October 2019

DEEPFAKES, HARMFUL SYNTHETIC MEDIA AND LEGAL CERTAINTY IN NEW ZEALAND

PURPOSE OF DISCUSSION AND MEETING

1. **For Government**, to make agencies aware of synthetic media technologies, their potential harms, and government responsibilities for those harms.
2. **For the public**, to generate certainty about individual rights and responsibilities with regard to potentially harmful uses of synthetic media technologies.
3. **As a result**, to deter harmful conduct and content by signalling clear standards to potential users.
4. To identify the kinds of evidential services that government agencies and the public might require, and where those services can be obtained, in New Zealand or internationally, for the purposes of an Online Safety Grant awarded by Netsafe.

BACKGROUND TO SUBJECT MATTER

This paper follows a New Zealand Law Foundation-funded report by Curtis Barnes and Tom Barraclough on whether New Zealand's legal system is prepared for the harms that might arise from synthetic media technologies.

Synthetic media technologies use computer systems to modify digital audio and visual information: they allow users to "make it look (or sound) like something happened when it didn't happen". These technologies cannot be separated from a wider context in which increased creation and consumption of audiovisual information is made possible by smartphones and the internet, and where there is wider distrust in the reliability of online information.

Synthetic media technologies present two problems:

- To the extent that the existence and capabilities of synthetic media technologies **are not widely understood**, they can be used to deceive people individually or at scale, or to appropriate an individual's audiovisual profile in harmful ways.
- To the extent that their existence and capabilities **are widely understood**, they create the potential for justified or unjustified doubt about the reliability of audiovisual evidence. This in turn may fuel unnecessary disputes or undermine important evidence of wrongdoing.

There has been some suggestion that existing laws are not sufficient to deal with synthetic media harms. We found that existing laws do touch on the harms that may flow from

synthetic media technologies, however because those harms are diverse, a range of government agencies carry some responsibility for legal regimes governing them. It is difficult to be confident about whether existing laws will provide effective remedies without closer attention to specific fact patterns in individual complaints.

This paper aims to summarise our findings as well as facilitate discussion between selected government agencies. Our purpose is to reach some kind of consensus on our findings. This will allow for gaps to be identified and filled.

Importantly, the certainty generated by this forum will also guide members of the public and other government agencies to know where to turn if they perceive that harm has been caused by synthetic media technologies.

A significant outcome of this meeting is that there will be certainty for the public, including a degree of deterrence where appropriate through signalling legal consequences for certain kinds of actions.

RECENT EXAMPLES

Below are some examples of synthetic media technologies that illustrate our thinking on the topic and provoke discussion about the harms that may arise from uses of those technologies. We also indicate when the technology came to prominent public attention.

It is vital to consider not only how these technologies could be used in isolation, but also in combination with each other.

DEEPNUDE (JULY/AUGUST 2019)

DeepNude refers to a software application that uses neural networks to take an image of an identifiable cis-gendered woman and generate an image of them as if they were naked. The app was available for $50. After being uncovered by Motherboard (a news organisation), it was taken down, but its source code remains available online and its creators recently attempted to sell intellectual property associated with the app.

FACEAPP (JULY/AUGUST 2019)

FaceApp was a software application that uses neural network technology to artificially age or de-age photographs of identifiable individuals. It raised significant discussion about privacy and the company that developed the app was based in Russia. There were many situations of people using the app on photographs of other people in their communities. There was at least one example of it being used in a tweet to undermine the credibility of prominent Guardian journalist and Pullitzer Prize-winner Carole Cadwalladr who broke the Cambridge Analytica stories detailed in Netflix documentary “the Great Hack”.

FACETIME FOR IOS 13 EYE CONTACT (JULY/AUGUST 2019)

Media organisations are widely reporting that the next version of FaceTime – a video-calling app native to Apple’s iOS platform – will utilise augmented reality technology to subtly alter a user’s face to give the impression the user is making eye contact with the other party to the conversation. This is thought to enhance the sense of personal connection between users which is currently undermined by the fact that users don’t

appear to be making eye contact with each other due to placement of the camera in relation to the phone's display.

REPLICA AI (MAY 2019)

Replica AI is a Brisbane-based company that "is creating a marketplace of the world's voices. We can replicate anyone's voice with a few minutes of their recorded speech. Protect, Licence, Monetize". In May, it released an example of its technology generating vocal clips of podcaster and media personality Joe Rogan, who has a regular audience of millions of people globally. Another demonstration features known personalities like Arnold Schwarzenegger.

ADOBE TECHNOLOGIES INCLUDING PHOTOSHOP

Photoshop is a software product made by Adobe, a large technology company. To "photoshop" something has entered common parlance in the same way as "googling" something. Photoshop requires time, training and experience to use well, but can be used to alter still images or individual frames in videos.

