



International Journal of Intellectual Property Management

ISSN online: 1478-9655 - ISSN print: 1478-9647

<https://www.inderscience.com/ijipm>

The danger of deepfakes, Indian laws and platform responsibility

Smarnika Srivastava

DOI: [10.1504/IJIPM.2024.10062876](https://doi.org/10.1504/IJIPM.2024.10062876)

Article History:

Received:	26 July 2023
Last revised:	19 January 2024
Accepted:	19 January 2024
Published online:	24 March 2025

The danger of deepfakes, Indian laws and platform responsibility

Smarnika Srivastava

Jindal Global Law School,
O.P. Jindal Global University,
Sonipat, Haryana, India
Email: Smarnika.srivastava@jgu.edu.in

Abstract: In the realm of digital media, deepfakes have become the most sophisticated means of spreading disinformation and indulging in harms such as pornography, commercial frauds, fake news, and political propaganda. This paper analyses a few current legal solutions in India which can be repurposed to apply to deepfakes, such as copyright, criminal laws, right to privacy, etc., however, it has been argued that these are insufficient to tackle these threats. In recent years, India has seen a shift towards platform responsibility and pro-active monitoring, in light of which, the author discusses the reasons why intermediaries should be responsible for moderating content. To do the same, platforms need to strike a balance between respecting the fundamental rights of users and due process principles while taking down harmful content. The author also suggests that criminal legislation should be updated to encompass the harmful uses of deepfakes so that platforms have a legal basis to respond to them, amongst other solutions. Further, some recommendations have been made in the paper on how platforms can update their policies and guidelines to combat the same.

Keywords: deepfakes; pro-active monitoring; content moderation policy; platform responsibility; intermediary liability; IT Rules 2021.

Reference to this paper should be made as follows: Srivastava, S. (2025) 'The danger of deepfakes, Indian laws and platform responsibility', *Int. J. Intellectual Property Management*, Vol. 15, No. 1, pp.60–76.

Biographical notes: Smarnika Srivastava is working as a Lecturer at the Jindal Global Law School, O.P. Jindal Global University, Sonipat, Haryana, India. She has an LLM in Intellectual Property Laws from the King's College, London, and an undergraduate law honours degree from the O.P. Jindal Global University, Sonipat.

1 Introduction: deepfakes – boon or bane?

It is often said that the camera does not lie. However, in today's technologically savvy world, where digital media is the easiest and fastest medium of communication, deepfakes have become an advanced way of disseminating misinformation (Caldwell et al., 2020; Kabbara, 2021; Satariano and Mozur, 2023). Deepfake technology is an advanced algorithm which generates hyper realistic videos using a person's face, voice or likeness using machine learning techniques. These fake videos/audios are then

manipulated to appear as if the person is saying or doing certain things (Ajder et al., 2019).

Like every coin has two sides, any new technology can either act as a boon or bane. Depending on the user's intention behind creating and disseminating deepfakes, it can either be used for good or evil. Some of its innocuous uses could be made by research institutions to create educational videos on technology, development of creative works such as films and documentaries, etc. (Cruz, 2020). One such use was made by the company Rephrase.ai, for developing their Rakhee campaign. They made a deepfake of a famous Indian actor, where his voice and face were manipulated to seem as if he is personally congratulating his fans by taking their name. The actor had legally licensed the company to use his likeness to make such videos (The Hard Copy, 2021). There are other Indian companies which use licensed deepfakes of many Indian celebrities to send special messages, such as birthday or anniversary wishes to their customers for money. This is how deepfakes are used in a positive manner as a strategy for brand building and marketing.

However, this paper only focuses on the harm perpetrated by the usage of deepfakes, which has been rated as the most serious AI crime threat in a study conducted in 2020 (Caldwell et al., 2020). Deepfakes have become a cause of great concern for the society. These can be used to manipulate public opinion by generating different types of content, such as fake pornographic videos or political videos of a person without their consent, which harms their life, both personally and professionally (Westerlund, 2019). Its believability is becoming increasingly worrisome because the technology is becoming better, leading to its easy distribution to reach a wide audience, thus leading to the dominant spread of misinformation (Kabbara, 2021). Technology pioneer Professor Hao Li warned that deepfakes will get so advanced that there will be no way to detect them anymore (Stankiewicz, 2019). Therefore, it has become imperative to debate on how to control the proliferation of unauthorised and harmful deepfakes and the role which platforms can play in responding to the threat.

