Legal Opinion on Deepfakes and Platform Responsibility

Abstract: In the digital media landscape, deepfakes are the latest, most sophisticated vehicles of disinformation and misinformation—and online platforms risk becoming the highways that enable their breakneck-paced dissemination. This Legal Opinion explores in depth the role that platforms can, and ought, to play in the response to deepfakes. First, I argue that current legal interventions do not go far enough to tackle deepfake threats; an effective response strategy must involve online platforms as the frontline responders in a deepfake-incited crisis. Next, in view of a normative shift towards greater platform responsibility, I discuss compelling legal and business reasons for platforms to tackle deepfake disinformation sooner rather than later. To help platforms achieve a balance between taking down deepfake content and respecting fundamental rights, I propose a number of ways for policymakers and regulators to update the legal framework. Finally, I consider the main practical challenges presented by deepfake content moderation, which range from identification of deepfakes to ensuring due process and transparency. I also evaluate the content policies of four major platforms, from which I draw some general suggestions for platforms seeking to update their own guidelines to contend with deepfakes. However, platform action alone is no panacea to the spread of disinformation and misinformation via deepfakes or otherwise. It is hoped that this Legal Opinion will serve as a basis for stakeholder collaboration towards that aim.

Introduction[1]

In the digital media landscape, deepfakes are the latest, most sophisticated vehicles of disinformation and misinformation—and online platforms risk becoming the highways that enable their breakneck-paced dissemination.[2] 'Deepfakes' are a genre of fake video/audio in which a person's likeness or voice is generated using machine learning techniques and then manipulated to make them appear to say or do certain things.[3] Since their entry into public consciousness, deepfakes have become the subject of great controversy and concern. By tapping into the human tendencies to vividly recall and share visual content like videos and images, malicious deepfakes have a unique potential to harm individuals and undermine broader democratic processes.[4]

Recognising the threat to the information ecosystem that deepfakes pose, numerous writers and policymakers have now openly called for platforms to take action.[5] This Legal Opinion builds upon their suggestions by exploring in depth the role that platforms can, and ought, to play in the response to deepfakes.

Chapter 1 assesses key threat scenarios involving deepfakes. I argue that existing legal interventions, though promising in some respects, do not go far enough to tackle the unique challenges posed by deepfakes. Instead, an effective defence strategy must involve online platforms as the likely frontline responders in a deepfake-incited crisis.

Chapter 2 examines the legal basis for platforms to take action against deepfakes. Traditionally, platforms enjoy strict or qualified immunity for user-generated content, but

[1] Many thanks to my supervisor, Professor Frederick Mostert, for helping to bring this project to life through his unparalleled support, insight, and enthusiasm.

[2] While there is no consensus definition of 'platform', the following description encapsulates the sense in which I use the term:

> "Online platforms share key characteristics including the use of information and communication technologies to facilitate interactions (including commercial transactions) between users, collection and use of data about these interactions, and network effects which make the use of the platforms with most users most valuable to other users."

'Online Platforms (Accompanying the Document Communication on Online Platforms and the Digital Single Market)' (European Commission) SWD(2016) 172 final.

[3] Henry Ajder and others, 'The State of Deepfakes: Landscape, Threats, and Impact' (Deeptrace 2019).

[4] Cristian Vaccari and Andrew Chadwick, 'Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News' (2020) 6 Social Media + Society 2.

[5] Refer to Chapter 1.3.

this orthodox legal position has become increasingly overshadowed by emerging norms of platform responsibility. In light of this development, I discuss a compelling legal and business case for platforms to tackle deepfake disinformation sooner rather than later. I conclude with a series of recommendations regarding what regulators and policymakers can do to better facilitate platforms' response to deepfakes.

Finally, Chapter 3 focuses on the content moderation measures and policies that platforms can implement in response to deepfakes. I discuss the main practical challenges presented by deepfake content moderation, which range from identification of deepfakes to ensuring due process and transparency. I also evaluate the content policies of four major platforms, from which I draw some general suggestions for platforms seeking to update their own guidelines to contend with deepfakes.

At the time of writing, several platforms have taken the initiative to respond to deepfakes by contributing to the development of detection technologies or updating their policies. The measures proposed herein are intended to bolster such good faith efforts. That being said, platform action alone is no panacea to the spread of disinformation and misinformation via deepfakes or otherwise. It is hoped that this Legal Opinion will serve as a basis for stakeholder collaboration towards that aim.

1. Deepfake Threats and Current Legal Solutions

The purpose of this chapter is to analyse, in turn, the potential applications of deepfakes by bad actors and the adequacy of current legal solutions. Non-consensual pornography, political disinformation, and commercial manipulation are three deepfake threats which share a common thread: harm is almost always caused by online dissemination. Although numerous legal solutions have been suggested in the literature, I argue that none of them will be able to protect claimants in a timely and cost-efficient manner. Instead, an effective response strategy must engage online platforms as the probable first line of defence against malicious deepfakes.

1.1. Key Deepfake Threats

Like any new technology, deepfakes are a double-edged sword. Depending on the intentions of the actors behind their creation and dissemination, the synthetic media technology can be used for good or evil. Benign use cases of deepfakes include the development of creative works by the entertainment industry,[6] the preservation of sources' anonymity for the purposes of documentary filmmaking,[7] and the creation of deepfake videos by research institutions to educate the public on the technology's existence.[8]

However, the focus of this Legal Opinion is on deepfake uses that threaten to harm individuals, society, or both. The ultra-realistic depictions made possible by machine learning techniques create a potential for deception that can be exploited by a range of bad actors. These concerns led experts in a 2020 study to rate deepfakes as the most serious AI crime threat.[9] It is to three such uses that we now turn.

[6] James Vincent, 'Disney's Deepfakes Are Getting Closer to a Big-Screen Debut' (*The Verge*, 29 June 2020) <https://www.theverge.com/2020/6/29/21306889/disney-deepfake-face-swapping-research-megapixel-resolution-film-tv> accessed 21 August 2020.

[7] Joshua Rothkopf, 'Deepfake Technology Enters the Documentary World' (*The New York Times*, 1 July 2020) <https://www.nytimes.com/2020/07/01/movies/deepfakes-documentary-welcome-to-chechnya.html> accessed 30 August 2020.

[8] Jeffery DelViscio, 'A Nixon Deepfake, a "Moon Disaster" Speech and an Information Ecosystem at Risk' (*Scientific American*) <https://www.scientificamerican.com/article/a-nixon-deepfake-a-moon-disaster-speech-and-an-information-ecosystem-at-risk1/> accessed 21 August 2020.

[9] M Caldwell and others, 'AI-Enabled Future Crime' (2020) 9 Crime Science 14.

1.1.1. Non-consensual Pornography

To date, pornography is the most notorious and prevalent use case for deepfakes. In fact, the term was coined in 2018 when Reddit user 'deepfakes' published AI-generated porn videos that superimposed the faces of celebrity actresses and artists onto the bodies of porn performers.

In an interview with *Vox*, actress Kristen Bell described her shock in discovering that her face had been featured in pornographic videos without her consent. However, deepfake porn is not limited to celebrities, as one Australian law graduate discovered that her photos from social media had been photoshopped onto nudes and then used to generate deepfake porn videos which were published along with her full name.[10]

Non-consensual deepfake pornography has been widely denounced as a violation of individuals' privacy and a form of harassment (sexual and otherwise). It undermines a person's rights to bodily autonomy and dignity and can cause psychological distress. Furthermore, a 2019 report by Deeptrace (now Sensity), a cybersecurity company specialising in visual threat intelligence, found that women are exclusively targeted as the subjects of deepfake porn. [11] This finding suggests that the phenomenon may also reinforce damaging gender inequalities in a broader societal context. The same report noted that as many as 96% of deepfakes online today are pornographic.[12] In other words, this deepfake threat is no mere hypothetical—it is a reality already experienced by many women, and likely to affect many more.

Several major online platforms have taken steps to combat non-consensual deepfake pornography. For instance, Reddit shut down its subreddit r/deepfakes and updated its content policy to ban such content. [13] Meanwhile, Pornhub promised to take down deepfake porn and made it impossible to search for 'deepfakes' on its website, though observers have criticised the platform for poor enforcement of the ban.[14]

[10] Vox, *The Most Urgent Threat of Deepfakes Isn't Politics* <https://www.youtube.com/watch?v=hHHCrf2-x6w> accessed 20 August 2020.

[11] Ajder and others (n 3) 2.

[12] ibid 1.

[13] Samantha Cole, 'Reddit Just Shut Down the Deepfakes Subreddit' (*Vice*, 7 February 2018) <https://www.vice.com/en_us/article/neqb98/reddit-shuts-down-deepfakes> accessed 28 March 2020.

[14] Samantha Cole, 'The Ugly Truth Behind Pornhub's "Year In Review"' (*Vice*, 18 February 2020) <https://www.vice.com/en_us/article/wxez8y/pornhub-year-in-review-deepfake> accessed 29 March 2020.

Far from being slowed by these efforts, deepfake porn appears to be a growing phenomenon aided by two trends. First, the open-source technology to develop deepfakes has already been disseminated through a combination of public and private channels, providing interested users with the means to generate their own content. Second, the majority of deepfake pornography videos available online are hosted by dedicated deepfake pornography websites, many of which have not made public commitments to combat non-consensual content as mainstream platforms have done.[15] The harms highlighted above are likely to persist, as deepfake porn videos continue to be shared via a range of internet platforms.

1.1.2. Political Disinformation[16] and Propaganda

Deepfakes can also be used in political contexts to propagate false or misleading information by exploiting the image and associated ethos of a recognised political figure. Granted, the 'fake news' phenomenon is not a new plague to politics: in May 2019, a doctored video that depicted US Speaker of the House Nancy Pelosi slurring her words went viral on Facebook and Twitter, without the need for any AI-assisted manipulation.[17] The advent of deepfakes, which can appear particularly convincing to viewers at first sight, now threatens to accelerate the speed at which such disinformation is disseminated.[18]

Well-known examples of deepfakes include Buzzfeed's video of Barack Obama voiced by actor Jordan Peele warning against the threat of deepfakes to politics,[19] and two videos produced by Future Advocacy purporting to show Jeremy Corbyn and Boris Johnson

[15] Ajder and others (n 3) 4.

[16] The Oxford English Dictionary defines disinformation as '[t]he dissemination of deliberately false information, esp. when supplied by a government or its agent to a foreign power or to the media, with the intention of influencing the policies or opinions of those who receive it; false information so supplied.' 'Disinformation, n.' <https://www.oed.com/view/Entry/54579> accessed 23 June 2020.

[17] The term 'cheap fakes' is sometimes used to distinguish this type of relatively unsophisticated video manipulation from actual deepfakes. See **Appendix A** for a visualisation of the spectrum of AV manipulation.

[18] A 2018 study found that false news on Twitter spread significantly faster and farther than true news stories, owing to human responses to the novelty of false news, rather than bot intervention: Soroush Vosoughi, Deb Roy and Sinan Aral, 'The Spread of True and False News Online' (2018) 359 Science 1146.

[19] Craig Silverman, 'How To Spot A DeepFake Like The Barack Obama-Jordan Peele Video' (*BuzzFeed*) <https://www.buzzfeed.com/craigsilverman/obama-jordan-peele-deepfake-video-debunk-buzzfeed> accessed 31 August 2020.

endorsing each other for the 2019 UK general election.[20] The creators published these deepfakes with the goal of raising awareness about nefarious political uses. Among other things, political deepfakes could undermine trust in institutions, disrupt diplomacy, and distort the outcome of elections. One study found that even where viewers were not misled, exposure to deepfakes caused them to feel heightened uncertainty, reducing trust in news shared on social media.[21]

Besides the threat to democracy posed by actual deepfakes, the technology's very existence can undermine political discourse when politicians or citizens cry 'deepfake' without proof. For instance, when the president of Gabon made a rare appearance on social media to deliver a 2019 New Year's address to the nation, social media commentators and politicians speculated that the video had been a deepfake being used to cover up the president's death or illness. This claim was later repeated by members of the military who attempted a coup against the government. In June 2019, a Malaysian politician also alleged that a leaked sex tape of himself was a deepfake. Contrary to these accusations, forensic analysis of both videos did not reveal any manipulation.[22] These examples illustrate a particularly insidious consequence that Chesney and Citron call the 'liar's dividend': deepfakes provide an escape from accountability by allowing people to deny the truth of audio-visual evidence.[23]

1.1.3. Commercial Fraud and Market Manipulation

Thirdly, deepfakes could have potentially disastrous economic consequences if deployed in a commercial context. For instance, criminals could use deepfake voice phishing techniques to convincingly impersonate well-known business leaders in order to defraud an unsuspecting business or scam customers of their money. Other concerning applications include deepfake social botnets—a more sophisticated variation of existing bots on social media—and fabrications of remarks by public figures.[24]

[20] 'Deepfakes' (*Future Advocacy*) <https://futureadvocacy.com/deepfakes/> accessed 31 August 2020.

[21] Cristian Vaccari and Andrew Chadwick (n 4).

[22] Ajder and others (n 3).

[23] Robert Chesney and Danielle Keats Citron, 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security' (2019) 107 California Law Review 1753, 1785.

[24] Jon Bateman, 'Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenarios' (Carnegie Endowment for International Peace) "Cybersecurity and the Financial System" 7 2.

In addition, deepfakes could be deployed to manipulate the financial markets. One only needs to observe how Elon Musk's tweets have previously affected Tesla's share price to recognise how the social media activity of prominent figures can have an immediate impact on corporations and their shareholders.[25] While market reversals may be possible, the longer it takes to uncover or confirm that a video is a deepfake, the greater the likelihood that a business will suffer lasting financial or reputational damage.[26]

1.2. Current Legal Solutions

So far, a variety of legal mechanisms have been proposed to address the harmful uses of deepfakes outlined above. Some jurisdictions have opted to introduce new deepfake-specific legislation, while others already provide forms of legal protection which could be repurposed to cover deepfakes. What follows is a non-exhaustive overview of the main legal solutions to date which have been discussed in the literature or implemented by policymakers.

