

Cedric Vanleenhove "The European Court of Justice in Bolagsupplysningen: The Brussels I Recast Regulation's jurisdictional rules for online infringement of personality rights further clarified" (2017) 34 Computer Law & Security Review 640 to 646.

Charlie Warzel "He Predicted the 2016 Fake News Crisis. Now He's Worried About An Information Apocalypse" (11 February 2018) Buzzfeed News <www.buzzfeednews.com/article/charliewarzel/the-terrifying-future-of-fake-news>.

Chris Welch "Samsung Galaxy's S10 has up to six cameras: here's what they all do" (20 February 2019) The Verge <www.theverge.com/2019/2/20/18233130/>.

Christina Cardoza "Google Cloud Text-to-Speech now generally available" (28 August 2018) SD Times <https://sdtimes.com/ai/google-cloud-text-to-speech-now-generally-available/>

Coco v A N Clark (Engineers) Ltd [1969] RPC 41.

Criminal Code Amendment (Sharing of Abhorrent Violent Material) Bill 2019.

Damian Schofield "The use of computer generated imagery in legal proceedings" (2016) 13 Digital Evidence and electronic Signature Law Review 3.

Dave Gershgorn "A West Virginia teen taught himself how to build a rapping AI using Kanye West lyrics" (18 March 2017) Quartz <https://qz.com/920091/a-west-virginia-teen-taught-himself-how-to-build-a-rapping-ai-using-kanye-west-lyrics/?curator=MusicREDEF>.

David Feldman Civil Liberties and Human Rights in England and Wales (2nd ed, Oxford University Press, Oxford, 2002) at 517-518.

David French "The Social Media Censorship Dumpster Fire" (1 March 2019) National Review <www.nationalreview.com/2019/03/the-social-media-censorship-dumpster-fire/>.

David Harvey, "Digital Property - Dixon v R [2015] NZSC 147, [2016] 1 NZLR 678" [2017] NZCLR 195

Dawie Olivier "Keynote - AI Day 2018" (22 April 2018) <youtube.com>.

Dawn Stover "Garlin Gilchrist: Fighting fake news and the information apocalypse: (2018) 74 Bulletin of the Atomic Scientists 283 to 288.

Deeptrace Lab <www.deeptracelabs.com/> and Tracer newsletter.

Defamation Act 1992.

Depiction of individual using digital or electronic technology: sexually explicit material Bill 2019 (Bill 564) <www.legiscan.com/CA/text/SB564/id/1926323/California-2019-SB564-Introduced.html>,

Derek Hawkin "The Cybersecurity 202: Doctored videos could send fake news crisis into overdrive, lawmakers warn" (31 July 2018) The Washington Post <https://www.washingtonpost.com/>.

Devin Coldeway "The future of photography is code" (23 October 2018) Techcrunch <https://techcrunch.com/2018/10/22/the-future-of-photography-is-code>.

Cristiano Lima "Tom Hanks did not wear this profane anti-Trump shirt" (7 October 2018) Politico <https://www.politico.com/interactives/2018/is-this-true/tom-hanks-antitrump-doctored/>

Devindra Hardawar "Magic Leap wants to create art, not just technology" (25 August 2018) Engadget <engadget.com>.

Di Wen et al "Face Spoof Detection With Image Distortion Analysis" (2015) 10 IEEE Transactions on Information Forensics and Security 74.

Dixon v R [2014] NZCA 329, [2014] 3 NZLR 504.

Dixon v R [2015] NZSC 147, [2016] 1 NZLR 678.

Donie O'Sullivan, et al. "Pentagons race against deepfakes" (2019) CNN <https://edition.cnn.com/interactive/2019/01/business/pentagons-race-against-deepfakes/>.

Douglas v Hello! Ltd [2005] EWCA Civ 595.

Erdelyi OJ. and Goldsmith J. (2018) Regulating Artificial Intelligence: Proposal for a Global Solution. New Orleans, USA: AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society, 2-3 Feb 2018.

Erin Griffith "The Other Tech Bubble" (16 December 2017) <wired.com>.