Currently, India does not have a specific legislation which regulates the use of deepfakes. Even though there are certain existing laws which can be repurposed to apply to deepfakes, these are insufficient to tackle its harmful uses. Many commentators have started recognising the need for platforms to proactively monitor content, in order to respond to deepfakes (Chesney and Citron, 2019; Cruz, 2020; Grimmelmann, 2015; Laidlaw, 2015; Mostert, 2020). The recent legislative changes in the IT Rules, 2021 also require intermediaries to do the same, however, due to lack of legislative or policy principles, it is unclear how this can be achieved. It thus becomes difficult for platforms to implement robust content moderation mechanisms which balance the rights of the users and due process, while upholding the democratic principles at large.

This paper is divided into six sections. In Section 2, the author has identified some of the key threats posed by deepfakes which need to be addressed and regulated, in the context of India. Since there is no existing law tailored to deepfakes in particular, the current legal solutions in Indian laws have been analysed in Section 3, which can be used to respond to deepfake threats. However, these current solutions are inadequate to combat the spread of deepfakes, as discussed in Section 4. Building upon the same, in Section 5, the author has delved into the debate on whether or not platforms should be responsible for proactively monitoring and moderating content, and the justification for their role in doing the same. Finally, in Section 6, suggestions have been given on how Indian legal

framework can be improved, through which platforms will be able to better respond to deepfake threats. Further, the paper also gives suggestions on content moderation policies of platforms which should balance the interests of the intermediaries, users, and democracy at large.

2 Threats posed by the use of unauthorised deepfakes

This chapter identifies and mentions some of the key ways in which the non-consensual usage of deepfakes has become a threat to the society.

2.1 Political propaganda during elections

There have been several instances around the world where deepfakes have been used to further political agendas by exploiting the likeness and image of famous political personalities. They could be portrayed as using a racial slur, taking a bribe, confessing complicity in a crime, etc. (Kabbara, 2021). In 2019, a doctored video of Nanci Pelosi, the US Speaker of the House went viral, which portrayed her slurring her speech (CBS News, 2019). In the UK, the think tank Future Advocacy released deepfake videos of Boris Johnson and Jeremy Corbyn which showed them to be endorsing each other for the post of prime minister. This was done to raise awareness on the threat surrounding political disinformation and tactics used by parties during elections to mislead the public (Future Advocacy, 2019).

India is not averse to the threat posed by deepfakes in the political scenario as well. Many doctored videos have been used during elections to either further political agendas or tarnish the reputation of an opposition leader. In 2020, there were several reports alleging that the Bhartiya Janta Party's New Delhi unit had used deepfake technology to produce videos which portrayed their leader speaking in English and Haryanvi languages to reach a larger voter base. These videos were reportedly an alteration of a video made in Hindi language in which the politician was speaking about a completely different issue but seemed to be real, due to the perfect lip-synching (Dasgupta, 2020).

Political deepfakes could have profound implications for journalism and quality of democracy. With respect to viewers, the disinformation could diminish the trust they place in online news, which negatively affects their faith in public discourse and social media, due to unsurety about which news to believe in. A study published in 2020 found that altered videos leave its viewers feeling "uncertain about the truthfulness of their content... and results in lower levels of trust in news on social media" (Viccari and Chadwick, 2020). Further, online news platforms may start publishing false news, with the aim of getting more readership and publicity. The implications for the person portrayed are also dangerous, since they could lose important positions of power due to their reputation being tarnished, which would ultimately be detrimental to the democratic process of elections in India. On the flip side, politicians could also allege a real video as a deepfake to escape political accountability. This is known as the 'the liar's dividend' (Chesney and Citron, 2019). If left unchecked, political deepfakes will prove to be dangerous in many ways.

2.2 Non-consensual pornography

Most of the deepfakes on the internet are pornographic in nature (Ajder et al., 2019). Deepfake pornography began spreading on the internet several years ago when a Reddit user shared altered videos of actresses where their faces were swapped on the bodies of porn actors (The Times of India, 2023). Since then, there are thousands of such unauthorised videos which have been uploaded on several websites by deepfake creators. Many famous people, such as actors, journalists or public figures are targeted in this way, which harms their image and brand. In fact, some websites make commercial use of this technology by offering users to create their own deepfakes, allowing them to generate unauthorised sexual videos of people without their consent or knowledge (The Times of India, 2023). This technology could also be used to indulge in crimes such as revenge pornography, blackmail and extortion. According to a 2019 report, women are the exclusive targets of deepfake pornography (Ajder et al., 2019). It harms a person's dignity, causes psychological distress, and undermines their right to bodily autonomy.

In order to combat the same, platforms namely Reddit and Pornhub have taken steps such as prohibiting such content by updating their content policy and taking down and banning it from being searched (Cole, 2018). However, in reality, Pornhub has not been able to implement the same (Cole, 2020). The severe harms associated with deepfake pornography will continue to persist if there are no checks and balances in place.