1.2.1. Image Rights

In the US, numerous states recognise a 'right of publicity' which protects an individual's right to control and commercialise various distinctive personal characteristics such as their name, likeness, and voice. The right of publicity is governed by state law through legislation, common law rights or a combination of both. Protection is typically available for commercial uses of a person's image, but some states go even further. The most notable of these is California's common law right to publicity, which protects the plaintiff's 'identity' from non-consensual use and appropriation of their name or likeness to a defendant's advantage, commercial or otherwise.[27]

It has been argued that the right of publicity is flexible enough to accommodate claims based on deepfakes.[28] In *In re NCAA Student-Athlete Name & Likeness Licensing*

[25] Russell Hotten, 'Elon Musk Tweet Wipes $14bn off Tesla's Value' *BBC News* (1 May 2020) <https://www.bbc.com/news/business-52504187> accessed 1 September 2020.

[26] Henry Ajder, 'Social Engineering And Sabotage: Why Deepfakes Pose An Unprecedented Threat To Businesses' (*Deeptrace*, 3 October 2019) <https://deeptracelabs.com/social-engineering-and-sabotage-why-deepfakes-pose-an-unprecedented-threat-to-businesses/> accessed 10 March 2020.

[27] *Eastwood v. Superior Court (National Enquirer, Inc.)* (1983) 149 Cal. App. 3d 409.

[28] Emma Perot and Frederick Mostert, 'Fake It till You Make It: An Examination of the US and English Approaches to Persona Protection as Applied to Deepfakes on Social Media' (2020) 15 Journal of Intellectual Property Law & Practice 32, 34.

Litigation[29] concerning video game depictions of college athletes, the defendant could not rely on the 'transformative use' defence as they had sought to portray the plaintiff as realistically as possible. This judicial treatment of expressive works as subject to the right to publicity bodes well for claimants seeking to invoke the right against deepfake uses of their identity—though their success in obtaining redress will depend on being able to identify an appropriate defendant, discussed at 1.3 below.

1.2.2. Passing Off

In the UK, a person seeking to protect against commercial misuse of their image through deepfakes can bring an action in the tort of passing off, if they can establish the elements of goodwill, misrepresentation and damage. Historically, the UK courts were reticent to recognise celebrity image protection through passing off: in the 1970s, the Swedish band ABBA brought a claim against the defendant's use of photos of them on merchandise, but failed to establish that customers would be deceived into thinking there was a commercial link between themselves and the defendant.[30]

Since then, however, celebrities have had more success in using passing off to defend their image against unauthorised commercial use. In *Irvine v Talksport Ltd*,[31] the court held that the tort of passing off covered cases of false celebrity endorsement where the claimant can prove a 'significant reputation or goodwill'. More recently, pop singer Rihanna brought a successful claim against Topshop for using her photograph on a T-shirt in a way that could mislead consumers into believing that she had authorised it.[32] Passing off should therefore provide a suitable cause of action for celebrity claimants in the UK who suffer damage to their goodwill or reputation due to misrepresentation via deepfakes.[33]

1.2.3. Defamation

The tort of defamation is another option for claimants, particularly where there has been non-commercial unauthorised use of their image. In the UK, defamation is governed by common law and the Defamation Act 2013. It covers a defendant's publication to a third party of material that is defamatory to the claimant. The term 'libel' is used to describe

[29] 724 F.3d 1268, 1279 (9th. Cir. 2013).
[30] *Lyngstad v Anabas Products Ltd* [1977] F.S.R. 62.
[31] [2003] EWCA Civ 423.
[32] *Robyn Rihanna Fenty v Arcadia Group Brands Ltd (T/A Topshop)* [2013] EWHC 2310 (Ch).
[33] Perot and Mostert (n 28) 36.

instances where the publication takes place in a written or other permanent form, which encompasses deepfake videos.

One hurdle to a defamation claim is the requirement of 'serious harm' to an individual's reputation imposed by s.1 Defamation Act 2013. In *Lachaux v Independent Print Ltd*,[34] the Supreme Court held that whether a statement 'has caused or is likely to cause serious harm' requires proof, as a matter of fact, of actual historic harm or probable future harm. This interpretation departs from the common law's lower threshold of harm, as defamatory character was previously determined by reference to the inherent character of the words and their *tendency* to cause harm.[35] Defamation thus offers a relatively limited sphere of protection to deepfake claimants.

Additionally, the rule of 'no prior restraint' makes it extraordinarily difficult to obtain interim injunctions, as the court will only grant one where it is satisfied that there is no arguable defence.[36] This may prove problematic in a deepfakes case where speed is of the essence in mitigating the harms of dissemination.

1.2.4. Privacy

An individual's right to privacy can also provide a cause of action against deepfakes in some jurisdictions.

In the UK, photos and videos are viewed as capable of interfering with a person's right to private life pursuant to Art.8 ECHR, since they allow the public to spectate upon whatever aspect of their life is captured by the medium.[37] A privacy claim will be assessed on a case-by-case basis, taking into account whether an individual has a reasonable expectation of privacy in the circumstances, and the claimant's right to privacy will have to be balanced against the competing right of the creator's freedom of expression and information.

However, as a shield against deepfakes, the right to privacy suffers from a major limitation. To determine an individual's reasonable expectation of privacy, a court will consider the source of the photos or videos used. Hence, Farish points out that a photograph of a college

[34] [2019] UKSC 27.
[35] *Thornton v Telegraph Media Group Ltd* [2011] 1 WLR 1985.
[36] *Bonnard v Perryman* [1891] 2 Ch 269.
[37] *Douglas & Ors v Hello! Ltd & Ors* [2005] EWCA Civ 595.

sports player which he consented to having featured on the sports team's website would likely receive a lower degree of protection than a photo uploaded by the same individual to his private social media account. This logic creates problems for individuals with a large public social media presence, who might find it difficult to argue they have a reasonable expectation of privacy that warrants protection against unauthorised deepfakes.[38]

Moreover, deepfake algorithms may blend audio-visual content from different sources such that it is impossible to determine the individual sources of the training material, nor the relative weight accorded to each when the final product was generated. A privacy claim in these circumstances would effectively require judges to confront the black box problem and apply the law to a hypothetical.[39]

1.2.5. Copyright

A copyright holder can file a request for the material to be taken down by a digital platform such as YouTube or Facebook, on the basis that the deepfake infringes their exclusive right of reproduction, or their moral rights as an author (available in many civil jurisdictions, but relatively limited in the UK and US).

However, copyright protection is only available to the original author of the work—in the case of deepfake videos, the videographer whose material was used to train the AI—or a later assignee of the copyright. As a result, copyright will not generally aid those who find themselves the unwilling subjects of deepfake videos.

1.2.6. *Sui Generis* Legislation

Several US states have passed legislation addressing specific uses of deepfakes. For instance, California now criminalises non-consensual deepfake pornography,[40] and also allows a political candidate featured in a deepfake to seek injunctive relief against any person who distributes such a deepfake within 60 days of election, unless they include a disclosure stating that the media has been manipulated.[41] Texas introduced a criminal

[38] Kelsey Farish, 'Do Deepfakes Pose a Golden Opportunity? Considering Whether English Law Should Adopt California's Publicity Right in the Age of the Deepfake' (2020) 15 Journal of Intellectual Property Law & Practice 40, 45.
[39] ibid.
[40] Cal. Civ. Code § 1708.86 (2019).
[41] Cal. Civ. Proc. Code § 35 (2019); Cal. Elec. Code § 20010 (2019).

offence for anyone who creates a deepfake video with the intent to influence an election or injure a candidate, and causes its publication or distribution within 30 days of an election.[42] Meanwhile, Virginia amended its existing criminal law on revenge pornography to include deepfakes.[43]

US federal legislators are also keeping an eye on the issue, with several deepfake-related bills being introduced in Congress in recent years. The most extensive of these is the DEEPFAKES Accountability Act of 2019, which if passed would require any deepfake image, audio or video to be labelled with a watermark.[44] Failure to do so could result in up to 5 years in prison, if the defendant had the requisite intent and knowledge.[45] Unlike existing state deepfakes legislation, the DEEPFAKES Act takes an omnibus approach attempting to cover all uses of deepfakes across different contexts, with limited exceptions.[46]

In China, legislators have taken a similar tack. A new regulation, effective from 1 January 2020, requires providers and users of online audio-visual news and information services to clearly label content generated using new technologies including deep learning and virtual reality. The Cyberspace Administration of China has stated that non-compliance can result in criminal liability.[47]

1.3. Analysis of Current Legal Solutions: Too Little, Too Late

While the technology to develop deepfakes is relatively new to the general public, scholars and legislators around the world increasingly view deepfakes as a priority for research and regulation. Yet **existing legal options are inadequate to address the threats of deepfakes used for non-consensual pornography, political purposes and commercial manipulation**, for two main reasons.

[42] Tex. Elec. Code § 255.004 (2019).
[43] Va. Code Ann. § 18.2-386.2 (2019).
[44] 'Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2019', H.R.3230, 116th Cong. (2019).
[45] Ibid.
[46] Matthew Ferraro, 'Deepfake Legislation: A Nationwide Survey—State and Federal Lawmakers Consider Legislation to Regulate Manipulated Media' (WilmerHale 2019).
[47] 'China Seeks to Root out Fake News and Deepfakes with New Online Content Rules' *Reuters* (29 November 2019) <https://www.reuters.com/article/us-china-technology-idUSKBN1Y30VU> accessed 1 April 2020.

First, most civil causes of action are designed to be initiated by individuals to obtain personal remedies, which makes them inherently limited in their ability to mitigate the harms caused by the rapid and broad dissemination of deepfakes. Once individual actors are made aware of a deepfake's existence, the onus lies wholly on them to justify takedown and claim damages. To make matters worse, once deepfake videos are disseminated online, they can be downloaded and re-uploaded by others, which can turn a claimant's quest for legal redress into a frustrating and seemingly futile endeavour akin to 'whack-a-mole'.

Second, bad actors online can often be difficult to track down. Even if the original creator of a deepfake is successfully identified, there may be practical issues in bringing a civil action for personal redress—for instance, if the creator is located in another jurisdiction, or if they lack the assets to pay the damages ordered by a court.[48]

Thus, from a practical standpoint, the fallout from a malicious deepfake can occur within the span of days or just hours. In contrast, court-ordered relief is likely to arrive months if not years after the fact, and will rarely be able to save an individual's reputation from irreparable damage, or halt the manipulation of political processes in a timely manner.

For these reasons, I suggest that **mitigation and even prevention of deepfakes harms can be achieved not by individual action alone, but through a broader legal and technological response that mobilises online platforms**. Multiple commentators have already recognised an urgent need for platforms to play an active role in the war against deepfake disinformation and misinformation.[49] Since the internet is the primary means by which a deepfake's reach can be maximised, interventions by platforms are crucial to ensure early detection and rapid action to limit their spread, thereby minimising the resulting harm to individuals and society at large. In the next chapter I consider the obligations and incentives that will shape the way that platforms respond to deepfakes.

[48] Jaani Riordan, *The Liability of Internet Intermediaries* (First edition, Oxford University Press 2016) ch 3.

[49] Edvinas Meskys and others, 'Regulating Deep Fakes: Legal and Ethical Considerations' (2020) 15 Journal of Intellectual Property Law & Practice 24, 31; Chesney and Citron (n 23) 1817; Perot and Mostert (n 28) 37; Mark Warner and Marco Rubio, 'Deepfakes Letter to Facebook' (*Scribd*) <https://www.scribd.com/document/428320935/Deepfakes-Letter-to-Facebook> accessed 10 March 2020.

2. A Legal Framework for Platform Intervention

The purpose of this chapter is to consider the state of the legal framework within which platforms could take action on deepfakes, with particular emphasis on existing obligations and incentives to moderate (or not moderate) user-generated deepfake content. Over time, legislative immunities for platforms have been eroded, while notions of platform responsibility have taken centre stage. Consequently, both legal and business incentives provide an impetus to tackle deepfake disinformation sooner rather than later. I conclude with several recommendations to improve the existing legal framework.

2.1. The Orthodox View: Limited or No Liability for User-generated Deepfake Content

In general, it will be difficult for claimants to establish that platforms owe primary liability for deepfake user-generated content that they host. As Chapter 1 demonstrated, deepfake claimants already face hurdles to demonstrate that their case is covered by an existing cause of action. The primary wrongdoer will typically be the user(s) who have posted the deepfake in question, not the platform itself.

Still, primary or secondary liability may attach to platforms in rare cases where the necessary elements of a cause of action are fulfilled. In the UK case of *Tamiz v Google*, it was argued that platforms can be the 'publishers' of defamatory content once notified.[50] The claimant sued Google for failing to take down comments on a Blogger post within a reasonable period after receiving complaints about their defamatory nature (Blogger being a Google-owned website). At first instance, Eady J held that Google could not be liable as a publisher, analogising the blog to a graffiti wall which Google had no responsibility to police. However, the Court of Appeal overturned the decision, suggesting that Blogger functioned more like a notice board subject to Google's power to block or remove material that did not comply with the terms of service. Thus, the Court considered that Google could be a secondary 'publisher' of defamatory material on Blogger in some circumstances, though it found otherwise on the facts.

[50] [2013] EWCA Civ 68.

Platforms have also been held liable for copyright infringement. In *Stichting Brein v Ziggo BV*, the Court of Justice of the European Union ('CJEU') held that the operators of the Pirate Bay, a P2P file-sharing website, were liable for an act of communication to the public of copyright-infringing material.[51] Due to the limited applicability of copyright protection to victims of malicious deepfakes, though, this cause of action is unlikely to be of widespread use in the deepfake context.

But even when primary or secondary liability can be established, lawsuits against platforms tend to have poor prospects of success. The caveat is that unlike other types of defendants, **platforms frequently enjoy an extra layer of statutory immunities limiting their exposure to liability in many jurisdictions**. Two main varieties of immunity can be distinguished .

2.1.1. Broad Immunity

The classic example of broad immunity is Section 230(c) of the US federal Communications and Decency Act 1996. Section 230 was enacted at a time when minimal government regulation was seen as a contributing factor to the success of the developing internet. This philosophy underpins the twin immunities conferred by subsection (c).