Adobe's work on synthetic media technologies is not limited to Photoshop. A demonstration on YouTube from 2016 illustrates the capabilities of "VoCo", which can create audio clips of human speech that sounds like an identifiable individual and can be generated from simple text.

iOS apps such as "Photoshop Fix" allow the use of image in-painting to remove details from photographs and "Photoshop Mix" allows the merging of photographs to insert elements into photos. These can be used on most smartphones without any special expertise or skill.

Commercial technologies also allow for "content-aware in-painting", that effectively allows relatively unskilled individuals to remove objects from images and use AI technologies to replace those objects with persuasive in-painting of what a viewer would otherwise expect to see in place of that object.

HOLLYWOOD: THE DIGITAL EFFECTS AND ENTERTAINMENT INDUSTRIES

The capabilities of synthetic media technologies are plain to see in entertainment industries using audio and visual effects artistry. Many of these technologies do not entail the same use of automation and neural networks as DeepFakes or other emerging technologies, but with enough human and capital resources of the right kind, a large motivated actor could readily create synthetic media artefacts that could mislead or deceive.

GAN SYNTHESIS: THISPERSONDOESNOTEXIST.COM / WHICHFACEISREAL.COM

Two websites illustrate the capabilities of Generative Adversarial Networks at generating highly photorealistic images of human faces. Thispersondoesnotexist.com illustrates the capabilities of the algorithms and whichfaceisreal.com allows users to test their ability to distinguish between a real photograph and a generated image, and also gives guidance as to indicia which allow discrimination between the images.

DIGITAL HUMANS AND HUMAN AVATARS

There is an emerging industry around the use of highly realistic avatars of real humans and mixtures of real humans.

Relatedly, an emerging area of research attempts to generate human gestures in a video avatar from audio clips of speech.

BACKGROUND BLUR IN SKYPE, VIRTUAL BACKGROUND FOR ZOOM

Video-conferencing software such as Skype and Zoom allows users to either blur or replace the area surrounding their face for privacy or novelty reasons. While generally innocuous, this illustrates the way that synthetic media technologies can be integrated into platforms that are otherwise intended to provide consistent and reliable audio and visual representations of events occurring in real time. It is feasible that synthetic media technologies could also be used to alter other relevant features in the frame.

COMPUTATIONAL PHOTOGRAPHY: SMARTPHONE CAMERAS

Smartphones capture, process and display light and sound information. In doing so, they alter that information. Frequently, a camera will enhance or reduce desirable or undesirable features, changing colour profiles, removing blemishes, filling in gaps caused by inferior sensor performance, etc.

DIRECT AND INDIRECT HARMS

Synthetic media can be used in a direct way to cause harm, for example by fraud or non-consensual pornography. It can also have less direct harms: specifically, a wider loss of faith in the reliability of audiovisual information can create doubt in situations where it did not exist previously.

Frequently, audio or video recordings are used to resolve disputes or create certainty about how contested facts actually occurred. If such recordings are no longer perceived as reliable, then they lose that capacity. Further, it is not necessary that an audiovisual artefact is conclusively shown to be manipulated to undermine its credibility. The mere suggestion that it can be undermined creates uncertainty or delays while the audiovisual record is verified. It can be difficult to prove a negative – ie that something has not been manipulated. All digital media is manipulated to some degree for accepted purposes.

LEGAL VS NON-LEGAL RESPONSES

A holistic response to synthetic media risks and opportunities must focus on legal and non-legal (or extra-legal) responses to potential harms. There is a risk that too much emphasis is placed on systems of law as being the primary means of response to perceived risks of synthetic media.

In practice, there will be many reasons why an arguable grievance is not brought to formal dispute resolution mechanisms, including:

1. where access to justice barriers such as cost, delay, poor access to or understanding of the law,

2. where the law is uncertain or ambiguous,
3. where the harm is *de minimis*, or the financial value of a remedy outweighs the cost of seeking the remedy,
4. where individuals "fall through the cracks" between government agencies,
5. where legal remedies would be (or are perceived to be) ineffective,
6. where a responsible agent cannot be identified, or
7. where there are other more effective means of dispute resolution.

Frequently, the primary form of pseudo-regulatory guidance about what is acceptable online is the use of community guidelines or terms of service adopted by social media platforms such as YouTube, Facebook or Twitter. This space is developing rapidly. Any governmental response should account for the existing content of these guidelines in order to be effective.

We propose that a strictly legal approach is worth exploring despite the increased prominence of extra-legal dispute resolution or regulatory mechanisms. The legal system will still need to provide a remedy in situations of extreme harm or where all other approaches are ineffective.