Event Horizon Telescope "Astronomers Capture First Images of a Black Hole" (2019) <www.eventhorizontelescope.org/>.

Films, Videos, and Publications Classification Act 1993.

Finances Africa “Le président gabonais Ali Bongo, mort ou vivant ? Vidéo deepfake?” (1 January 2019) <www.youtube.com/watch?v=62vkG7xfC18&feature=youtu.be&t=25>.

FireEye Intelligence "Suspected Iranian Influence Operation Leverages Network of Inauthentic News Sites & Social Media Targeting Audiences in U.S., UK, Latin America, Middle East" (21 August 2018) FireEye Intelligence <www.fireeye.com>.

Francesco Marconi, Till Daldrup “How the Wall Street Journal is preparing its journalists to detect deepfakes” (15 November 2018) NiemanLab <https://www.niemanlab.org/2018/11/how-the-wall-street-journal-is-preparing-its-journalists-to-detect-deepfakes/>.

Franklin Foer “The Era of Fake Video Begins” (May 2018) The Atlantic <www.theatlantic.com/magazine/archive/2018/05/realitys-end/556877/>.

Gabe Cohn “AI Art at Christie’s Sells for $432,500” (25 October 2018) The New York Times <www.youtube.com/watch?v=Y_9TZHGswVA>.

Gavin Ellis “Deep Fake” New Zealand Listener (New Zealand, 16 February 2019) at 14.

General Data Protection Regulation 2016/679.

Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeeo Jaitly, Andrew Senior "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups" (2012) 29 IEEE Signal Processing Magazine 82 to 97.

Grant Blank "WHO CREATES CONTENT?: Stratification and content creation on the Internet" (2013) 16 Information Communication & Society 590 to 612.

Grant Blank, Darja Groselj "Dimensions of Internet use: amount, variety, and types: (2014) 17 Information Communication & Society 417-435.

Greg Cross and Dr Mark Sagar "Soul Machines - AI-Day 2018" (22 April 2018) <youtube.com>.

Gunther Pernul et al "All Your Voices are Belong to Us: Stealing Voices to Fool Humans and Machines" (paper presented at 20th European Symposium on Research in Computer Security, Vienna, September 2015).

HDCA s 6 Principles 6 and 7.

Hosking v Runting [2004] NZCA 34 (25 March 2004); [2005] 1 NZLR 1; (2004) 7 HRNZ 301.

Houdini “Ubisoft | Far Cry 5 | Houdini Connect” <www.youtube.com/watch?v=k8ChCR8vBGk&feature=youtu.be&t=93>.

Electronic Frontiers Foundation “We don’t need new laws for faked videos, we already have them” <https://www.eff.org/deeplinks/2018/02/we-dont-need-new-laws-faked-videos-we-already-have-them>

Human Rights Watch “Singapore: Reject Sweeping ‘Fake News’ Bill” (3 April 2019) <https://www.hrw.org/news/2019/04/03/singapore-reject-sweeping-fake-news-bill> ;

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaaroun Courville, Yoshua Bengio "Generative adversarial nets." (paper presented at Neural Information Processing Systems 27, Montreal, Canada, December 2014).

Idealog “Soul Machines ‘Digital Humans’” (7 September 2017) <https://www.youtube.com/watch?v=rRsBMEwflz8>.

International Covenant on Civil and Political Rights, art 17; Universal Declaration of Human Rights, art 12; European Convention on Human Rights, art 8; American Convention on Human Rights, art 11(2).

Isobel Asher Hamilton “Scarlett Johansson says trying to stop people making deepfake porn videos of her is ‘a lost cause’” (31 December 2018) Business Insider Australia <https://www.businessinsider.com.au/scarlett-johansson-stopping-deepfake-porn-of-me-is-a-lost-cause-2018-12?r=US&IR=T>.

J.M. Porcup “What are deepfakes? How and why they work” (1 August 2018) CIO New Zealand <https://www.cio.co.nz/article/644646/what-deepfakes-how-why-they-work/>.