Section 230(c)(1) ensures that providers of any 'interactive computer service' are deemed not to be the speaker or author of any third-party content.[52] The immunity provided in this respect can be viewed as absolute and unconditional, as there are no exceptions outlined by the legislation. Moreover, courts have interpreted the provision broadly to effectively exempt platforms from a wide range of liabilities.[53]

Whereas the first paragraph of Section 230(c) shields platforms when they *make available* user-provided content, the second paragraph protects platforms from incurring liability when they take down or otherwise *restrict the availability* of such material. Section 230(c)(2) provides that an interactive computer service provider will not be liable for actions taken in 'good faith' to restrict access to material they determine to be 'obscene,

[51] Case C-610/15 *Stichting Brein v Ziggo BV* ECLI:EU:C:2017:456.

[52] 47 U.S.C. § 230(c)(1).

[53] US federal courts have upheld the Section 230(c)(1) defence for platforms in *Blumenthal v. Drudge*, 992 F. Supp 44 (D.D.C. 1998) (defamation); *Doe v. MySpace, Inc.,* 528 F.3d 413 (5th Cir. 2008) (fraud and negligence); *Tiffany (NJ) Inc. v. eBay Inc.* 600 F.3d 93 (2nd Cir. 2010) (trade mark infringement).

lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable'.[54] The section is sometimes referred to as the 'Good Samaritan' provision.

Platform defendants in the US frequently invoke Section 230(c)(1) to plead unqualified immunity. In contrast, few rely on the second limb's Good Samaritan protection, likely because the 'good faith' requirement would require a great deal more evidence and expense to prove in court.[55]

A version of broad immunity is also conferred by Section 5 of the UK Defamation Act 2013, which provides platform operators with a defence from liability if they can show the content was posted by a third party. The defence will only be defeated if the poster cannot be identified by the claimant, the claimant submits a notice of complaint to the platform, and the platform fails to respond in accordance with the relevant Regulations.[56] However, if the poster objects to the content's removal and refuses to disclose contact details to the platform, s.5 appears to shield the platform unconditionally from liability as a publisher or authoriser.

With respect to deepfakes, a broad immunity regime could theoretically lead to under-blocking as platforms lack any legal incentives to moderate content. Granted, in practice, there may be market incentives favouring moderation, such as a desire to build long-term confidence and trust in relationships with other businesses and consumers.[57] Even so, a broad immunity regime fails to provide guidance to platforms that do wish to take down content for any reason, commercial or otherwise. Thus, unless it is supplemented by additional principles enshrined in hard law or soft law instruments, broad immunity risks fostering inconsistent and divergent approaches by different platforms towards deepfake content. Legal certainty for platforms is achieved at the expense of considerable uncertainty for victims of malicious deepfakes.

[54] § 230(c)(2).

[55] Jess Miers, 'SCL Webinar: An Overview of Section 230 and Content Moderation - Online Intermediary Liability in the US' (Society for Computers and Law, 6 August 2020).

[56] Defamation (Operators of Websites) Regulations 2013, SI 2013/3028.

[57] Eric Goldman, 'An Overview of the United States' Section 230 Internet Immunity' (*Oxford Handbook of Online Intermediary Liability*, 4 May 2020) 164.

2.1.2. Qualified Immunity

On the other hand, a qualified immunity approach is taken by the EU E-Commerce Directive.[58] The Directive offers safe harbours from liability to 'information society service providers' depending on their level of activity when processing information: acting as mere conduits (Art.12), caching (Art.13), and storing information (Art.14). Online platforms such as Facebook, Twitter, YouTube and Medium fall within the Art.14 category as they provide hosting services for user-generated content.

To be exempt from liability, two conditions must be met. First, the platform's activities must be of a 'mere technical, automatic and passive nature, which implies that the [platform] has neither knowledge of nor control over the information which is transmitted or stored'.[59] In *L'Oreal v eBay*, concerning eBay's liability for trade mark infringement by third-party sellers on its online marketplace, the CJEU emphasised that merely setting the terms of service and providing general information to sellers would not disapply the exemption.[60] On the other hand, where the operator had provided assistance by optimising presentation of sales or promoting those offers, it would have taken an active role bringing it beyond the scope of the Art.14 safe harbour. Thus, platforms in the EU would not be exempt from liability for deepfakes if they had actively promoted the visibility of that content.

The second condition for the exemption is a lack of actual or constructive knowledge of illegal activity or information. Once a platform becomes aware of such illegality, though, it must act expeditiously to remove or disable access to the content.[61] A failure to do so may expose the platform to primary liability for damages.

Arguably, conferring statutory immunity based on a platform's lack of knowledge creates a perverse incentive for platforms to take a hands-off approach to moderating deepfakes

[58] Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('E-Commerce Directive') OJ L 178.
[59] Rec.42.
[60] *Case C-324/09 L'Oréal SA and Others v eBay International AG and Others,* EU:C:2011:474 (12 July 2011).
[61] Typically, knowledge is generated in the form of a Notice-and-Action mechanism, explained in Chapter 3.

and other content, as ignorance allows them to escape liability.[62] One suggestion to remedy this situation is to introduce a Good Samaritan protection for content moderation, akin to Section 230(c)(2) CDA.[63] However, while a Good Samaritan clause might remove disincentives, there is no guarantee it would positively incentivise greater moderation of deepfake content.

Furthermore, the Directive only requires the removal of *illegal* content. Where deepfake content is blatantly illegal, then platforms have an obligation to remove it once notified. A grey area exists with regard to deepfakes that are potentially harmful but not expressly made illegal under national law. Such content would not be covered by any specific legal obligation but could possibly be subject to private ordering of platforms in accordance with their own terms of service.

2.2. The New View: Platforms Have Responsibilities as Information Gatekeepers

Evidently, the immunity regimes discussed above do not provide incentives that would encourage platforms to respond to deepfakes unless they are manifestly illegal. Indeed, qualified immunities conditioned on knowledge could even disincentivise content moderation. Nevertheless, there is a growing appetite among legislators and commentators for platforms to play a central role in combating the spread of malicious deepfakes, as well as other types of disinformation, misinformation, and illegal content. Broadly speaking, the discourse is shifting away from the traditional approach of immunising platforms from liability, and towards a new paradigm where platforms are expected to assume greater responsibility for the user-generated content they host, including deepfakes.[64]

Note, however, that 'platform responsibility' takes on a different meaning depending on the context of its use. The term is invoked by scholars and policymakers in two distinct and interrelated ways:

[62] Miriam C Buiten, Alexandre de Streel and Martin Peitz, 'Rethinking Liability Rules for Online Hosting Platforms' (CRC TR 224 2019) Discussion Paper No. 074 16 <https://www.ssrn.com/abstract=3350693> accessed 26 July 2020.

[63] Buiten, de Streel and Peitz (n 62).

[64] Giancarlo F Frosio, 'Reforming Intermediary Liability in the Platform Economy: A European Digital Single Market Strategy' (2017) 112 Northwestern University Law Review 19.

- In the first sense, a responsibility to **vigilantly monitor and address illegal content**; and
- In the second sense, a responsibility to **respect human rights, including individuals' right to freedom of expression, right to privacy** and, arguably, their **right to due process**.

Below, I examine how both notions of platform responsibility bear upon the question of how platforms ought to address deepfakes.

2.2.1. A Responsibility to Monitor and Address Illegality

The first notion of platform responsibility arises from a perception that platforms are '**under-blocking**', i.e. not doing enough to halt the dissemination of illegal content. That sentiment underpins **a trend of recent legal developments aiming to compel platforms to monitor user-generated content more closely than ever before**. For instance, Article 17 of the EU's Copyright Directive disapplies the E-Commerce Directive's safe harbours and imposes a standard of liability which asks whether an online content-sharing service provider has made 'best efforts' to obtain authorisation before making available to the public a copyright work, and has acted in accordance with 'high industry standards of professional diligence'.[65] Providers must not only act expeditiously to remove copyright infringing material but also make best efforts to prevent future uploads of such content.[66] Commentators have pointed out that this last requirement would in practice require platforms to implement automated filtering systems in order to identify re-uploads of the illegal material.[67] The requirement therefore appears to contradict Article 15 of the E-Commerce Directive, which prohibits EU member states from imposing general obligations on platforms to monitor content. In any case, this development clearly signals that platform ignorance will no longer function as an exemption to copyright infringement liability in the EU.

Recent EU jurisprudence mirrors the legislative shift towards an elevated responsibility for platforms. In *Glawischnig-Piesczek*, an Austrian politician sued Facebook for failing

[65] Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC (Text with EEA relevance.) OJ L 130/92 ('Copyright Directive').
[66] Ibid.
[67] Giancarlo Frosio and Sunimal Mendis, 'Monitoring and Filtering: European Reform or Global Trend?' (*Oxford Handbook of Online Intermediary Liability*, 4 May 2020).

to remove content that was found to be defamatory by a national court. The CJEU ruled that even where a hosting service provider such as Facebook was exempted from liability via the Art.14 safe harbour, it could nonetheless be subject to a stay-down injunction under national law that would require it to remove or block access to information whose content was identical or equivalent to the illegal material, 'irrespective of who requested the storage of that information'.[68]

At first sight, the practical implications of *Glawischnig-Pieszcek* seem incompatible with the Art.15 prohibition. However, according to the CJEU, Art.15 precludes *general* monitoring obligations, but in principle permits an order in a *specific* case to prevent future uploads of illegal material. In effect, *Glawischnig-Piesczek* establishes that platforms can be required under national law to proactively screen out certain illegal content. Remarkably, the Court suggested that platforms need not carry out an independent assessment of the legality of such content 'since the [platform] has recourse to automated search tools and technologies'.[69] Apparently, the CJEU expects platforms to be able to operationalise this type of monitoring obligation through the use of such technologies as a matter of course.

Over in the US, lawmakers have also begun to make inroads into Section 230 CDA's blanket immunity via the 2018 passage of the Stop Enabling Sex Traffickers Act (SESTA) and Allow States and Victims to Fight Online Sex Trafficking Act (FOSTA) (collectively, 'FOSTA-SESTA').[70] FOSTA-SESTA created exclusions to Section 230 not only with respect to criminal prosecution of federal sex trafficking crimes, but also for civil causes of action based on behaviour that violates the aforementioned federal criminal law.[71] In other words, platform operators in the US can now be held liable for user-generated content they publish if it is covered by laws prohibiting sex trafficking and prostitution. FOSTA-SESTA represents a stark change from the philosophy behind Section 230, and its passage has emboldened activists calling for similar regulation concerning other areas of content moderation.[72]

[68] Case C-18/18 *Eva Glawischnig-Piesczek v Facebook Ireland Limited* Judgment (OJ), 22/11/2019, para 37.
[69] Ibid, para 46.
[70] 'Allow States and Victims to Fight Online Sex Trafficking Act of 2017', H.R.1865, 115th Cong. (2017).
[71] Eric Goldman, 'The Complicated Story of FOSTA and Section 230' (2019) 17 First Amendment Law Review 279.
[72] Emily Stewart, 'The next Big Battle over Internet Freedom Is Here' (*Vox*, 23 April 2018) <https://www.vox.com/policy-and-politics/2018/4/23/17237640/fosta-sesta-section-230-internet-freedom> accessed 25 August 2020.

The trend described above is not unique to the EU and US. In the *Baidu* case in China, hosting providers were ordered to proactively monitor for infringements of content that had been viewed and downloaded over a certain number of times.[73] In Brazil, the Superior Court of Justice ordered Google to carry out removal of a copyright-infringing video as well as similar videos uploaded by other users under different titles.[74] More recently, the UK government released its Online Harms White Paper addressing a swathe of problematic content from child abuse to revenge pornography, and proposing inter alia to impose a 'duty of care' on online platforms that would be enforced by a regulator.[75]

The economic justification cited in favour of many of these developments is that platforms are the lowest cost avoiders of harm caused by online dissemination of content.[76] With regard to deepfakes, the discussion in Chapter 1 demonstrates that this argument does hold true, as rapid removal or restriction of harmful deepfakes at their source is only possible via platform engagement far sooner than an individual would be able to obtain redress in court. Platforms themselves possess more resources and are better placed to carry out such operations at scale than the individual claimants in each deepfake case. Thus, insofar as deepfakes are found to be illegal on any basis, platforms may increasingly find themselves saddled with obligations to monitor and remove such content on the basis of court orders or legislation.

Legal reasons aside, **there is also a strong business incentive for companies to win public trust by being seen to take action against disinformation and misinformation**. Short-lived apps such as YikYak provide a cautionary tale: platforms that take an utterly laissez-faire approach, allowing any kind of user-generated content to remain unmoderated and unfiltered, have not proven sustainable or attractive business ventures in the long term.[77] Thus, platforms may have a good commercial reason to adopt a more hands-on approach towards deepfake content.

[73] *Zhong Qin Wen v Baidu* [2014] Gao Min Zhong Zi no. 2045 (Ch.)
[74] *Superior Court of Justice Fourth Panel Google Brazil v Dafra* [24 March 2014] Special Appeal no. 1306157/SP (Bra.).
[75] 'Online Harms White Paper' (*GOV.UK*) <https://www.gov.uk/government/consultations/online-harms-white-paper/online-harms-white-paper> accessed 23 August 2020.
[76] Garry A Gabison and Miriam C Buiten, 'Platform Liability in Copyright Enforcement' (2020) 21 Columbia Science & Technology Law Review 237, 250.
[77] Goldman (n 57) 164.

2.2.2. A Responsibility to Respect Human Rights

Even as platforms come under increasing pressure to 'do more' from various quarters, they also face criticism for doing *too much* in certain situations. The same automated filtering systems endorsed by the CJEU in *Glawischnig-Piesczek* have also been lambasted by commentators for indiscriminately taking down content that was neither illegal nor in violation of platforms' terms of service.