RISKS OF USING LAW TO INTERVENE

The New Zealand Bill of Rights Act 1990 grants the right to freedom of expression in s 14, including the right to receive and impart information of any kind in any form. The NZBORA must be seen in light of ss 4, 5 and 6 which affect the application of the NZBORA in New Zealand and applies only to actors caught by s 3 of the Act.

The NZBORA needs to also be seen in the context of a range of other laws in New Zealand that protect definable interests in specific contexts. Many laws may protect freedom of expression or individual privacy, for example, without being explicitly drafted in terms of "the right to freedom of expression".

Discussions about synthetic media regulation have strong overlaps with wider discussions about content moderation practices and standards in the modern world.

HIGH PRIORITY QUESTIONS

We identify the following specific questions to generate discussion and consensus on the state of the New Zealand legal system in responding to synthetic media.

(1) Does your agency have any responsibility for harmful use of synthetic media?

(2) You're confronted with an allegation that you ought to exercise your jurisdiction in some way in relation to an audio/visual record on the basis that it is "fake" or has been "manipulated", or is in some other way unreliable.

   a. Is the "fakeness" or "reliability" of the audiovisual record relevant to your jurisdiction?

   b. Is it for you or for the applicant to establish this fact?

c. Where do you turn to in order to establish whether this is the case or not?

d. Is this an expert question?

e. What evidential standard are you required to apply?

(3) Do the Evidence Act 2006, the Evidence Regulations 2007 (relating to video records), and ss 258-259 of the Crimes Act cover the conceivable uses of altered video evidence in a proceeding covered by the Evidence Act?

(4) Would the use of the DeepNude app or other synthetic media technologies to generate an image of a person as if they were naked, without their consent:

a. be caught by the definition in s 216G of the Crimes Act?

b. be subject to the Privacy Principles by falling foul of s 56(2) of the Act?

(5) Does s 249 of the Crimes Act criminalise use of a computer system to use synthetic media technologies with dishonest intent?

(6) What is the public consensus on whether synthesis of non-consensual pornographic material is different from capturing non-consensual intimate images?

(7) Does the use of synthetic media technologies to generate information about an identifiable individual constitute "collection" under the Privacy Act?

(8) Is a synthetic media artefact about an identifiable individual personal information under the Privacy Act?

(9) Is the Electoral Act at ss 199A, 197 sufficient to deal with influence campaigns around elections?

(10) Are there any other potential problems you can identify based on the way your government department operates? Consider its day-to-day operation, as well as powers and duties granted to it under legislation or regulation.

RESULTING NEED FOR EVIDENTIAL SERVICES

Answering the legal questions above will indicate what kinds of evidential services are required to prove matters in issue in any complaints or tribunal-based proceeding.

Based upon our own analysis, we suggest that the following kinds of evidential services will be required:

- Identification evidence of the kind used in CCTV, perhaps including facial recognition software.
- Semantic evidence (is the content of the image consistent internally and with its alleged context).
- Expert technological detection of manipulation (eg pixel-by-pixel analysis).
- Image hashing and use of reverse image search to identify originals.

(11) Where would you currently seek evidence of this kind? Are there any other disciplines or areas of expertise that you believe could assist you in determining evidential questions about synthetic media?

LIST OF STAKEHOLDERS AND PARTICIPANTS

The following are suggested as interested parties who should be invited to attend or have input in some form or other.

GOVERNMENT OR PUBLIC FUNCTION

1. Office of Film and Literature Classification
2. Office of the Privacy Commissioner
3. NZ Police
4. Electoral Commission
5. Broadcasting Standards Authority
6. Media Council
7. Human Rights Commissioner
8. Commerce Commission
9. Netsafe
10. Department of Prime Minister and Cabinet
11. Department of Internal Affairs
12. Advertising Standards Authority
13. Ministry of Justice

NON-GOVERNMENT ENTITY OR PRIVATE FUNCTION

There are a range of other institutions who could add to this discussion from traditional media, social media, private enterprise and academia. At some point, they will be included, however we wish to generate some government consensus on these issues first to provide structure to subsequent discussion.

CONCLUSION

Your feedback and interaction with us will be fundamental to how New Zealand responds to synthetic media in the near future. It will also be invaluable for us as we move towards production of guides and resources for the public and government agencies as an output of our Online Safety Grant, funded by Netsafe. New Zealand has demonstrated it can be an international leader in addressing online harms.

Tom Barraclough and Curtis Barnes
Brainbox Ltd

www.brainbox.institute
info@brainbox.institute

BRAINBOX
EMERGING TECH + LEGAL RESEARCH