Jane Adams "'Distributed Courts': AVL in New Zealand’s Courts" LawTalk (912, Online ed, New Zealand).

Jeremy Hsu “Can AI Detect Deepfakes to Help Ensure Integrity of U.S. 2020 Elections?” (28 February 2019) IEEE Spectrum <https://spectrum.ieee.org/tech-talk/robotics/artificial-intelligence/will-deepfakes-detection-be-ready-for-2020>. See also:

Jon Fingas "AI-altered video makes it look like you can dance" (27 August 2018) <engadget.com>.

Jordan G Teicher "What do Facial Recognition Technologies Mean for our Privacy?" (18 July 2018) New York Times <https://www.nytimes.com/2018/07/18/lens/what-do-facial-recognition-technologies-mean-for-our-privacy.html?smid=tw-nytimesphoto&smtyp=cur>.

Jose Manuel Perez Tornero et al "How to confront fake news through literacy? State of the art" (2018) 26 DOXA COMUNICACION 211 to 235.

Joshua Rothman "In the Age of A.I., is Seeing Still Believing?" (5 November 2018) The New Yorker <https://www.newyorker.com/magazine/2018/11/12/in-the-age-of-ai-is-seeing-still-believing>.

Jurgen Schidhuber "Deep Learning in Neural Networks: An Overview" (2015) 61 Neural Networks 85–117.

K Bouatouch and C Bouville Photorealism in Computer Graphics (1st Ed, Springer-Verlag, Berlin Heidelberg, 1992).

Karoun Demirjian "Top Senate intel Democrat proposes measures to counter influence campaigns on social media" (30 July 2018) The Washington Post <https://www.washingtonpost.com>.

Katie Kenny "Q+A: Troll hunter Ginger Gorman on the Christchurch mosque shootings and cyberhate" Stuff.co.nz (3 April 2019) <https://www.stuff.co.nz/national/christchurch-shooting/111743226/qa-troll-hunter-ginger-gorman-on-the-christchurch-mosque-shootings-and-cyberhate>.

Kim Willsher, Oliver Holmes "Conmen Made €8M By Impersonating French Minister - Israeli Police" The Guardian </www.theguardian.com/world/2019/mar/28/conmen-made-8m-by-impersonating-french-minister-israeli-police>.

Kinnunen, Tomi & Lorenzo-Trueba, Jaime & Yamagishi, Junichi & Toda, Tomoki & Saito, Daisuke & Villavicencio, Fernando & Ling, Zhenhua "A Spoofing Benchmark for the 2018 Voice Conversion Challenge: Leveraging from Spoofing Countermeasures for Speech Artifact Assessment" (paper presented at Odyssey, Les Sables d'Ollone, France, 2018).

Lange v Atkinson [2000] 3 NZLR 385 (CA) at 396.

Lee McIntyre "Lies, damn lies and post-truth" (19 November 2018) The Conversation <https://theconversation.com/lies-damn-lies-and-post-truth-106049>.

Lisa Pitney (Vice President of Government Relations at Walt Disney Corporation) to various Senators (of the New York State Assembly ) requesting their opposition to Bill A.8155B (8 June 2018) <www.rightofpublicityroadmap.com/sites/default/files/pdfs/disney_opposition_letters_a8155b.pdf

Lucas Theis "Fast Face-swap Using Convolutional Neural Networks" (paper presented at IEEE International Conference on Computer Vision, Venice, 2017).

Lucinda Southern "How Reuters is training reporters to spot 'deepfakes'" (26 March 2019) Digiday <https://digiday.com/media/reuters-created-a-deepfake-video-to-train-its-journalists-against-fake-news/>.

Lyrebird "Lyrebird - Create a digital copy of your voice" (4 September 2017) <www.youtube.com/watch?v=YfU_sWHT8mo&feature=youtu.be>.

Maiella Moon "Australia's new law threatens social media companies with jail, fines" (4 April 2019) Engadget <www.engadget.com/2019/04/04/australia-laws-social-media-fines-jail/>.