A key example of this phenomenon, known as **over-blocking**, arose during the early stages of the COVID-19 pandemic. In March 2020, Facebook, Twitter and YouTube sent home their human content moderators and increased their reliance on automated technologies to detect and remove content in violation of their policies. At the same time, these companies warned of the possibility that the algorithms would identify more 'false positives' resulting in erroneous content take-downs.[78] Sure enough, the New York Times reported on an instance when Facebook's automated systems wrongly flagged posts by volunteers sewing handmade masks for frontline workers, issuing warnings and even threatening to ban the groups and accounts in question.[79]

In response to criticisms that they are blocking free speech, privately-owned platforms invoke their own rights to expression which in principle allow them to freely set and enforce the terms and conditions governing the content they do and don't allow. Because platforms are private parties and not government bodies, by definition they do not engage in censorship and therefore cannot violate individuals' rights to freedom of expression. In the US, this type of argument forms the basis of platforms' 'First Amendment defence' against liability.[80]

In reality, however, **many platforms face public and regulatory pressure to self-regulate in accordance with soft law instruments.**[81] In 2017, the EU Commission

[78] 'Social Media Giants Warn of AI Moderation Errors as Coronavirus Empties Offices' *Reuters* (18 March 2020) <https://www.reuters.com/article/us-health-coronavirus-google-idUSKBN2133BM> accessed 23 August 2020.

[79] Mike Isaac, 'Facebook Hampers Do-It-Yourself Mask Efforts' *The New York Times* (5 April 2020) <https://www.nytimes.com/2020/04/05/technology/coronavirus-facebook-masks.html> accessed 23 August 2020.

[80] The US Supreme Court recently affirmed that 'merely hosting speech by others is not a traditional, exclusive public function and does not alone transform private entities into state actors subject to First Amendment constraints': *Manhattan Community Access Corp. v. Halleck*, 139 S. Ct. 1921 (2019).

[81] Soft law includes 'soft rules that are included in treaties, nonbinding or voluntary resolutions, recommendations, codes of conduct, and standards. See: 'Soft Law' (*obo*)

released its Communication on 'Tackling Illegal Content Online: Towards an Enhanced Responsibility of Online Platforms'. The Communication outlined a set of recommendations for platforms to effectively remove illegal content, improve transparency in content policies and notice-and-action procedures, and provide safeguards for fundamental rights by instituting counter-notice procedures.

Additionally, and of relevance to deepfakes, the EU Commission published a voluntary Code of Practice on Disinformation that requires its signatories to commit to five pillars of action:

1) de-monetise purveyors of disinformation
2) close fake accounts
3) invest in technological means to dilute visibility of fake info by promoting findability of trustworthy accounts
4) educate consumers and raise digital literacy, and
5) provide support to the research community.[82]

All measures to curb disinformation must be taken 'within the legal framework provided by the Charter of Fundamental Rights of the European Union (CFREU) and the European Convention on Human Rights (ECHR)'. The Code makes specific mention of the right to freedom of expression enshrined in Art.11 CFREU and Art.10 ECHR, which is described therein as 'an indispensable enabler of sound decision-making in free and democratic societies'.[83]

To date, the Code has been signed by Microsoft, Mozilla, Facebook, Google, Twitter, and numerous advertising companies. Each signatory provided a roadmap and submitted a report describing the steps it took toward compliance. Despite the apparent progress, the EU has hinted it will consider further legislative or regulatory action if self-regulation does not work as well as hoped.[84]

<https://www.oxfordbibliographies.com/view/document/obo-9780199796953/obo-9780199796953-0040.xml> accessed 25 August 2020.

[82] EU Commission, 'Code of Practice on Disinformation' (*Shaping Europe's digital future - European Commission*, 26 September 2018) <https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation> accessed 24 June 2020.

[83] ibid.

[84] Tarlach McGonagle, 'Free Expression and Internet Intermediaries: The Changing Geometry of European Regulation' (*Oxford Handbook of Online Intermediary Liability*, 4 May 2020) 484.

While self-regulation by definition relies on the cooperation of platforms, it has also been suggested that **platforms can become subject to human rights obligations as a consequence of their engagement in activities of democratic significance**. Online platforms can be viewed as gatekeepers controlling the flow of information, according to Emily Laidlaw. A subset of these gatekeepers, whom Laidlaw terms 'Internet Information Gatekeepers' (IIGs), have a significant impact on democratic discourse—for instance, by controlling participation in democratic culture. Laidlaw argues that IIGs incur human rights responsibilities by virtue of their de facto functions, even without voluntarily assuming those obligations**.** As she notes, the framework for IIG responsibility bears resemblance to the way corporate social responsibility arises based on a company's sphere of influence as articulated in the United Nations Global Compact, an important instrument of soft law.[85]

IIGs exist along a spectrum such that the degree of human rights obligations they attract depends on the importance of their functions to a democratic society. Laidlaw classifies them into three categories:

- **Macro-gatekeepers** are essential or inevitable to facilitate democratic discourse, serving as 'choke points' of information flows. The essentiality of their function would therefore attract the strongest human rights obligations. Examples: ISPs, search engines.
- **Authority gatekeepers** are a level below, in that their impact on democratic discourse may be central but not inevitable. Examples: Facebook, Twitter.
- **Micro-gatekeepers** have the power to moderate content and control information flows which are of 'democratic significance', but on a relatively small scale such that their impact is more limited. Example: individual news websites like Huffington Post.[86]

Based on this classification, the most likely candidates to address the dissemination of deepfakes are authority gatekeepers like Facebook, YouTube and Twitter—platforms whose considerable reach gives them a central if not essential role in democratic discourse. Gatekeeping theory indicates that these platforms should be subject to an intermediate

[85] Emily B Laidlaw, *Regulating Speech in Cyberspace: Gatekeepers, Human Rights and Corporate Responsibility* (Cambridge University Press 2015) 47.
[86] Laidlaw (n 85).

degree of obligations, inviting more scrutiny and accountability than micro-gatekeepers but allowing for greater latitude than we would give to macro-gatekeepers.

An important question remains: which human rights are at stake? In the online gatekeeping context, Laidlaw suggests that IIGs have **responsibilities to protect individuals' rights to freedom of expression, privacy, and association**.[87] The relevance of these rights is obvious: content moderation has been described, imperfectly yet insightfully, as 'private censorship'.[88] In the everyday world of online content moderation, platforms act as the all-in-one law enforcers, adjudicators, and administrative agencies.[89] When authority gatekeepers decide to restrict certain content or to remove a user's account, they are de facto making an impact on democratic discourse by curtailing a person's expression, regardless of their justification or legal defences against liability.

In addition, I submit that respect for **the right to due process** also belongs on the list of gatekeeper responsibilities. Content moderation decisions, whether made by humans or algorithms, are far from infallible, as illustrated earlier by Facebook's erroneous removals of volunteer posts. For this very reason, Facebook is developing its own 'supreme court', an oversight board designed to handle appeals from users contesting wrongful removal of content as well as users requesting the removal of content.[90]

Due process is enshrined in Principle 5 of the Manila Principles on Intermediary Liability, a set of best practice standards for the regulation of online content that were developed by an international coalition of NGOs. Inter alia, the principle states that platforms should supply 'user content providers with mechanisms to review decisions to restrict content in violation of the intermediary's content restriction policies'.[91] Furthermore, platforms must reinstate content upon successful appeal by the user or as directed by a court order following judicial review.[92]

[87] ibid 48.

[88] Kyle Langvardt, 'Regulating Online Content Moderation' (2018) 106 Georgetown Law Journal 1353.

[89] Niva Elkin-Koren and Maayan Perel, 'Guarding the Guardians: Content Moderation by Online Intermediaries and the Rule of Law' (*Oxford Handbook of Online Intermediary Liability*, 4 May 2020) 674.

[90] Jane Wakefield Cellan-Jones Rory, 'Facebook's "supreme Court" Members Announced' *BBC News* (6 May 2020) <https://www.bbc.com/news/technology-52558559> accessed 23 August 2020.

[91] The text of Principle 5 is reproduced in **Appendix B.** The full Manila Principles can be found at 'Manila Principles' <https://www.manilaprinciples.org/> accessed 12 July 2020.

[92] ibid.

Going a step further, Frederick Mostert has argued for the development of 'digital due process' principles to underpin a new legal infrastructure for the online world, through which the rule of law and fundamental rights can be upheld.[93] Given the 'virality, volume and velocity of online speech', as Mostert aptly observes, content moderation issues of under-blocking and over-blocking are not only probable but inevitable as platforms attempt to respond to the dual pressures of platform responsibility. Viewed thus, respect for due process is necessarily a pre-requisite to facilitate protection of users' rights to freedom of expression, privacy, and association.

2.3. Recommendations to Improve the Legal Framework in Response to Deepfakes

Under the emerging framework of platform responsibility, platforms will face growing pressure to oversee and respond to user-generated deepfake content. But reconciling the two facets of responsibility is no easy task. As platforms scale up their moderation efforts at the urging of courts or policymakers, they simultaneously risk blocking too little harmful content and erroneously taking down content such that they restrict users' legitimate exercise of free speech. This phenomenon is known as the 'moderator's dilemma'.[94]

To address the moderator's dilemma in the context of deepfakes, we need a robust legal framework that enables and incentivises platforms to appropriately act upon deepfake content. I propose the following recommendations to facilitate platform action that will strike a balance between competing rights: on the one hand, protecting the individual or

[93] Mostert identifies the following principles as the cornerstones of digital due process:

1. a fair and public review by an independent and impartial panel or competent court within a reasonable time;
2. a proper prior notification of the review;
3. an opportunity for a user or notifier to respond and present evidence in respect of a takedown or a stay-up inaction by a platform;
4. the right to legal representation;
5. the right to appeal to an appeals panel, alternative dispute resolution tribunal or competent court;
6. notifiers may at any stage in the process seek access to competent courts;
7. the right to receive a decision which clearly articulates the reason for that decision; and
8. the right to an effective remedy including, for example, stay-up or takedown.

See Frederick Mostert, '"Digital Due Process": A Need for Online Justice' (2020) 15 Journal of Intellectual Property Law & Practice 378, 388.

[94] Goldman (n 57) 157.

society under threat from the deepfake content, and on the other, safeguarding the poster's and public's fundamental rights to free speech and due process.

First, legislators should enact or amend criminal laws to explicitly prohibit the most serious deepfake threats outlined in Chapter 1. US states like Virginia, California and Texas have already updated their laws to cover specific uses of deepfakes, from revenge pornography to misleading political statements prior to an election. Other jurisdictions should strongly consider taking similar legislative action to bring deepfakes within the scope of their law. A statutory offence would reduce the burden for deepfake claimants who might not definitively be covered by existing causes of action—for instance, those in jurisdictions that do not recognise image rights.

From the platform perspective, criminalising specific deepfake uses will give platforms a sound legal basis to update and enforce their policies on prohibited content. Platform liability frameworks set out in the E-Commerce Directive and Section 230 CDA already contain built-in exceptions or conditions for immunity that oblige platforms to restrict criminal content and collaborate with authorities on such matters. Statutory deepfake offences thus provide a firm scaffold on which platforms can build their own policies.

Also, creating specific deepfake offences will attenuate the criticisms that platforms are unilaterally and opaquely restricting user expression based on what they deem 'harmful' rather than outright 'illegal'. Although the illegality of content is not always clear-cut, what is 'harmful' remains even more subjective and lacking in specific guidance from legislation or case law.[95] Updating criminal legislation will ensure that platforms respond to illegal deepfake content because they are legally required to do so. Moreover, freedom of speech concerns will be alleviated in many jurisdictions because speech rights can be restricted in accordance with the law as is necessary in a democratic society, particularly to prevent crime.[96] The harms of deepfakes are already well recognised in the literature, but now is the time to expressly prohibit them as a matter of law.

[95] Tambiama Madiega, *Reform of the EU Liability Regime for Online Intermediaries: Background on the Forthcoming Digital Services Act : In-Depth Analysis.* (European Parliamentary Research Service 2020) 10 <https://op.europa.eu/publication/manifestation_identifier/PUB_QA0420239ENN> accessed 23 June 2020.

[96] Art.10(2) ECHR.

A word of caution, however: this Legal Opinion does **not** recommend imposing a blanket ban on all uses of deepfakes. Given that deepfakes can be used in both harmful and beneficial ways, it would be far more difficult to justify the interference with users' rights when deepfakes are clearly labelled as such by the user and not intended to cause harm. Benign deepfakes could even enhance consumers' digital literacy by making them aware of the technology's existence. In contrast, a total ban might inadvertently result in reduced awareness and greater vulnerability to truly deceptive uses of deepfakes.

Second, regulators should develop or modify soft law instruments to provide guidance to platforms on how to respond to deepfakes. Such instruments should clearly outline:

- the types of deepfakes that should, or should not, be restricted
- the applicable legal bases for restricting deepfakes
- examples of deepfake content moderation best practices[97]
- fundamental rights safeguards that platforms can implement to protect users' freedom of expression and privacy, and
- ways in which platforms can ensure a respect for due process through their policies and notice-and-action procedures.

Third, law enforcement authorities should develop a blocklist of websites that predominantly host illegal deepfake content. Assuming national criminal law provides a sufficient legal basis to do so, authorities could maintain a blocklist comprised of, inter alia, websites dedicated to deepfake pornography or online platforms that deliberately host deceptive deepfake 'news'.

The concept of an illegal website blocklist has already been successfully implemented by the City of London Police's IP Crimes Unit ('PIPCU') as part of their anti-counterfeiting initiative, Operation Ashiko. Upon notification by rightsholders or detection of websites engaging in IP crimes like trading in counterfeit goods, PIPCU works with Nominet—the UK's top-level domain name registry—to suspend the domains in question within 48 hours.[98] A similar tactic could be employed to combat illegal deepfakes.

[97] See Chapter 3 for a discussion of content moderation measures.

[98] Presentation by Detective Constable Weizmann Jacobs, City of London Police Intellectual Property Crime Unit (City of London Police, 16 October 2019).

Fourth, lawmakers and regulators should consult with stakeholders to determine how to provide appropriate incentives or disincentives for platforms to moderate deepfake content as desired. Certainly, market forces alone could incentivise platform intervention in respect of deepfakes, but consumers gain no certainty from relying solely on individual platform operators to implement their own policies and practices as they see fit.