2.3 Commercial frauds

If deployed commercially, deepfakes and bots could have potentially disastrous consequences. Criminals could impersonate a famous business leader by using deepfake voice phishing technique to scam customers financially or indulge in defrauding a business (Cruz, 2020). The influence wielded by any prominent person's activity can swiftly impact corporations and those associated with it. By the time it is uncovered whether the video is a deepfake or an original, the damage would already be done. For example, in 2019, a staff member of a British energy firm, who fell victim to a deepfake voice impersonating the CEO of the parent company, was fraudulently made to transfer \$250,000 (Tiwari, 2023).

These harms discussed above are pressing issues highlighting the danger of deepfakes, which need to be regulated through legal mechanisms. In India, till yet, there is no legislation which specifically addresses the issue of deepfakes. However, there are existing laws which can be repurposed to apply to the same. These have been discussed in the next chapter.

3 Current legal solutions in India to combat the harms of deepfake technology

In this chapter, the author analyses the current legal solutions in India for combating the potential harmful uses of deepfake technology, as previously discussed.

3.1 *Right to privacy*

Article 12 of the Universal Declaration of Human Rights Act, 1948, recognises the right to privacy as a basic human right. According to it, no person's honour or reputation can be arbitrarily or unlawfully harmed. The right to privacy has also been recognised as a fundamental right to life under Article 21 of the Constitution of India in the *Justice KS Puttaswamy v. Union of India* (2017) judgement. According to it, an individual's right to privacy can and must include an action against the right to control the dissemination of personal information. The author believes that this would include the non-consensual and unauthorised usage of their deepfakes, as it can harm their reputation and interfere with their bodily autonomy. Therefore, any person who creates and distributes unauthorised deepfakes of a person would be violating their privacy, on account of harming their reputation and honour. This would be applicable to all three types of deepfakes discussed in Section 2, since those involve harm to the victim's reputation and honour.

In India, since there is no exclusive law pertaining to image or personality rights, these are also protected by the right to privacy and publicity (Rana and Sharan, 2022). The right to publicity, which encompasses the right to control the commercial exploitation of one's personality, has been recognised as a part of right to privacy in *R. Rajagopal v. State of Tamil Nadu* (1995). The Supreme Court stated that a person has the right to protect their reputation by preventing the unauthorised use of their name, image, or likeness for commercial, advertisement, entertainment, or any other purposes without their consent. Therefore, if any person creates and distributes deepfakes of another person using the above aspects, they could be liable for violation of the right to publicity.

The *Digital Personal Data Protection Bill, 2022* also provides for the protection of the personal data of natural persons, which could either be collected online or offline. It puts restrictions on any use other than lawful and consensual use of such data (Jain and Jha, 2020). Individuals can demand for the correction and erasure of information collected by a data fiduciary, who can be charged with significant penalties for data breaches. The Bill also proposes to setup a Data Protection Board where consumers can file for a complaint in case the data fiduciary does not give a satisfactory response. Since the Bill is yet to be passed as an Act in India, its impact is left to be seen, however, the author submits that its applicability can be extended to the unauthorised circulation and usage of deepfakes.

3.2 *Criminal laws*

The Indian Penal Code (IPC), 1860 also has certain provisions which can apply to non-consensual distribution of deepfakes. Such content which is created with the intention to harass, intimidate, or defame someone would come under the purview of Section 499 for the offense of defamation. This would be applicable to unauthorised and fake pornographic or political deepfakes. Additionally, Section 292 on obscenity would also be applicable to the circulation and distribution of pornographic deepfakes. Further, if a deepfake is created with the intent to deceive or defraud its viewers, this could fall under the provision of Section 415 for fraud and cheating, such as impersonation and dishonest inducement.

3.3 Copyright and passing off

Copyright laws protect original works of authorship. If an individual takes a ‘selfie’ of themselves, they would have a copyright in it, since it would be “the result of author’s own skill, labour and effort” (Ahuja, 2017). Even if someone else records/clicks the picture consensually, the subject can claim joint authorship in it and it is not necessary that the person who clicks the shutter button of the camera would be the sole author of the work, to the exclusion of the subject (Simone, 2022). The author argues the level of creativity expended in clicking pictures or taking videos, such as angle, lighting, pose, etc. would fulfil the requirement of originality under Indian copyright laws.

The author submits that deepfakes, which involve the unauthorised manipulation or alteration of existing photos and videos, can be said to infringe the copyright owner’s right to reproduction of the copyrighted material if a substantive portion is reproduced. Deepfakes are creations based on pre-existing works¹, and thus, these can be considered as derivative work (Cruz, 2020). Since copyright law grants the permission to make derivative works exclusively to the copyright owner, if someone creates and distributes deepfakes without permission, it will be considered as copyright infringement (Cruz, 2020).