Malicious Deep Fake Prohibition Bill 2018 (S.3805) <www.govinfo.gov/content/pkg/BILLS-115s3805is/pdf/BILLS-115s3805is.pdf>.

Mariusz Flasiński Introduction to Artificial Intelligence (Springer International Publishing, Switzerland, 2016).

Martin Giles "The GANfather: The man who's given machines the gift of imagination" (21 February 2018) MIT Technology Review <https://www.technologyreview.com/s/610253/the-ganfather-the-man-whos-given-machines-the-gift-of-imagination/>.

Matthias Niessner "Face2Face: Real-time Face Capture and Reenactment of RGB Videos (CVPR 2016 Oral) 17 March 2016 <www.youtube.com/watch?v=ohmajJTcpNk&feature=youtu.be>.

Max Towle "Deepfakes: When seeing is no longer believing" (18 May 2018) Radio New Zealand <https://www.radionz.co.nz/news/the-wireless/375262/deepfakes-when-seeing-is-no-longer-believing>.

Melanie Navamanikkam, "Truth in Advertising: Should America Ban Photoshop?" University of Cincinnati Law Review (21 June 2017).

Mike Elgan "Here comes 'antidisinformation as a service'" (26 August 2018) CIO NZ <www.cio.co.nz>/.

Monkey Cage "Fake news is about to get a lot worse" (3 April 2018) The Washington Post <https://www.washingtonpost.com/news/monkey-cage/wp/2018/04/03/fake-news-is-about-to-get-a-lot-worse-that-will-make-it-easier-to-violate-human-rights-and-get-away-with-it/>.

Morgan Wright, "The age of deepfakes: When seeing is no longer necessarily believing" (23 January 2019) The Hill <https://thehill.com/opinion/technology/426536-the-age-of-deepfake-when-seeing-is-no-longer-necessarily-believing>.

MPAA "Memorandum in Opposition to New York Assembly Bill A.8155B (Morelle, Right of Publicity)" <www.rightofpublicityroadmap.com/sites/default/files/pdfs/mpaa_opposition_to_a8155b.pdf>.

Namrata Maheshwari "Indian High Court opens the door to the own name defence" (2018) 13 Journal of Intellectual Property Law & Practice 527 to 529.

NBCUniversal "Memorandum in Opposition to New York Assembly Bill A08155B (Right of Publicity)" 8 June 2018 <https://www.rightofpublicityroadmap.com/sites/default/files/pdfs/nbc_opposition_a8155b.pdf>.

Neil Davey "Automation apocalypse: Proof that brands have lost sight of the human experience" (24 April 2017) <mycustomer.com>.

Nesli Erdogmus "Spoofing Face Recognition With 3D Masks" (2014) 8 IEEE Transactions on Information Forensics and Security 1084 to 1097.

New York Assembly Bill A08155 2018. <nyassembly.gov/leg/?default_fld=&leg_video=&bn=A08155&term=2017&Summary=Y&Text=Y>.

New Zealand Bill of Rights Act 1990.

New Zealand Law Commission "Review of the Privacy Act 1993: Review of the Law of Privacy Stage 4" Issues Paper 17, March 2010 at 3.5, available from: <www.lawcom.govt.nz/>.

Niam Yaraghi "Regulating free speech on social media is dangerous and futile" (21 September 2018) Brookings Institute <www.brookings.edu/blog/techtank/2018/09/21/regulating-free-speech-on-social-media-is-dangerous-and-futile/>.

NVIDIA "GauGAN: Changing Sketches into Photorealistic Masterpieces" (18 March 2019) <www.youtube.com/watch?v=p5U4NgVGAwg&feature=youtu.be> ; NVIDIA "Research at NVIDIA: AI Reconstructs Photos with Realistic Results" (22 April 2018) <www.youtube.com/watch?v=gg0F5JjKmhA>.

Oscar Schwartz "You thought fake news was bad? Deep fakes are where the truth goes to die" The Guardian (12 November 2018) <https://www.theguardian.com/technology/2018/nov/12/deep-fakes-fake-news-truth>.