But regulation presents its own challenges. When setting regulatory standards, consider which types of platforms would be regulated: small or big? Macro-, authority, or micro-gatekeepers? A one-size-fits-all approach may lack the necessary nuance to tackle the different challenges faced by well-resourced major platforms and smaller-scale website operators.

The type of legal measure matters, too. With hard law alone, it may be difficult to tailor incentives to all types of platforms. Soft law measures, on the other hand, may suffer from a lack of 'teeth' to ensure platform compliance. Consequently, policymakers and regulators may find it useful to implement a combination of both to achieve their aims.

Fifth, it is crucial to avoid the temptation to implement drastic changes to liability frameworks without first conducting further research into their potential impact. A knee-jerk removal of existing platform immunities in respect of deepfake content might do more harm than good. Fear of liability may lead platforms to act in risk-averse ways—for instance, by removing particular channels of discourse entirely rather than risking liability. In an illustration of this scenario, the passage of FOSTA-SESTA prompted Craigslist to remove its infamous 'Personal' ads section,[99] halting its use by sex trafficking rings but also forcing voluntary sex workers off the platform.[100] What if potential liability for, say, political deepfakes caused a platform like Twitter to make it impossible to share video clips of politicians at all? For smaller platforms in particular, the economically rational course of action is to simply eliminate their risk of exposure to lawsuits, even if that means shutting down forums for users to legitimately exercise their freedom of speech. Whatever their sentiments on platform immunity, legislators should bear in mind that changing entire platform liability structures can have ramifications beyond their original policy goals.

[99] Stewart (n 72).

[100] Goldman (n 71) 291.

3. Moderating Deepfake Content

At this juncture, it is clear that platforms' immunity from primary or secondary liability for deepfake-related claims will not necessarily insulate them from pressures to monitor and moderate deepfake content. In this third and final chapter, I explore the practical challenges that platforms will encounter when moderating deepfakes. The content policies of several major platforms are used as case studies to understand current industry practice regarding deepfakes, and to identify gaps to address.

Content moderation can be defined as 'the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse'.[101] Although moderation practices vary from platform to platform, the following taxonomy may be helpful as a background to the discussion:

- Approaches to moderation: **manual** (human moderators), **automated** (software), or a **combination** of both.
- Models of moderation: **centralised** (the platform itself hires and trains moderators, either internally or by outsourcing to a contractor) or **decentralised** (the task of moderation is delegated to users themselves).
- Stages of moderation: **ex ante** (screening prior to publication), **ex post proactive** (automated filtering of published material), **ex post reactive** (moderator action in response to user reports concerning published material).[102]

In the specialised context of deepfake moderation, the main challenges that arise are: (1) identification of deepfakes, (2) drafting and enforcement of content policies, (3) ensuring due process and preventing abuse, and (4) transparency and accountability.

3.1. Identifying Deepfakes

The first step to moderating deepfakes is being able to accurately identify them as such. At the time of writing, researchers have identified tell-tale signs that can help human viewers to recognise synthetic media. Possible indicators include blinking patterns,

[101] James Grimmelmann, 'The Virtues of Moderation' (2015) 17 Yale Journal of Law and Technology 47.

[102] Spandana Singh, 'Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content' (Open Technology Institute (New America) 2019) 8.

unusual shading, and unnatural facial hair.[103] Yet the sophistication of deepfakes is constantly advancing; in time, deepfakes may become so subtle that they cease to be detectable by the human eye. In fact, the goal of the machine learning framework used in deepfake generation is for the 'generative' neural network to learn to output deepfakes so convincing that the 'discriminative' network can no longer reliably discern which are real or fake.[104]

Multi-stakeholder efforts to **improve deepfake detection technologies** are currently underway. In 2019, the Deepfake Detection Challenge ('DFDC') was launched by a group of technology companies along with the research community as a way of crowdsourcing solutions.[105] The results of the competition demonstrate that technological solutions for detection still have a long way to go. Researchers used a public dataset to train their models and achieved up to 82.56% accuracy of detection. But when tested on a black-box dataset of previously unseen videos, the best-performing model had an accuracy rate of just 65.18%. Nevertheless, the DFDC provides an important proof-of-concept that machine learning tools can be used to successfully detect deepfakes at all, and the organisers are hopeful that further progress can be made to improve accuracy rates.[106]

Others argue that detection alone may never be sufficient to counter synthetic media, given the inevitable arms race between deepfake-generating technologies and the tools meant to detect them. An alternative and complementary approach focuses on **improving the standard of content attribution**, in order to increase trust and transparency in 'real' media. In a recent white paper, researchers at Adobe propose a new industry standard for content attribution to be developed in collaboration with media organisations and other stakeholders. The central idea is to ensure that attribution metadata created at the point of capture remains intact, tamper-proof and accessible no matter the mode of transmission. Adobe's model relies on a trusted authority to carry out

[103] 'Project Overview ‹ Detect DeepFakes: How to Counteract Misinformation Created by AI' (*MIT Media Lab*) <https://www.media.mit.edu/projects/detect-fakes/overview/> accessed 27 August 2020.

[104] For an accessible explanation of Generative Adversarial Networks, see Joey Mach, 'Deepfakes: The Ugly, and The Good' (*Medium*, 2 December 2019) <https://towardsdatascience.com/deepfakes-the-ugly-and-the-good-49115643d8dd> accessed 27 August 2020.

[105] 'Facebook, Microsoft, and Others Launch Deepfake Detection Challenge' (*VentureBeat*, 11 December 2019) <https://venturebeat.com/2019/12/11/facebook-microsoft-and-others-launch-deepfake-detection-challenge/> accessed 27 August 2020.

[106] Devin Coldewey, 'Facebook's "Deepfake Detection Challenge" Yields Promising Early Results' (*TechCrunch*) <https://social.techcrunch.com/2020/06/12/facebooks-deepfake-detection-challenge-yields-promising-early-results/> accessed 27 August 2020.

verification and time-stamping, with the possibility of using Distributed Ledger Technology as an additional tool for secure information storage.[107] If successfully implemented, the Content Authenticity Initiative could provide consumers with greater assurance that they can trust in the authenticity of verified video.

Regardless of whether platforms rely on detection, attribution, or a combination of both approaches, **deepfake identification is a prerequisite to the sort of ex ante or ex post proactive monitoring envisaged in a *Glawischnig-Piesczek*-style stay-down injunction**. Once deepfakes are detected and verified as false, automated filtering technologies can subsequently be used to identify them and prevent further dissemination. For instance, digital hash technology is presently used to filter copyright-infringing content and child sexual abuse material (CSAM). A unique hash, or digital signature, is generated for known images and videos and stored in a database. Whenever a new piece of content is uploaded to a platform, its hash is added to the database and screened against previous entries to determine if there is a match.[108] Examples of content filtration using digital hashing include PhotoDNA for Video, a tool developed by Microsoft that is now widely used to tackle CSAM,[109] and audio/video fingerprinting by YouTube as part of its Content ID system.[110]

Deepfake identification will also require adjustments to platforms' ex post reactive moderation, i.e. Notice-and-Action (N&A) systems that rely on users to flag problematic content for further review. Platforms would first have to update their N&A procedures to allow users to notify the platform of potential deepfakes. Subsequently, deepfake detection technology and/or other methods of forensic analysis would need to be deployed to verify the video's authenticity or lack thereof.

[107] Leonard Rosenthal and others, 'Setting the Standard for Digital Content Attribution' (The Content Authenticity Initiative 2020) <https://documentcloud.adobe.com/link/track?uri=urn%3Aaaid%3Ascds%3AUS%3A2c6361d5-b8da-4aca-89bd-1ed66cd22d19#pageNum=1> accessed 27 August 2020.

[108] Singh (n 102) 12.

[109] 'How PhotoDNA for Video Is Being Used to Fight Online Child Exploitation | Microsoft On The Issues' (*On the Issues*, 12 September 2018) <https://news.microsoft.com/on-the-issues/2018/09/12/how-photodna-for-video-is-being-used-to-fight-online-child-exploitation/> accessed 27 August 2020.

[110] *YouTube Content ID* (2010) <https://www.youtube.com/watch?v=9g2U12SsRns> accessed 27 August 2020.

3.2. Content Policies

Once a deepfake can be accurately identified, the next challenge for a platform is to determine how best to respond to this information. When is a deepfake problematic enough to warrant action? Should the content be removed or restricted in other ways? What penalties, if any, should the user who posted the deepfake incur? Answering each of these difficult questions would require a wealth of empirical research and analysis beyond the scope of this Legal Opinion. However, I submit that the current industry practice with regard to deepfake moderation provides a useful starting point from which to evaluate, in theory, the strengths and shortcomings of various approaches.

A survey of four platforms' content policies reveals a smorgasbord of ideas on how to moderate deepfakes. I analyse the salient features of each one before concluding with some general recommendations.

3.2.1. YouTube

When content is found to be contrary to YouTube's Content Guidelines, it will be subject to removal and the user who posted the content will receive a warning. Multiple violations count towards YouTube's three-strikes system leading to eventual termination of a user's channel.[111]

Two sections of the Content Guidelines are relevant to deepfakes: the ban on impersonation, and the ban on manipulated media under its policy against spam and deceptive practices.

First, YouTube bans content 'intended to impersonate a person or channel'.[112] The ban on channel impersonation is primarily designed to protect the rights of trade mark holders from content which may cause confusion about the source of goods advertised, or the intentional copying of the overall look and feel of another channel. More relevant to deepfakes is the concept of 'personal impersonation', which YouTube defines as '[c]ontent

[111] 'Community Guidelines Strike Basics - YouTube Help' <https://support.google.com/youtube/answer/2802032?hl=en-GB> accessed 28 August 2020.

[112] 'Policy on Impersonation - YouTube Help' <https://support.google.com/youtube/answer/2801947?hl=en&ref_topic=9282365> accessed 17 July 2020.

intended to look like someone else is posting it'.[113] This definition may apply to a situation in which a user posts a deepfake video meant to deceive viewers into believing they are a different person. However, it is equally conceivable to imagine a scenario in which an anonymous user posts a deepfake video of a high-profile YouTuber—generated using the reams of publicly available video footage—that then goes viral because of the realistic content, rather than any intention by the poster themselves to impersonate the video's subject. YouTube's Policy on Impersonation would not necessarily capture this latter deepfake scenario.

Additionally, YouTube updated its policy on 'Spam, deceptive practices & scams' to include a prohibition against 'Manipulated Media' which it defines as:

> "Content that has been technically manipulated or doctored in a way that misleads users (beyond clips taken out of context) and may pose a serious risk of egregious harm."[114]

The phrase 'technically manipulated or doctored' is broad enough to cover deepfakes as a core instance, but to make this point even clearer, YouTube could consider explicitly including a deepfake video as an example of manipulated media in its policy.

As for what constitutes 'misleading' material, the term encompasses videos that present a factually inaccurate version of reality, e.g. by feigning the death of a government official or fabricating events. Deepfakes would clearly fall under this description.

Arguably, the most problematic element of YouTube's prohibition is the concept of a 'serious risk of egregious harm'. The notion of 'harm' in the legal context has often proven as slippery to define as 'reasonable person' or 'unfair competition' and ultimately, as J. Thomas McCarthy has observed, it is unproductive to try and define these terms in the abstract.[115] Instead, the law relies on a corpus of case law containing examples—real and hypothetical—to illustrate these otherwise nebulous concepts. In a similar vein, YouTube should provide examples of when and how it determines that a video poses a 'serious risk of egregious harm'. Currently, YouTube's examples of manipulated media simply restate

[113] ibid.
[114] 'Spam, Deceptive Practices & Scams Policies - YouTube Help' <https://support.google.com/youtube/answer/2801973?hl=en> accessed 17 July 2020.
[115] *McCarthy on Trademarks and Unfair Competition* § 1:8 (5th ed.)

the entire phrase (e.g. 'Inaccurately translated video subtitles that inflame geopolitical tensions creating *serious risk of egregious harm*') without elaborating on its meaning. This circular definition cannot provide the necessary guidance for moderators to enforce the Content Guidelines, nor for users to appeal wrongful moderation decisions. YouTube's policy would benefit from an explanation of what types of harm would warrant action; for a start, they might consider including the deepfake threats described in Chapter 1.

3.2.2. Facebook

Facebook's Community Standards prohibit manipulated media which meet the following criteria:

> "Video that has been edited or synthesized, beyond adjustments for clarity or quality, in ways that are not apparent to an average person, and would likely mislead an average person to believe that a subject of the video said words that they did not say
> AND
> is the product of artificial intelligence or machine learning, including deep learning techniques (e.g., a technical deepfake), that merges, combines, replaces, and/or superimposes content onto a video, creating a video that appears authentic."[116]

Exceptions to the prohibition include satire and parody, and also videos which have been edited to omit words that were said or in which the words have been reordered.[117]

The most striking feature of Facebook's policy is the narrowness of its criteria for applying the prohibition. Indeed, while the specificity of the language might be thought to provide greater certainty for moderators and users, it results in a ban too narrow to protect against the full spectrum of deepfakes. Case in point: Facebook's ban applies exclusively to video, whereas deepfakes exist in both video and audio form.[118]

Unlike other platforms, Facebook makes no mention of a criterion of 'harm' in its policy, allowing it to sidestep the thorny definitional issues noted previously. Instead, to trigger the prohibition, the video must mislead the average person to believe the subject said

[116] 'Community Standards' (*Facebook*) <https://www.facebook.com/communitystandards/manipulated_media/> accessed 24 June 2020.
[117] ibid.
[118] 'The Best, Most Impressive Audio Deepfakes on the Web | Digital Trends' <https://www.digitaltrends.com/news/best-audio-deepfakes-web/> accessed 28 August 2020.

words they did not. In short, application of the policy hinges upon a predominantly factual assessment of whether something did or did not happen.