Copyright owners also have a claim under moral rights (Garg, 2021). As per Section 57 of the Copyright Act, 1957, moral rights are infringed if the work is modified or distorted in a way which damages its copyright owner’s integrity or reputation. This would apply to unauthorised deepfakes used for purposes such as pornography, commercial frauds, etc.

Platforms can also be held liable for copyright infringement. In *My Space Inc. v. Super Cassettes Industries Ltd.* (2016), the Delhi High Court held that an intermediary would be liable for copyright infringement under Section 51 of the Copyright Act, 1957, if it can be proved that they have ‘actual knowledge’ of the infringing content on their website and do not remove it within 36 hours of being notified. However, a major drawback is that a relief under copyright laws is not always available to the victims of unauthorised deepfakes, where they have been clicked or recorded without their knowledge.² Therefore, this cause of action would not usually be used in such cases. Further, intermediaries also have certain statutory protections (discussed in Subsection 3.5), which restricts their liability in India.

The author further submits that the tort of passing, which occurs when one party misrepresents their business, goods, or services as that of another party, leading to deception or confusion amongst consumers, would also be applicable to deepfakes. Therefore, if there is *damage to the goodwill or reputation* of a famous person, such as a celebrity, due to *misrepresentation* via the unauthorised usage of deepfakes, then they would have a suitable cause for action under passing off (Cruz, 2020). This would include a situation where a deepfake is presented in a way which creates a false association with a brand or person. The affected person would then have the right to seek remedy, such as injunction or damages, against the other party.

3.4 Information technology law and intermediary liability: old and new regime

The IT Act, 2000 deals with and defines various offences related to the breach of data and privacy of an individual and also has provisions for regulating the activities of electronic and social media intermediaries.

Although the IT Act, 2000 does not mention deepfakes specifically, but based on the nature and intent behind the legislation of some of the provisions, it can be argued that these would apply to deepfakes. Section 66D deals with punishment for cheating and impersonating someone else by using a computer device or resource. Therefore, if a person who makes a deepfake of a person where they are portrayed as saying or doing an act which would be used for the purpose of cheating would be liable under this provision. This would include acts such as asking for money from friends and family or indulging in commercial frauds such as defrauding a business by impersonating someone else. There are some provisions in the IT Act which can act as a deterrent against the generation and use of deepfake pornography as well. Section 66E deals with the punishment for violating the privacy of a person by non-consensually capturing, transmitting, or publishing images of their sexual parts. Section 67 deals with the punishment for publishing or transmitting obscene material through electronic means. Thus, if someone circulates or uploads deepfake videos of a person where they are engaging in nudity or sexual activity, such an act could amount to the potential violation of these provisions.

Another important provision is Section 79 of the IT Act, 2000, which encapsulates intermediary liability and safe harbour principles. It provides a safe harbour for intermediaries, according to which, they would not be deemed liable for any third-party information hosted on their website, provided that they fulfil certain conditions. These conditions have been laid down in the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021 (IT Rules, 2021). Before discussing the IT Rules, 2021, it is imperative to mention that under the old IT Rules, 2011, intermediaries were only required to publish rules and regulations, privacy policy and user agreement for access or usage of their platform. If the intermediary merely acts as a conduit, lacks the actual knowledge regarding the unlawful content and disables or removes access to it expeditiously upon being notified, they would not be held liable for hosting the same, as per the safe harbour principle. There was no provision which legally required the intermediaries to proactively monitor the content posted on their platform. In the case of *Kamlesh Vasvani v. Union of India* (2017), the Supreme Court held that imposing an obligation on intermediaries to monitor and filter objectionable content would be impractical and violate the freedom of speech and expression of internet users.

However, the new IT Rules, 2021 require intermediaries to also observe 'due diligence' in order to be exempt from liability, in addition to the requirements mentioned in the old rules. The rules have created a new category of 'significant social media intermediaries' (SSMI) who have a user base of 5 million users and above. Under Rule 4(2), it is mandatory for SSIMs to identify the first originator of a message/post if ordered by an executive authority or competent court. Some technical experts argue that complying with the requirement would mean breaking end-to-end encryption on messaging services such as WhatsApp (Mendiratta, 2021). The rules have placed additional restrictions on users for sharing of certain information, such as that which is 'harmful to a child', 'insulting on the basis of gender' and "knowingly and intentionally communicates any information which is patently false or misleading in nature but may reasonably be perceived as a fact." The author believes that this would directly apply to the situation where deepfakes are shared for indulging in unauthorised pornography, commercial frauds, and political propaganda, since these are 'false' and 'misleading' but may be perceived as the truth by its audience. For tackling with such content, SSIMs have to establish technological measures for identifying, removing, and restricting such content. Additionally, Rule 4(1)(p) requires intermediaries to take down non-consensual

sexually explicit content within 24 hours of receiving a complaint from the victim, which would be immensely beneficial for victims of deepfake pornography.