Paul McDonald "So, about our name..." (13 September 2007) <blog.bodega.ai>.

Peters v The Electoral Commission [2016] NZHC 394.

Petra Butler: "A Dworkinian Right to Privacy in New Zealand" in Salman Khurshid, Lokendra Malik and Veronica Rodriguez-Blanco (eds) Dignity in the Legal and Political Philsophy of Ronald Dworkin (Oxford University Press, India, 2018) pp 433-465

Petra Butler "The Case for a Right to Privacy in the New Zealand Bill of Rights Act" (2013) 11(1) New Zealand Journal of Public and International Law Special Issue - 21st Birthday of the New Zealand Bill of Rights Act 1990 pp 213-256.

Plato The Republic (2nd ed, Penguin Group, London, 2007) 53 to 76 and 335 to 349.

Privacy Act s 42, Privacy Bill at Clause 62.

Privacy Bill 2018 (34-2) Clause 6.

Protection from Online Falsehoods and Manipulation Bill 2019 (10/2019) <https://sso.agc.gov.sg/Bills-Supp/10-2019/Published/20190401?DocDate=20190401>.

R v Williams [2007] NZCA 52, [2007] 3 NZLR 207.

Robert Chesney, Danielle Citron "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security" (2018) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3213954>

Robert Chesney, Danielle Citron "Deep fakes: A Looming Crisis for National Security, Democracy and Privacy?" (21 February 2018) Lawfare <https://www.lawfareblog.com/deep-fakes-looming-crisis-national-security-democracy-and-privacy>

Robert Chesney, Danielle Citron "Deepfakes and the New Disinformation War" (January 2019) Foreign Affairs <www.foreignaffairs.com/articles/world/2018-12-11/deepfakes-and-new-disinformation-war>.

Robert Hranitzky "How to remove objects in video with Content-Aware Fill in Adobe After Effects" (3 April 2019) <www.youtube.com/watch?v=gg0F5JjKmhA>.

Rokoko "Online demo of Smartsuit Pro" (17 October 2017) <www.youtube.com/watch?v=Y_9TZHGswVA>.

Ronald Poppe "A survey on vision-based human action recognition" (2010) 28 Image and Vision Computing.

Rosina Zapparoni "Propertising Identity: Understanding the United States Right of Publicity and Its Implications - Some Lessons for Australia" (2004) 23 MelbULawRw; (2004) 28(3) Melbourne University Law Review 690.

SAG-AFTRA "SAG-AFTRA Backs Legislation to End Nonconsensual Digital Sex Scenes and Nudity" (28 March 2019 <www.sagaftra.org/sag-aftra-backs-legislation-end-nonconsensual-digital-sex-scenes-and-nudity>.

Sam Gregory "Deepfakes and Synthetic Media: What should we fear? What can we do?" WITNESS Blog <https://blog.witness.org/2018/07/deepfakes/>

Sam Gregory "Deepfakes and Synthetic Media: Survey of Solutions against Malicious Usages" WITNESS Blog <https://blog.witness.org/2018/07/deepfakes-and-solutions/>.

Samantha Cole "'Deep Voice' Software Can Clone Anyone's Voice With Just 3.7 Seconds of Audio" (8 March 2018) Motherboard <https://motherboard.vice.com/en_us/article/3k7mgn/baidu-deep-voice-software-can-clone-anyones-voice-with-just-37-seconds-of-audio>

Samantha Cole "People Are Using AI to Create Fake Porn of Their Friends and Classmates" (27 January 2018) Motherboard <https://motherboard.vice.com/en_us/article/ev5eba/ai-fake-porn-of-friends-deepfakes>

Samantha Cole "AI-Assisted Fake Porn Is Here And We're All Fucked" (12 December 2017) Motherboard (VICE) <https://motherboard.vice.com/en_us/article/gydydm/gal-gadot-fake-ai-porn>.

Scholz, L. H. (2016). Privacy as quasi-property. Iowa Law Review, 101(3), 1113-1141

Senator Mark R. Warner Potential Policy Proposals for Regulation of Social Media and Technology Firms (US Senate, draft white paper).