Again, this deception requirement is underinclusive. Deepfakes can deceive viewers by misrepresenting characteristics of a person’s speech other than the words themselves. On its face, the policy will not cover a deepfake which re-enacts a person speaking words they actually said, albeit changing the emotion and tone of the speech to fabricate the context. Additionally, deepfake misrepresentations need not be verbal at all. Consider the example of non-consensual deepfake pornography which falsely depicts a person to have acted in a way that they did not—no speech is necessary to mislead the average viewer.

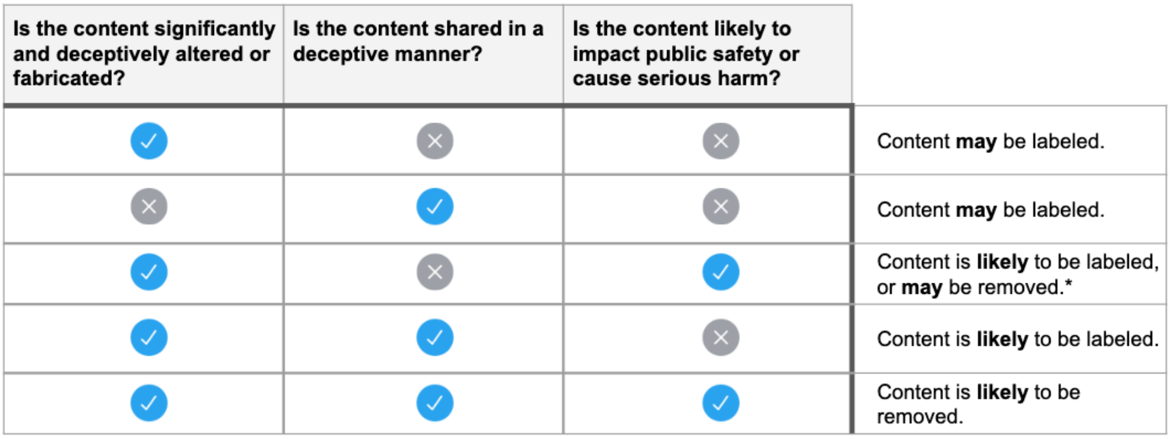
These examples illustrate the shortcomings of an overly specific policy. While specificity might make it simpler for Facebook's content moderators to determine what is and is not in violation, the policy’s formalistic approach risks leaving users without remedy if they have suffered harm from misleading deepfake audio or content that lies beyond the limited remit of ‘manipulated media’ as defined by Facebook.

3.2.3. Twitter

In February 2020, Twitter introduced its ‘Synthetic and manipulated media policy’ to combat the spread of misinformation and disinformation, including deepfakes. Twitter’s policy covers media that meets one or more of the following criteria: 1) significantly and deceptively altered or fabricated; 2) shared in a deceptive manner; and/or 3) likely to impact public safety or cause serious harm.[119]

A unique feature of Twitter’s policy is that the platform adjusts its response depending on which combination of the three criteria is exhibited by the content. As Table 1 illustrates, the severity of action will correspond with the risk of harm that the deepfake poses.

[119] ‘Synthetic and Manipulated Media Policy’ (*Twitter*) <https://help.twitter.com/en/rules-and-policies/manipulated-media> accessed 10 March 2020.



Is the content significantly and deceptively altered or fabricated?	Is the content shared in a deceptive manner?	Is the content likely to impact public safety or cause serious harm?	
✓	✗	✗	Content **may** be labeled.
✗	✓	✗	Content **may** be labeled.
✓	✗	✓	Content is **likely** to be labeled, or **may** be removed.*
✓	✓	✗	Content is **likely** to be labeled.
✓	✓	✓	Content is **likely** to be removed.

Table 1: Criteria used by Twitter to determine whether Tweets or media should be subject to labelling or removal pursuant to its 'Synthetic and manipulated media policy'.[120]

On one end of the scale, **labelling** is the response of choice where the content is *either* significantly and deceptive altered or fabricated, or shared in a deceptive manner based on the context of the media, [121] but not both. The policy emphasises through the word '*may*' that labelling will be carried out on a discretionary case-by-case basis, rather than a blanket application.

Where content is both synthetic or manipulated content and is likely to mislead, but not likely to cause harm, Twitter's policy is that such content is *likely* to be labelled. The language of likelihood here signals a recognition of gradated severity from the previous cases.

The most severe response, **removal**, is triggered only where two or more criteria are met, of which one must be a likelihood of impact on public safety or serious harm. Thus, media meeting criteria 1 and 3 are likely to be labelled and may be subject to removal. Where all three criteria are met, the content in question is 'likely to be removed'.

One advantage of Twitter's multi-step assessment is that it responds flexibly to all kinds of deepfakes. First, the scope of the policy applies to a broad range of synthetic media (criterion 1: 'significantly and deceptively altered or fabricated') and thus capable of capturing a range of deepfake content. Second, response severity is tailored based on a

[120] ibid.
[121] Relevant factors include 'text of the Tweet accompanying or within the media; metadata associated with the media; information on the profile of the account sharing media; websites linked in the Tweet, or in the profile of the account sharing media': ibid.

contextual assessment of deceptiveness (criterion 2) and harm or impact on public safety (criterion 3). By prima facie casting a wide net and then determining whether the content should be labelled, removed, or simply left alone, Twitter aims to achieve a degree of proportionality in its interference with user expression. This is a laudable step towards the operationalisation of platform responsibility.

Another strength of Twitter's policy lies in its interaction with other areas of content moderation. Where synthetic media overlaps with Twitter's prohibition on non-consensual nudity (which mandates removal), Twitter can take action on either basis. Moreover, Twitter states that where synthetic media is involved, they will err on the side of removal in borderline cases that might not be subject to moderation under other Twitter rules.

However, Twitter will need to contend with significant practical challenges to implement its policy. For one thing, the complexity of the rules will require considerable training for moderators and may be difficult to scale. It may also create uncertainty in the N&A procedure: will takedown notices issued to users provide a comparable level of detail to the current policy? How, if at all, can a user challenge a moderator's assessment and subsequent decision to impose a particular type of sanction? The success of Twitter's policy against deepfakes will depend in large part on how it tackles these issues.

Moreover, while labelling of content may be a lesser restriction on expression than outright removal, the task of designing an appropriate label is far from straightforward. In May 2020, Twitter's decision to label one of US President Trump's tweets as 'potentially misleading' stirred up controversy and led to the president's signing of an executive order targeting platforms' immunity for editorial decisions.[122] The incident highlights how 'every label — whether for text, images, video, or a combination — comes with a set of assumptions that must be independently tested against clear goals and transparently communicated to users', as researchers from the Partnership on AI note.[123] Thus, when designing labels for deepfakes and other forms of manipulated media, platforms should consider how to combat the influence of disinformation/misinformation without drawing

[122] 'Trump Signs Executive Order Targeting Twitter after Fact-Checking Row' *BBC News* (28 May 2020) <https://www.bbc.com/news/technology-52843986> accessed 21 June 2020.

[123] Emily Saltz and others, 'It Matters How Platforms Label Manipulated Media. Here Are 12 Principles Designers Should Follow.' [2020] *The Startup Magazine* <https://medium.com/swlh/it-matters-how-platforms-label-manipulated-media-here-are-12-principles-designers-should-follow-438b76546078> accessed 14 August 2020.

unnecessary attention to it, and also seek to educate users in a non-confrontational manner. Depending on the reason for labelling the deepfake, Twitter could also consider supplementing its response with other measures to reduce the content's visibility or shareability, such as downranking the content or adding an overlay.[124]

3.2.4. Reddit

Reddit takes a two-pronged approach to tackle deepfakes in its Account and Community Restrictions. In 2018, the platform banned involuntary pornography whether real or faked, including deepfakes as an instance of the latter.[125] The infamous r/deepfakes subreddit was shut down on the basis of this policy.[126]

In addition, Reddit prohibits the impersonation of an individual or entity in a misleading or deceptive manner, including by use of manipulated media. In contrast to YouTube, Reddit adopts a broader definition of impersonation that expressly includes 'deepfakes or other manipulated content presented to mislead, or falsely attributed to an individual or entity'. Notably, the platform exempts satire and parody from the ban on impersonation, ostensibly to provide refuge for legitimate creative uses of deepfakes but adds that they 'will always take into account the context of any particular content'. [127]

In a post accompanying the update of its policy against impersonation, a Reddit moderator stated that the ban 'doesn't apply to all deepfake or manipulated content-- just that which is actually misleading in a malicious way. Because believe you me, we like seeing Nic Cage in unexpected places just as much as you do.'[128]

However, no mention of 'malice' is made in the actual policy. This apparent inconsistency raises concern as to whether the spirit of the rule is made sufficiently clear by the policy language, which focuses only on whether the content is misleading or deceptive. Further

[124] ibid.

[125] 'Do Not Post Involuntary Pornography' (*Reddit Help*) <http://reddit.zendesk.com/hc/en-us/articles/360043513411> accessed 29 August 2020.

[126] Cole (n 13).

[127] 'Reddit Account and Community Restrictions - Do Not Impersonate an Individual or Entity' (*Reddit Help*) <https://www.reddithelp.com/en/categories/rules-reporting/account-and-community-restrictions/do-not-impersonate-individual-or> accessed 17 July 2020.

[128] 'R/Redditsecurity - Updates to Our Policy Around Impersonation' (*reddit*) <https://www.reddit.com/r/redditsecurity/comments/emd7yx/updates_to_our_policy_around_impersonation/> accessed 17 July 2020.

clarification and examples would be helpful to explain what kind of misleading material will be considered a violation and subjected to removal.

3.2.5. Recommendations for Deepfake Content Policies

- **Clearly explain the terms of the policy, using examples where necessary**. Terms like 'harm', 'synthetic media' and 'manipulated media' lack a consensus definition and are subject to different interpretations. To assist both users and moderators, platforms should elaborate on the meaning of these terms as used within their policy, and provide examples to illustrate how they might assess whether a piece of content meets each of the listed criteria for the policy to apply.
- **Evaluate the scope of the policy** against potential deepfake applications and consider whether the terminology is sufficiently broad to catch different instances of use.
- **Consider whether actions short of removal would be appropriate**. Labelling or reducing content visibility may constitute a more proportionate response than removal in some cases of deepfake moderation. These measures could also provide a better way to accommodate borderline uses of deepfakes or specific exceptions such as satire and parody. However, platforms should also bear in mind that such measures will also add a layer of complexity to the enforcement of the content policy and ensuing appeals.
- **Clarify how any exceptions to the policy will be assessed and applied**. Where applicable, platforms should explain the criteria for a deepfake to meet the satire and parody exemption (for instance, clear labelling by the user), and how users could rely on these bases to appeal a deepfake moderation decision.

3.3. Due Process and Abuse Prevention

As a deepfake takedown request is being processed, **platforms should adhere to principles of due process, in accordance with their obligations of platform responsibility**. Drawing from the Manila Principles as a guide,[129] at the minimum, platforms should:

[129] 'Manila Principles' (n 91).

- Provide users with a clear notice explaining the reasons for the moderation decision. Takedown notices should mention the reasons for the decision and bring attention to ways that the decision can be appealed.
- Provide an appeals procedure which is transparent, fair, financially accessible, and user-friendly. The appeals procedure should provide the user with an effective right to be heard within an appropriate timeframe.
- Avoid disclosing personally identifiable information about a user, unless ordered to do so by a judicial authority.
- Reinstate content that has been subject to wrongful takedown if a user successfully overturns the initial moderation decision by way of the platform's appeals procedure or by judicial review.

An additional challenge for any N&A system is the risk of abusive notification. In fact, a large volume of unsubstantiated deepfake reports could exacerbate the problems of deepfake moderation by increasing the incidence of wrongful takedowns and undermining consumer trust in platforms. **Thus, careful guidelines and safeguards would need to be developed alongside a deepfake N&A system in order to prevent abusive notification**. One possible countermeasure is the imposition of penalties for misrepresentation in a takedown request, but so far these have not been shown to effectively deter abuse of N&A in the copyright context.[130] When designing their N&A system to accommodate deepfakes, platforms should consider what kind of information would be necessary to substantiate a deepfake takedown request. Furthermore, platforms should seek to minimise the incentives and opportunities for complainants to benefit from frivolous or abusive notification—for instance, ensuring that complainants cannot use the system to extort financial benefits from the respondents targeted by a deepfake notice.[131]

[130] Aleksandra Kuczerawy, 'From "Notice and Takedown" to "Notice and Stay Down": Risks and Safeguards for Freedom of Expression' (*Oxford Handbook of Online Intermediary Liability*, 4 May 2020) 530.

[131] YouTube's Content ID system illustrates the potential for abuse of a N&A procedure. Previously, copyright holders on YouTube could file manual claims on any video using their content, whatever the length of the clip, and choose to either block the video or redirect revenue to themselves instead of the uploader. The latter option created a financial incentive in favour of abusive manual claims. Subsequently, in September 2019, YouTube changed their policy to remove the possibility of monetising claims based on 'very short or unintentional uses of music'. While YouTube acknowledged that the update might temporarily lead to increased blocking of content, the platform hoped to facilitate a culture shift in the way rightsholders use Content ID. See Hayleigh Bosher, 'YouTube Takes Copyright Law into Their Own Hands with New Policy on Music Infringement' (*The IPKat*) <http://ipkitten.blogspot.com/2019/09/youtube-takes-copyright-law-into-their.html> accessed 27 August 2020.

3.4. Transparency and Accountability

Finally, **platforms need to remain transparent and accountable to their users at every stage of their deepfake response**. Because platforms act as the 'new governors' in today's digital world,[132] increasing the availability of information and strengthening accountability mechanisms will be fundamental to ensure that platforms serve the interests of their users and of democratic societies at large.