It can be concluded that under the new rules, intermediaries are now required to proactively monitor content posted on their platform. There are a few debates on whether intermediaries should be required to proactively monitor and moderate content, especially because this can be a threat to the fundamental rights of the citizens, such as freedom of speech and expression and right to privacy. This has been discussed in Section 5.

4 Inadequacy of the current legal solutions to combat the spread of deepfakes

The above-mentioned legal mechanisms in India would arguably be applicable to deepfakes, yet these existing options are inadequate to fully address the threats posed by its harmful usage and circulation.

Owing to their nature, some content, such as pornography or election propaganda, spreads faster than other types of deepfakes, once uploaded on the internet (Sisario and Vega, 2013). The right to privacy, criminal laws and IP laws can be used by victims to file a court case for obtaining personal remedies, such as compensation and permanent injunctions. Needless to say, it is a tedious process where the court's decision can take a few years to be rendered. Meanwhile, the content is still available for viewing and sharing on the internet. Therefore, these solutions are insufficient to alleviate the damage caused by the rapid spread of deepfakes.

One can argue that this spread can be controlled through issuing take down notices under the IT Act, 2000 and copyright laws.³ It is fast, cheap and time efficient to do the same. However, the author submits that there are various problems with the takedown system. Once a deepfake content is uploaded on the internet, it is subject to the whac-a-mole phenomenon. This means that it can easily be downloaded and re-uploaded on different websites. This would make it difficult for the victim to keep track of the many websites their deepfake gets uploaded on and subsequently, send a takedown notice every time it 'pops up' on the internet (Sisario and Vega, 2013). Therefore, it is difficult to contain its spread. Another issue is that deepfakes can be created and uploaded from anywhere in the world, which makes it challenging to determine the intermediary to whom the notice is to be issued. Moreover, some jurisdictions may not have takedown provisions equivalent to those available in the victim's jurisdiction, which might not allow them to send such a notice in the first place.

Even if one is eligible by law to send a takedown notice, this can only be done in cases where the victim has the ownership of copyright in the original content. However, there are many cases where the victim gets clicked or recorded without their consent and then those images or videos are altered to make deepfakes. In such a scenario, copyright would lie with the person behind the camera and not the victim, which will make it impossible for the latter to send a takedown notice to the intermediary.

Another hurdle is that technology is continuously evolving, which is making it difficult to detect and combat deepfakes effectively. This makes it harder to implement robust mechanisms to protect individuals from the unauthorised usage of their deepfakes. Even though the new IT Rules, 2021 have made placed a legal burden on the intermediaries to proactively monitor and moderate content, it does not specify how this

can be achieved.⁴ Moreover, deepfakes can be created and uploaded from anywhere in the world, which creates a complexity in determining which jurisdiction the offence falls under as well as coordinating international legal processes for investigation and prosecution (Cruz, 2020). I also believe that there is a serious lack of awareness amongst the general public and legal professionals regarding this technology and the threats posed by it, which hinders the development of effective strategies and policies to address it.

Considering these issues from a pragmatic perspective, the harm caused by a malicious deepfake materialises in the span of a few days, hours or even minutes. In comparison, courts take a lot of time to pass an order and give relief to the victim, causing irreparable damage to their reputation. By the time one recognises a deepfake employed for commercially or politically defrauding people or harming one's reputation, severe harm is already meted out.

In order to tackle the online dissemination of harmful deepfake content, the author believes that a comprehensive approach, which includes legal and technological measures for engaging online platforms, is essential. Given that the internet serves as a primary conduit for amplifying the impact and reach of a harmful deepfake, content moderation through an improved platform intervention framework can become crucial for early detection and swift measures can be taken to restrict their dissemination. This would greatly help in containing problems such as whac-a-mole phenomenon and jurisdictional issues, since there would not always be a need for victims to issue a takedown notice/approach the court for relief. In order to suggest how intermediaries can indulge in moderating content, it is first important to argue why they should be responsible for it in the first place. This has been discussed in Section 5.

5 Should intermediaries be responsible for moderating content?

As per the provisions laid down in the old IT Rules, 2011, intermediaries were not legally required to pro-actively monitor, identify, and remove the unlawful content hosted on their website. This arguably leads to the neglect of user safety and privacy and the dissemination of unauthorised deepfakes involving fake news⁵, non-consensual pornography⁶, hate speech, and commercial frauds, causing severe harm to the persons concerned, as well as its viewers.