Shannon K. Crawford, Kyra Phillips, Allie Yang "Seeing but not believing: Inside the business of "deepfakes" (10 December 2018) ABC News <www.abcnews.go.com/Technology/believing-inside-business-deepfakes/story?id=59731790>.

Simon Gibbs, Costas Arapis, Christian Breieneder, Vali Lalioti, Sina Mostafawy, Josef Speier "Virtual Studio: An Overview" (1998) 5 IEEE MultiMedia 18 to 35.

Siwei Lyu, Hany Farid "How realistic is photorealistic?" (2005) 53 IEEE Transactions on Signal Processing 845 to 850.

Stevan Price "What the new public interest defence really means for media and defamation" (2 August 2018) The Spinoff <https://thespinoff.co.nz/media/02-08-2018/heres-what-the-public-interest-defence-really-means-for-media-and-defamation/>.

Supasorn Suwajanakorn et al "What Makes Tom Hanks Look Like Tom Hanks" (paper presented at IEEE International Conference on Computer Vision, Santiago, Chile, 2015).

Taiamiti Edmunds, Alice Caplier "Face spoofing detection based on colour distortions" (2017) 7 IET Biometrics 27.

Taylor v Corrections [2018] NZHRRT 35.

Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., & Nießner, M. "Face2face: Real-time face capture and reenactment of rgb videos." (2016) Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2387-2395.

Thomas B Moeslund, Adrian Hilton, Volker Kruger "A survey of advances in vision-based human motion capture and analysis" (2006) 104 Computer Vision and Image Understanding 90 to 126.

Thomas Thorn "Background noise reduction: one of your smartphone's greatest tools" <www.techradar.com/au/news/phone-and-communications/mobile-phones/background-noise-reduction-one-of-your-smartphone-s-greatest-tools-1229667>.

Ting-Chun Wang et al "Video-to-Video Synthesis" <https://arxiv.org/abs/1808.06601>.

Todd Haselton "After almost a decade and billions in outside investment, Magic Leap's first product is finally on sale for $2,295. Here's what it's like." (8 August 2018) CNBC <www.cnbc.com>.

Tony Ezzat, Tomaso Poggio "Visual speech synthesis by morphing visemes" (2000) 38 International Journal of Computer Vision 45.

Truepic <https://blog.witness.org/2018/07/deepfakes-and-solutions/>.

United States Constitution, First, Third, Fourth, Fifth and Ninth Amendments

Visual Investigations, The New York Times <https://www.nytimes.com/interactive/2018/world/visual-investigations.html>,

Watson v Capital & Coast District Health Board [2015] NZHRRT 27 (7 July 2015).

Will Knight "The Defense Department has produced the first tools for catching deepfakes" (7 August 2018) MIT Technology Review <https://www.technologyreview.com/s/611726/the-defense-department-has-produced-the-first-tools-for-catching-deepfakes/?utm_medium=social&utm_campaign=owned_social&utm_source=twitter.com>

Will Knight "Fake America Great Again" (17 August 2018) MIT Technology Review <https://www.technologyreview.com/s/611810/fake-america-great-again/>.

William L Prosser "Privacy" (1960) 48 Cal LR 383.

WITNESS <https://witness.org/>

Woodrow Hartzog "The Public Information Fallacy" (2019) 99 Boston University Law Review 459.

Yann LeCun, Yoshua Bengio, Geoffrey Hinton "Deep Learning" (2015) 521 Nature 463 to 444.

Yasmin Green "Fake video will soon be good enough to fool entire populations" (12 January 2019) Wired <https://www.wired.co.uk/article/deepfake-videos-security>.

Ying Zhang et al "Automated Face Swapping and Its Detection" (paper presented to IEEE 2nd International Conference on Signal and Image Processing, Singapore, August 2017).

Zack Whittaker "OpenAI built a text generator so good, it's considered too dangerous for release" (February 2019) TechCrunch <www.techcrunch.com/2019/02/17/openai-text-generator-dangerous/>.