At the deepfake identification stage, questions about algorithmic accountability will arise, as platforms will likely rely on a combination of deepfake detection technologies and automated filtering technologies to flag content for review. Automated moderation, like other forms of algorithmic decision-making ('ADM'), can be subject to systematic biases or errors that could lead to over-blocking. They also tend to underperform relative to human moderators when required to make context-sensitive assessments of content.[133] Possible ways of counteracting the risks of ADM here include the involvement of human oversight in the process of decision-making after initial identification; greater transparency through disclosure of information about the algorithms used to detect and filter deepfakes; and 'black-box tinkering', which would test how a platform would respond to a realistic and representative set of content submitted for review.[134]

Transparency is also crucial in the drafting and implementation of platforms' content policies. The terms and scope of any restriction on deepfakes should be made clear, and platforms should also be transparent about how they intend to enforce their policies. In the future, platforms should aim to release regular transparency reports (or update their current practice in this regard) to include statistics on the prevalence of deepfake content,[135] accuracy of deepfake detection, enforcement of relevant content policies, user appeals of deepfake moderation decisions, and the outcomes of such appeals. These

[132] Kate Klonick, 'The New Governors: The People, Rules, and Processes Governing Online Speech' (2018) 131 Harvard Law Review 1598.

[133] Singh (n 102).

[134] Elkin-Koren and Perel (n 89) 676.

[135] Prevalence can refer to 'an estimate of the percentage of the total content-views…that are of violating content', or alternatively, the number of 'violating content' posts as a percentage of total posts. Ideally, reporting both metrics would provide the most complete picture. 'Report of the Facebook Data Transparency Advisory Group' 18 <https://law.yale.edu/sites/default/files/area/center/justice/document/dtag_report_5.22.2019.pdf> accessed 17 July 2020.

reports could benefit not only users and researchers, but also platforms themselves by engendering trust in their deepfake moderation practices.

Likewise, respect for due process in the N&A and appeals procedure requires platforms to embrace transparency. From takedown notices issued to a deepfake-posting user, to explanations of moderation decisions and appeal outcomes, it is essential that platforms provide all the necessary information for users to exercise their rights and seek review.

Lastly, the establishment of an independent body to review content moderation decisions (deepfake-related and otherwise) upon appeal could facilitate a separation of powers that would enhance platforms' accountability to their users. Decisions from an external body could provide crucial feedback to inform the development of platforms' practices and policies in the same way that judicial decisions can often influence legislation. This type of independent body could come in the form of a platform-specific group provided with the power to adjudicate individual cases, such as Facebook's oversight board.[136] Alternatively, it might be conceptualised as a broader Social Media Council jointly created and representative of a far larger group of stakeholders including various social media companies, journalists, regulators, lawyers, academics, and so on.[137]

[136] 'Welcoming the Oversight Board' (*About Facebook*, 6 May 2020) <https://about.fb.com/news/2020/05/welcoming-the-oversight-board/> accessed 31 August 2020.
[137] 'Social Media Councils: Consultation' (*ARTICLE 19*) <https://www.article19.org/resources/social-media-councils-consultation/> accessed 29 August 2020.

Conclusion

In December 2018, the number of deepfakes identified online was recorded at 7,964. By June 2020, that number had reached 49,081.[138] These statistics highlight a daunting truth: deepfakes are here to stay, and their numbers are growing exponentially. Now that deepfakes have navigated their way from fringe internet forums into mainstream awareness and use, the Rubicon cannot be uncrossed.

As deepfakes continue to evolve and reach new audiences, malicious and harmful applications of the technology will inevitably increase. Individuals and organisations who find themselves targeted may eventually seek legal recourse—but first, they will turn to online platforms as their port of call.

The trend towards platform responsibility and the imposition of monitoring obligations provides an impetus for private platforms to act upon deepfakes even if they cannot be held liable for the harms caused by such content. At the same time, platforms who act as gatekeepers to democratic discourse should strive to adhere to human rights principles.

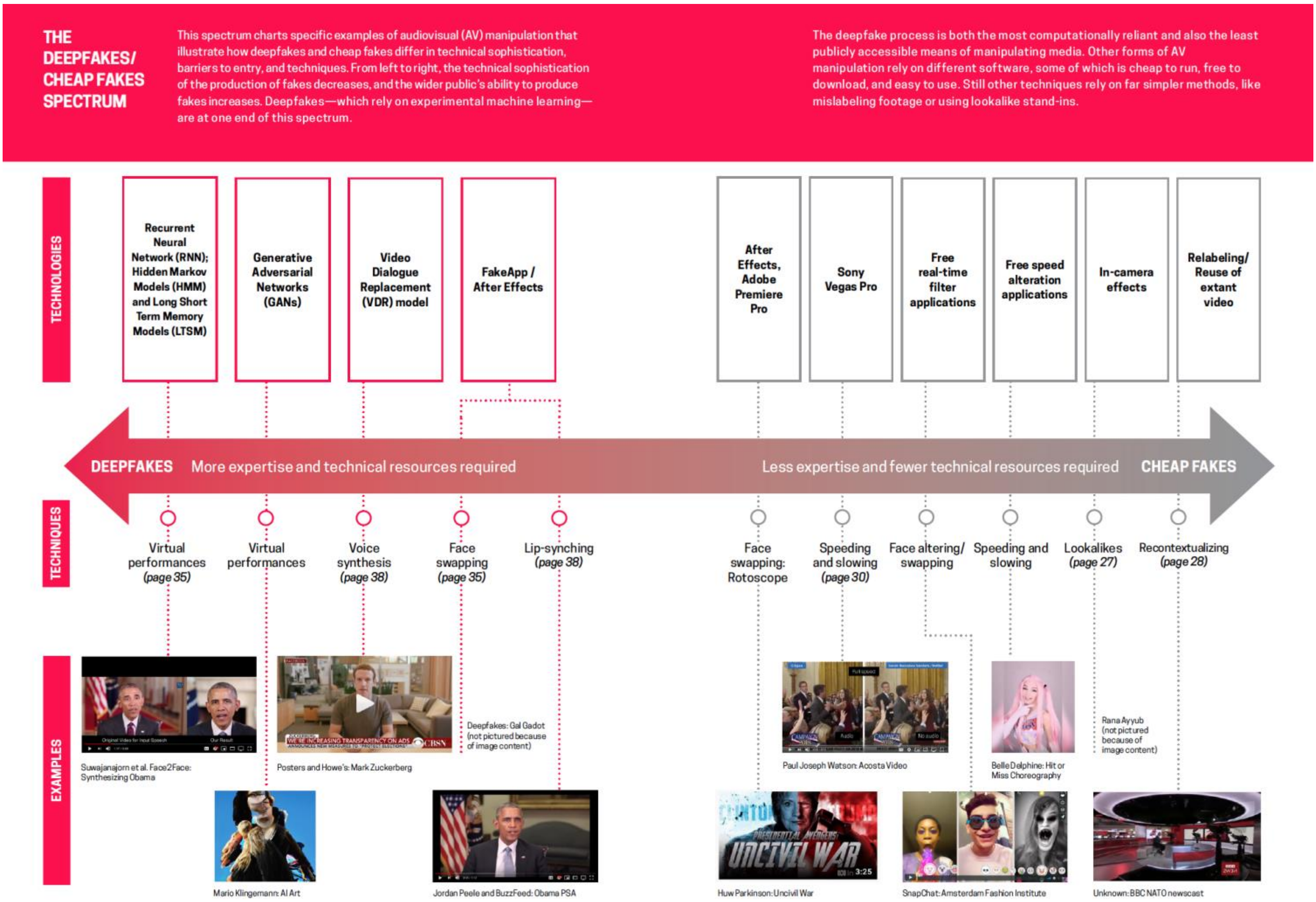
To help platforms achieve this delicate balance, governments should upgrade the legal framework to provide clearer guidance for platform action. Platforms, in turn, will need to moderate deepfakes proactively and reactively to satisfy both regulators and end users. The proposals I have set out in this Legal Opinion are designed to further both aims, taking into account legal and practical considerations.

Above all, the deepfake phenomenon is a microcosm of the broader struggle that embroils our modern information ecosystem. A central conflict between purveyors of deception and harm, and those who seek to stop them, threatens to inflict collateral damage on internet users' rights and freedoms. Platforms are, by virtue of their functions and positioning in modern society, at the frontline of the defensive campaign. But they will need the support of a full-fledged, multi-stakeholder effort to combat the spread of disinformation and misinformation through synthetic media and other means.

[138] 'Deepfake Threat Intelligence: A Statistics Snapshot from June 2020' (*Sensity*, 3 July 2020) <https://sensity.ai/deepfake-threat-intelligence-a-statistics-snapshot-from-june-2020/> accessed 30 August 2020.

Appendices

Appendix A: Spectrum of Audio-Visual Manipulation[139]



THE DEEPFAKES/ CHEAP FAKES SPECTRUM

This spectrum charts specific examples of audiovisual (AV) manipulation that illustrate how deepfakes and cheap fakes differ in technical sophistication, barriers to entry, and techniques. From left to right, the technical sophistication of the production of fakes decreases, and the wider public's ability to produce fakes increases. Deepfakes—which rely on experimental machine learning—are at one end of this spectrum.

The deepfake process is both the most computationally reliant and also the least publicly accessible means of manipulating media. Other forms of AV manipulation rely on different software, some of which is cheap to run, free to download, and easy to use. Still other techniques rely on far simpler methods, like mislabeling footage or using lookalike stand-ins.

TECHNOLOGIES

Recurrent Neural Network (RNN); Hidden Markov Models (HMM) and Long Short Term Memory Models (LTSM)

Generative Adversarial Networks (GANs)

Video Dialogue Replacement (VDR) model

FakeApp / After Effects

After Effects, Adobe Premiere Pro

Sony Vegas Pro

Free real-time filter applications

Free speed alteration applications

In-camera effects

Relabeling/ Reuse of extant video

DEEPFAKES More expertise and technical resources required

Less expertise and fewer technical resources required CHEAP FAKES

TECHNIQUES

Virtual performances *(page 35)*

Virtual performances

Voice synthesis *(page 38)*

Face swapping *(page 35)*

Lip-synching *(page 38)*

Face swapping: Rotoscope

Speeding and slowing *(page 30)*

Face altering/ swapping

Speeding and slowing

Lookalikes *(page 27)*

Recontextualizing *(page 28)*

EXAMPLES

Suwajanajorn et al. Face2Face: Synthesizing Obama

Posters and Howe's: Mark Zuckerberg

Deepfakes: Gal Gadot (not pictured because of image content)

Mario Klingemann: AI Art

Jordan Peele and BuzzFeed: Obama PSA

Paul Joseph Watson: Acosta Video

Belle Delphine: Hit or Miss Choreography

Rana Ayyub (not pictured because of image content)

Huw Parkinson: Uncivil War

SnapChat: Amsterdam Fashion Institute

Unknown: BBC NATO newscast

[139] Britt Paris and Joan Donovan, 'Deepfakes and Cheap Fakes' (Data & Society 2019) 10 <https://datasociety.net/library/deepfakes-and-cheap-fakes/>.

Appendix B: Excerpt from Manila Principles on Intermediary Liability[140]

Principle 5: Laws and content restriction policies and practices must respect due process

1. Before any content is restricted on the basis of an order or a request, the intermediary and the user content provider must be provided an effective right to be heard except in exceptional circumstances, in which case a post facto review of the order and its implementation must take place as soon as practicable.

2. Any law regulating intermediaries must provide both user content providers and intermediaries the right of appeal against content restriction orders.

3. Intermediaries should provide user content providers with mechanisms to review decisions to restrict content in violation of the intermediary's content restriction policies.

4. In case a user content provider wins an appeal under (b) or review under (c) against the restriction of content, intermediaries should reinstate the content.

5. An intermediary should not disclose personally identifiable information about a user without an order by a judicial authority. An intermediary liability regime must not require an intermediary to disclose any personally identifiable user information without an order by a judicial authority.

6. When drafting and enforcing their content restriction policies, intermediaries should respect human rights. Likewise, governments have an obligation to ensure that intermediaries' content restriction policies respect human rights.

[140] 'Manila Principles' (n 91).

Bibliography

Cases

Bonnard v Perryman [1891] 2 Ch 269. 9

Case C-18/18 *Eva Glawischnig-Piesczek v Facebook Ireland Limited Judgment* (OJ), 22/11/2019 .... 18, 19, 21, 31

Case C-324/09 *L'Oréal SA and Others v eBay International AG and Others,* EU:C:2011:474 (12 July 2011) 16

Case C-610/15 *Stichting Brein v Ziggo BV* ECLI:EU:C:2017:456 14

Douglas & Ors v Hello! Ltd & Ors [2005] EWCA Civ 595 9

Eastwood v. Superior Court (National Enquirer, Inc.) (1983) 149 Cal. App. 3d 409. 7

In re NCAA Student-Athlete Name & Likeness Licensing Litigation, 724 F.3d 1268, 1279 (9th. Cir. 2013) 8

Irvine v Talksport Ltd [2003] EWCA Civ 423 8

Lachaux v Independent Print Ltd [2019] UKSC 27 9

Lyngstad v Anabas Products Ltd [1977] F.S.R. 62. 8

Manhattan Community Access Corp. v. Halleck, 139 S. Ct. 1921 (2019) 21

Robyn Rihanna Fenty v Arcadia Group Brands Ltd (T/A Topshop) [2013] EWHC 2310 (Ch). 8

Superior Court of Justice Fourth Panel Google Brazil v Dafra [24 March 2014] Special Appeal no. 1306157/SP (Bra.) 20

Tamiz v Google [2013] EWCA Civ 68 13

Thornton v Telegraph Media Group Ltd [2011] 1 WLR 1985. 9

Zhong Qin Wen v Baidu [2014] Gao Min Zhong Zi no. 2045 (Ch.) 20

Statutes

's Defamation Act 2 .... 8, 9, 15

Cal. Civ. Code § 1708.86 (2019) 10

Cal. Civ. Proc. Code § 35 (2019) 10

Cal. Elec. Code § 20010 (2019) 10

Defamation Act 2013 15

Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC (Text with EEA relevance.) OJ L 130/92 ('Copyright Directive') 18

Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('E-Commerce Directive') OJ L 178 .... 16, 18, 19, 26

Section 230(c) of the Communications and Decency Act (47 U.S.C. § 230(c)) 14, 15, 17