On the other hand, the IT Rules, 2021 deem platforms as legally liable to proactively monitor the content on their website and have been heavily criticised. Automated content moderation systems have their limitations as they may struggle to detect nuanced forms of harmful content, such as unauthorised deepfakes which portray a person indulging in scamming customers. This can lead to under-identification. On the flip side, these systems can also result in overblocking, by flagging and removing legitimate and lawful deepfake content which does not violate any user policies (Dave, 2020). An example of this is when during COVID-19, Facebook's automated systems issued warnings and wrongly flagged groups of volunteers who were sewing handmade masks for frontline workers (Isaac, 2020). It could be that the system starts blocking legitimate and authorised deepfakes of celebrities posted by businesses and brands to endorse or advertise their products, or, even permitted deepfakes posted for educational purposes. The author submits that if not implemented correctly, this has the potential to violate the freedom of speech and expression of users, who may start self-censoring out of fear of

their content being flagged or removed. This will limit their ability to voice their opinions, engage in public discourse or share controversial ideas.

It is imperative to mention the pros and cons of proactive monitoring, since the author believes that balancing the need for it to address the sharing of harmful deepfakes, along with safeguarding user rights, privacy and freedom of expression is a significant challenge. It is also important to answer whether platforms are obligated to indulge in self-regulation, before giving suggestions on how unauthorised deepfakes can be regulated.

According to Laidlaw (2015), online platforms, by virtue of their de facto functions, can significantly impact democratic discourse due to their ability to control and gatekeep informational flow on the internet. There are three kinds of intermediaries. ‘Authority gatekeepers’ are platforms which have a central and inevitable impact on democratic discourse due to their considerable reach, such as Facebook, Instagram, YouTube, Twitter, etc. (SSMIs); ‘macro gatekeepers’ provide the route of informational flow and are essential for facilitating democratic discourse, such as search engines (Google, Yahoo, etc.); and ‘micro gatekeepers’ who are democratically significant, although on a small scale (individual news websites such as TheWire or Hindustan Times). Taking away from this theory, the author submits that macro and authority gatekeepers, such as SSMIs (e.g., WhatsApp) should be subject to a moderate level of obligations, which entails being subject to closer examination and responsibility. Laidlaw (2015) has further suggested that platforms are responsible for protecting the right to privacy and association, and the freedom of speech and expression of its users. Considering the impact that the new IT Rules, 2021 will have on content moderation, intermediaries and platforms will effectively assume the role of law enforcers, judges as well as administrative bodies at the same time. The author argues that when they chose to delete a user’s account, restrict certain content, or trace the first originator of a particular message/post, they affect democratic discourse by restricting their freedom of speech and expression and their right to privacy.

This should be balanced by the right to due process, as laid down in the Manila Principles on Intermediary liability, to which India is a signatory. Principle 5 states that platforms should offer users with the means through which they can review decisions made to restrict their content which is allegedly in violation of content restriction policies. These could be authorised deepfakes for business, advertisements, or educational purposes which have been overidentified and mistakenly flagged. If the user is able to successfully appeal, such lawful content should be reinstated. For example, Meta’s oversight board is equipped to hear appeals from users requesting removal of certain content or those contesting the wrongful removal of content (Wakefield, 2020). Mostert (2020) observes that due to the rapid and extensive nature of online speech, challenges such as over blocking and under blocking are inevitable where platforms will attempt to respond to the dual pressures of platform responsibility. Thus, upholding due process is an essential requirement to ensure the safeguarding of user’s rights to privacy, freedom of expression, and association.

6 Building a legal framework for regulating deepfakes, platform liability and intervention

6.1 Proposals for improving the legal framework to govern unauthorised deepfakes

In the evolving landscape of platform responsibility, there is an increasing pressure on platforms to monitor and address user-generated deepfake content (Cruz, 2020). As platforms start enhancing content moderation efforts, there is a simultaneous risk of under blocking and over blocking, thus impeding the users' rightful exercise of free speech. This phenomenon is known as the 'moderator's dilemma', which underscores the complex nature of balancing content moderation and the protection of users' rights (Goldman, 2020). In the context of deepfakes, platforms will have to strike a balance between safeguarding the user's right to freedom of speech and expression and due process, and on the other hand, protecting society against the harmful effects of unauthorised deepfake content (Mostert, 2020).

Firstly, for tackling the challenges posed by the moderator's dilemma in the context of deepfakes, the author believes that there is an ardent need for a comprehensive legal framework that enables and encourages platforms to take appropriate actions to respond to it. Criminal legislation should therefore be modified to expressly prohibit the gravest dangers associated with deepfakes, which have been discussed in Section 2. Several US states, such as California, Virginia and Texas have revised their legislations to address the specific instances of deepfakes (Poritz, 2023) and the author believes that the same should also be considered by India. The primary wrongdoer would be the user who shares the unauthorised deepfake and not the platform itself. If legislative measures are taken to encompass deepfakes within the purview of criminal law, then, from an individual's perspective, it would alleviate the burden faced by claimants who are seeking recourse which is not covered by any existing causes of actions (such as image rights, which is not recognised in India yet). Such an individual could approach the court, which could order the SSIMs to track the first originator of a particular post/message, as per the new IT Rules, 2021. From the perspective of intermediaries, bringing certain uses of deepfakes under the purview of criminal laws would provide platforms with a strong legal foundation to update and enforce their content policies (Cruz, 2020). This would help intermediaries to monitor and moderate content, as required under the new IT Rules, 2021, post which they would be able to claim the safe harbour principle under the IT Act, 2000. Therefore, statutory deepfake offences would serve as a solid framework on which intermediaries can build their own content policies. Moreover, this would also address the criticism that platforms opaquely and unilaterally restrict user content by subjectively determining what is 'harmful' rather than strictly 'illegal'. While it can be said that the legality of content is not always clearly defined, the notion of what is considered as 'harmful' is much more subjective, due to the lack of legislative guidance. Updating criminal legislature would ensure that intermediaries are responding to deepfake content as they are legally obligated to do so.⁷ This may also alleviate concerns regarding freedom of speech and expression as speech rights can be appropriately restricted within the boundaries of law, especially when it comes to crime prevention (Cruz, 2020). The same has been laid down under Article 19(2) of the Constitution of India, according to which the citizen's right to freedom of speech and expression cannot prevent the state to

make laws which impose reasonable restrictions in the interests of public order, morality, etc.

Secondly, the author submits that it would be immensely helpful for platforms if guidelines are developed on how to respond to and tackle the harmful uses of deepfakes. These could include approaches through which platforms can uphold due process with their policies, implement safeguards for fundamental rights such as freedom of speech and expression of the users and outline the category of deepfakes that ought to be restricted or exempted from restrictions. Some exemplary practices for content moderation concerning deepfakes could also be included (Mostert, 2020).

Thirdly, law enforcement agencies should develop a blocklist of sites that are primarily responsible for hosting illegal deepfake content. If sufficient legal basis is provided by national criminal laws, authorities could curate a blocklist including websites specifically dedicated to deepfake pornography or online platforms intentionally hosting deceptive news involving deepfakes (Cruz, 2020). City of London Police's IP Crimes Unit (PIPCU) has effectively implemented this initiative in collaboration UK's top level domain name registry 'Nominet'. Once any individual notifies PIPCU or when they detect any websites engaging in Intellectual Property Crimes, both entities work to suspend the domains within 48 hours (Batchelor, 2020). Through a cooperation of law enforcement agencies and domain name registries, a similar strategy could be employed to combat illegal deepfakes in India.

Fourthly, the author submits that even though the problem surrounding illegal uses of deepfakes is crucial to solve, it is important to keep in mind that liability frameworks should not be drastically changed without conducting thorough research into their potential consequences. Fearing liability, platforms may start to adopt risk averse approaches by choosing to completely eliminate certain channels of discourse. An example of this is when Craigslist, the American company which operates websites for classified advertisements, removed the 'personal ads' section after the passage of Stop Enabling Sex Traffickers Act (SESTA) in 2018 (Stewart, 2018). Even though this step halted its use made by sex trafficking rings, it also removed voluntary sex workers from the platform (Goldman, 2019). Consider a scenario where, as a response to political deepfakes, Twitter imposes a ban on the sharing of video clips of politicians. Additionally, the author believes that smaller platforms are more likely to eliminate forums for users who are legally exercising their freedom of speech, rather than risking an exposure to lawsuits. These aspects should be kept in mind while making policy changes.

6.2 Suggestions for platforms to moderate deepfakes

Content moderation can be defined as those governance mechanisms that organise and oversee participation within a community, with the objective of promoting cooperation and preventing instances of misconduct (Grimmelmann, 2015). There are some key challenges which need to be addressed when it comes to moderating deepfake content.

With regards to identifying deepfakes⁸, some possible indicators could include unnatural facial hair, blinking patterns and unusual shading (MIT Media Lab, 2023). However, with deepfakes getting more sophisticated due to advancement in technology, many times they are unrecognisable to the human eye. In an effort to fight its spread, companies like Facebook and Microsoft, along with academics from top universities,

spearheaded the Deepfake Detection Challenge in 2019. This initiative proved the success of detecting deepfakes through the use of machine learning tools and that its accuracy rate can be improved (Coldewey, 2020). Once identified, automated filtering technologies can be used to prevent their further dissemination. Changes could also be made to the notice and takedown system by updating this procedure to allow individuals to notify the platform regarding unauthorised deepfakes, post which, detection technologies can be used to verify its authenticity (Cruz, 2020).