Tex. Elec. Code § 255.004 (2019). 11

Va. Code Ann. § 18.2-386.2 (2019). 11

Treaties

Art.10(2) ECHR 26

Art.8 ECHR 9

Charter of Fundamental Rights of the European Union 22

Regulations

Defamation (Operators of Websites) Regulations 2013, SI 2013/3028 15

Secondary Sources

Ajder H, 'Social Engineering And Sabotage: Why Deepfakes Pose An Unprecedented Threat To Businesses' (*Deeptrace*, 3 October 2019) <https://deeptracelabs.com/social-engineering-and-sabotage-why-deepfakes-pose-an-unprecedented-threat-to-businesses/> accessed 10 March 2020

——, 'The State of Deepfakes: Landscape, Threats, and Impact' (Deeptrace 2019)

Bateman J, 'Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenarios' (Carnegie Endowment for International Peace) "Cybersecurity and the Financial System" 7

Bosher H, 'YouTube Takes Copyright Law into Their Own Hands with New Policy on Music Infringement' (*The IPKat*) <http://ipkitten.blogspot.com/2019/09/youtube-takes-copyright-law-into-their.html> accessed 27 August 2020

Buiten MC, de Streel A and Peitz M, 'Rethinking Liability Rules for Online Hosting Platforms' (CRC TR 224 2019) Discussion Paper No. 074 <https://www.ssrn.com/abstract=3350693> accessed 26 July 2020

Caldwell M and others, 'AI-Enabled Future Crime' (2020) 9 Crime Science 14

Cellan-Jones JW Rory, 'Facebook's "supreme Court" Members Announced' *BBC News* (6 May 2020) <https://www.bbc.com/news/technology-52558559> accessed 23 August 2020

Chesney R and Citron DK, 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security' (2019) 107 California Law Review 1753

'China Seeks to Root out Fake News and Deepfakes with New Online Content Rules' *Reuters* (29 November 2019) <https://www.reuters.com/article/us-china-technology-idUSKBN1Y30VU> accessed 1 April 2020

Cole S, 'Reddit Just Shut Down the Deepfakes Subreddit' (*Vice*, 7 February 2018) <https://www.vice.com/en_us/article/neqb98/reddit-shuts-down-deepfakes> accessed 28 March 2020

——, 'The Ugly Truth Behind Pornhub's "Year In Review"' (*Vice*, 18 February 2020) <https://www.vice.com/en_us/article/wxez8y/pornhub-year-in-review-deepfake> accessed 29 March 2020

'Community Guidelines Strike Basics - YouTube Help' <https://support.google.com/youtube/answer/2802032?hl=en-GB> accessed 28 August 2020

'Community Standards' (*Facebook*) <https://www.facebook.com/communitystandards/manipulated_media/> accessed 24 June 2020

Cristian Vaccari and Andrew Chadwick, ‘Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News’ (2020) 6 Social Media + Society

‘Deepfake Threat Intelligence: A Statistics Snapshot from June 2020’ (*Sensity*, 3 July 2020) <https://sensity.ai/deepfake-threat-intelligence-a-statistics-snapshot-from-june-2020/> accessed 30 August 2020

‘Deepfakes’ (*Future Advocacy*) <https://futureadvocacy.com/deepfakes/> accessed 31 August 2020

DelViscio J, ‘A Nixon Deepfake, a “Moon Disaster” Speech and an Information Ecosystem at Risk’ (*Scientific American*) <https://www.scientificamerican.com/article/a-nixon-deepfake-a-moon-disaster-speech-and-an-information-ecosystem-at-risk1/> accessed 21 August 2020

Devin Coldewey, ‘Facebook’s “Deepfake Detection Challenge” Yields Promising Early Results’ (*TechCrunch*) <https://social.techcrunch.com/2020/06/12/facebooks-deepfake-detection-challenge-yields-promising-early-results/> accessed 27 August 2020

‘Disinformation, n.’ <https://www.oed.com/view/Entry/54579> accessed 23 June 2020

‘Do Not Post Involuntary Pornography’ (*Reddit Help*) <http://reddit.zendesk.com/hc/en-us/articles/360043513411> accessed 29 August 2020

Elkin-Koren N and Perel M, ‘Guarding the Guardians: Content Moderation by Online Intermediaries and the Rule of Law’ (*Oxford Handbook of Online Intermediary Liability*, 4 May 2020)

Emily Saltz and others, ‘It Matters How Platforms Label Manipulated Media. Here Are 12 Principles Designers Should Follow.’ [2020] *The Startup Magazine* <https://medium.com/swlh/it-matters-how-platforms-label-manipulated-media-here-are-12-principles-designers-should-follow-438b76546078> accessed 14 August 2020

EU Commission, ‘Code of Practice on Disinformation’ (*Shaping Europe’s digital future - European Commission*, 26 September 2018) <https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation> accessed 24 June 2020

‘Facebook, Microsoft, and Others Launch Deepfake Detection Challenge’ (*VentureBeat*, 11 December 2019) <https://venturebeat.com/2019/12/11/facebook-microsoft-and-others-launch-deepfake-detection-challenge/> accessed 27 August 2020

Farish K, ‘Do Deepfakes Pose a Golden Opportunity? Considering Whether English Law Should Adopt California’s Publicity Right in the Age of the Deepfake’ (2020) 15 Journal of Intellectual Property Law & Practice 40

Ferraro M, ‘Deepfake Legislation: A Nationwide Survey—State and Federal Lawmakers Consider Legislation to Regulate Manipulated Media’ (WilmerHale 2019)

Frosio G and Mendis S, ‘Monitoring and Filtering: European Reform or Global Trend?’ (*Oxford Handbook of Online Intermediary Liability*, 4 May 2020)

Frosio GF, ‘Reforming Intermediary Liability in the Platform Economy: A European Digital Single Market Strategy’ (2017) 112 Northwestern University Law Review 19

Gabison GA and Buiten MC, ‘Platform Liability in Copyright Enforcement’ (2020) 21 Columbia Science & Technology Law Review 237

Goldman E, ‘The Complicated Story of FOSTA and Section 230’ (2019) 17 First Amendment Law Review 279

——, ‘An Overview of the United States’ Section 230 Internet Immunity’ (*Oxford Handbook of Online Intermediary Liability*, 4 May 2020)

Grimmelmann J, ‘The Virtues of Moderation’ (2015) 17 Yale Journal of Law and Technology

Hotten R, ‘Elon Musk Tweet Wipes $14bn off Tesla’s Value’ *BBC News* (1 May 2020) <https://www.bbc.com/news/business-52504187> accessed 1 September 2020

‘How PhotoDNA for Video Is Being Used to Fight Online Child Exploitation | Microsoft On The Issues’ (*On the Issues*, 12 September 2018) <https://news.microsoft.com/on-the-issues/2018/09/12/how-photodna-for-video-is-being-used-to-fight-online-child-exploitation/> accessed 27 August 2020

Isaac M, ‘Facebook Hampers Do-It-Yourself Mask Efforts’ *The New York Times* (5 April 2020) <https://www.nytimes.com/2020/04/05/technology/coronavirus-facebook-masks.html> accessed 23 August 2020

Jess Miers, ‘SCL Webinar: An Overview of Section 230 and Content Moderation - Online Intermediary Liability in the US’ (Society for Computers and Law, 6 August 2020)

Joshua Rothkopf, ‘Deepfake Technology Enters the Documentary World’ (*The New York Times*, 1 July 2020) <https://www.nytimes.com/2020/07/01/movies/deepfakes-documentary-welcome-to-chechnya.html> accessed 30 August 2020

Klonick K, ‘The New Governors: The People, Rules, and Processes Governing Online Speech’ (2018) 131 Harvard Law Review 1598

Kuczerawy A, ‘From “Notice and Takedown” to “Notice and Stay Down”: Risks and Safeguards for Freedom of Expression’ (*Oxford Handbook of Online Intermediary Liability*, 4 May 2020)

Laidlaw EB, *Regulating Speech in Cyberspace: Gatekeepers, Human Rights and Corporate Responsibility* (Cambridge University Press 2015)

Langvardt K, ‘Regulating Online Content Moderation’ (2018) 106 Georgetown Law Journal 1353

Leonard Rosenthal and others, ‘Setting the Standard for Digital Content Attribution’ (The Content Authenticity Initiative 2020) <https://documentcloud.adobe.com/link/track?uri=urn%3Aaaid%3Ascds%3AUS%3A2c6361d5-b8da-4aca-89bd-1ed66cd22d19#pageNum=1> accessed 27 August 2020

Mach J, 'Deepfakes: The Ugly, and The Good' (*Medium*, 2 December 2019) <https://towardsdatascience.com/deepfakes-the-ugly-and-the-good-49115643d8dd> accessed 27 August 2020

Madiega T, *Reform of the EU Liability Regime for Online Intermediaries: Background on the Forthcoming Digital Services Act : In-Depth Analysis*. (European Parliamentary Research Service 2020) <https://op.europa.eu/publication/manifestation_identifier/PUB_QA0420239ENN> accessed 23 June 2020

'Manila Principles' <https://www.manilaprinciples.org/> accessed 12 July 2020

McGonagle T, 'Free Expression and Internet Intermediaries: The Changing Geometry of European Regulation' (*Oxford Handbook of Online Intermediary Liability*, 4 May 2020)

Meskys E and others, 'Regulating Deep Fakes: Legal and Ethical Considerations' (2020) 15 Journal of Intellectual Property Law & Practice 24

Mostert F, '"Digital Due Process": A Need for Online Justice' (2020) 15 Journal of Intellectual Property Law & Practice 378

'Online Harms White Paper' (*GOV.UK*) <https://www.gov.uk/government/consultations/online-harms-white-paper/online-harms-white-paper> accessed 23 August 2020

'Online Platforms (Accompanying the Document Communication on Online Platforms and the Digital Single Market)' (European Commission) SWD(2016) 172 final

Paris B and Donovan J, 'Deepfakes and Cheap Fakes' (Data & Society 2019) <https://datasociety.net/library/deepfakes-and-cheap-fakes/>

Perot E and Mostert F, 'Fake It till You Make It: An Examination of the US and English Approaches to Persona Protection as Applied to Deepfakes on Social Media' (2020) 15 Journal of Intellectual Property Law & Practice 32

'Policy on Impersonation - YouTube Help' <https://support.google.com/youtube/answer/2801947?hl=en&ref_topic=9282365> accessed 17 July 2020

'Project Overview ‹ Detect DeepFakes: How to Counteract Misinformation Created by AI' (*MIT Media Lab*) <https://www.media.mit.edu/projects/detect-fakes/overview/> accessed 27 August 2020

'Reddit Account and Community Restrictions - Do Not Impersonate an Individual or Entity' (*Reddit Help*) <https://www.reddithelp.com/en/categories/rules-reporting/account-and-community-restrictions/do-not-impersonate-individual-or> accessed 17 July 2020

'Report of the Facebook Data Transparency Advisory Group' <https://law.yale.edu/sites/default/files/area/center/justice/document/dtag_report_5.22.2019.pdf> accessed 17 July 2020

Riordan J, *The Liability of Internet Intermediaries* (First edition, Oxford University Press 2016)

‘R/Redditsecurity - Updates to Our Policy Around Impersonation’ (*reddit*) <https://www.reddit.com/r/redditsecurity/comments/emd7yx/updates_to_our_policy_around_impersonation/> accessed 17 July 2020

Silverman C, ‘How To Spot A DeepFake Like The Barack Obama-Jordan Peele Video’ (*BuzzFeed*) <https://www.buzzfeed.com/craigsilverman/obama-jordan-peele-deepfake-video-debunk-buzzfeed> accessed 31 August 2020

Singh S, ‘Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content’ (Open Technology Institute (New America) 2019)

‘Social Media Councils: Consultation’ (*ARTICLE 19*) <https://www.article19.org/resources/social-media-councils-consultation/> accessed 29 August 2020

‘Social Media Giants Warn of AI Moderation Errors as Coronavirus Empties Offices’ *Reuters* (18 March 2020) <https://www.reuters.com/article/us-health-coronavirus-google-idUSKBN2133BM> accessed 23 August 2020

‘Soft Law’ (*obo*) <https://www.oxfordbibliographies.com/view/document/obo-9780199796953/obo-9780199796953-0040.xml> accessed 25 August 2020

‘Spam, Deceptive Practices & Scams Policies - YouTube Help’ <https://support.google.com/youtube/answer/2801973?hl=en> accessed 17 July 2020

Stewart E, ‘The next Big Battle over Internet Freedom Is Here’ (*Vox*, 23 April 2018) <https://www.vox.com/policy-and-politics/2018/4/23/17237640/fosta-sesta-section-230-internet-freedom> accessed 25 August 2020

‘Synthetic and Manipulated Media Policy’ (*Twitter*) <https://help.twitter.com/en/rules-and-policies/manipulated-media> accessed 10 March 2020

‘The Best, Most Impressive Audio Deepfakes on the Web | Digital Trends’ <https://www.digitaltrends.com/news/best-audio-deepfakes-web/> accessed 28 August 2020

‘Trump Signs Executive Order Targeting Twitter after Fact-Checking Row’ *BBC News* (28 May 2020) <https://www.bbc.com/news/technology-52843986> accessed 21 June 2020

Vincent J, ‘Disney’s Deepfakes Are Getting Closer to a Big-Screen Debut’ (*The Verge*, 29 June 2020) <https://www.theverge.com/2020/6/29/21306889/disney-deepfake-face-swapping-research-megapixel-resolution-film-tv> accessed 21 August 2020

Vosoughi S, Roy D and Aral S, ‘The Spread of True and False News Online’ (2018) 359 Science 1146

Vox, *The Most Urgent Threat of Deepfakes Isn’t Politics* <https://www.youtube.com/watch?v=hHHCrf2-x6w> accessed 20 August 2020

Warner M and Rubio M, ‘Deepfakes Letter to Facebook’ (*Scribd*) <https://www.scribd.com/document/428320935/Deepfakes-Letter-to-Facebook> accessed 10 March 2020

‘Welcoming the Oversight Board’ (*About Facebook*, 6 May 2020) <https://about.fb.com/news/2020/05/welcoming-the-oversight-board/> accessed 31 August 2020

YouTube Content ID (2010) <https://www.youtube.com/watch?v=9g2U12SsRns> accessed 27 August 2020