After identifying deepfakes, platforms have various ways to respond to the same. YouTube, as per its Content Guidelines, bans content which is ‘intended to impersonate a person or channel’ (YouTube Help, 2023a) as well as doctored or manipulated content which misleads the users and causes extreme harm. The author argues that deepfakes would clearly fall under both of these categories. Similarly, through its Community Standards, Facebook also forbids the usage of manipulated media such as videos that have been altered in a manner which is unapparent, which misleads an average person to believe that the depicted individual uttered words which, in reality, they did not utter and are rather a production of artificial intelligence (Meta Transparency Center, 2023). Some general recommendations which the author believes could aid in combating unauthorised deepfakes would be for intermediaries to evaluate the scope of their policies and its application to potential deepfakes by considering whether the terminology used by them is broad enough to encompass its different uses.

Platforms should also clarify to the users the criterion for exceptions (such as, parody) and how a decision to moderate the ban of such content can be appealed; for example, the labelling of content as satire or parody could be made mandatory.⁹ Safeguards should also be developed to prevent the abuse of notice and takedown system, which could lead to wrongful takedowns and resultantly, a multitude of appeals. Imposing penalties for misrepresentation could be a countermeasure to this (Kuczerawy, 2020).

Additionally, while responding to such content, platforms should keep in mind to adhere to due process principles. The Manila Principles could serve as a guide to the same. Users should be provided with a clear and comprehensive notice, explaining the grounds for the decision and highlight the available avenues for appeal. The process for appeal should be transparent, affordable, fair, user-friendly, and timely. If a user has been successful at overturning the moderation decision, appropriate measures should be taken to restore their content within a reasonable time frame (Cruz, 2020). Lastly, since in today’s digital world, platforms are the new governors of speech, they should make all information available to users and work on their measures towards accountability, which will not just serve the interest of their users, but also the democracy at large (Klonick, 2018).

7 Conclusions

Deepfakes are becoming a huge threat to the society, with malicious actors indulging in its harmful uses, such as deepfake pornography, commercial frauds, fake news, and political propaganda. Currently, India lacks a specific law for the regulation of deepfakes. Although existing legal mechanisms, such as criminal, privacy, copyright, and information technology laws can be repurposed to apply to deepfakes, these are insufficient in responding to and curbing its spread. Individuals may file cases in court under these laws, however, it can take a few months, if not years, to render the final

decision. Meanwhile, the unauthorised content is still available for views on the internet. The takedown system is also subject to certain drawbacks, such as the whac-a-mole phenomenon, jurisdictional issues, copyright ownership issues, etc. This highlights the need for a robust mechanism to address these challenges.

With the recent legislative changes in the IT Rules, 2021, there is a shift towards pro-active monitoring and platform responsibility, which serves as a driving force for intermediaries to address the proliferation of unauthorised deepfakes. However, it is unclear how platforms can achieve this. When platforms indulge in moderating content, it affects the freedom of speech and expression and the right to privacy of the users. It is therefore important for them to balance the need for moderation with the need to adhere to fundamental rights of the users as well as the principles of due process. This is a central conflict between perpetrators of harm and intermediaries, which has an effect on internet users in general (Klonick, 2018).

For striking this balance, a comprehensive legal framework would enable platforms to respond to deepfakes as a legal obligation, by building their own content policies. This would alleviate any democratic concerns as the law would clearly define the contours of what is and is not allowed. Apart from this, some legal guidelines for content moderation must also be developed to help platforms uphold due process and measures towards accountability and transparency. However, it is also important for government authorities to do a thorough research into the potential consequences before making any policy changes, in order to alleviate concerns surrounding under blocking and over blocking.

In today's day and age, platforms are the new governors of speech. Hence, it is important for them to balance the interests of the users and democracy, at large. Incorporating these suggestions and proposals would ensure that pre-existing solutions such as right to privacy, intermediary liability, and intellectual property rights such as copyright and passing off can be applied to deepfakes in a better way, while addressing the loopholes regarding copyright ownership, takedown, jurisdiction, and whac-a-mole, amongst others. Having these clear guidelines, policies and legislative changes would ensure that our country is able to respond to the threat of deepfakes in a comprehensive manner.

References

- Ahuja, V.K. (2017) 'Concept of originality', *Law Relating to Intellectual Property Rights*, 3rd ed., p.35, Lexis Nexis, India.
- Ajder, H et al. (2019) *The State of Deepfakes: Landscape, Threats and Impact*, Deepttrace [online] https://regmedia.co.uk/2019/10/08/deepfake_report.pdf (accessed 21 July 2023).
- Batchelor, L. (2020) *Nominet and PIPCU Tackle Cybercrime with Launch of New Landing Pages for Suspended Criminal Domains*, Nominet UK [online] <https://www.nominet.uk/nominet-and-pipcu-tackle-cyber-crime-with-launch-of-new-landing-pages-for-suspended-criminal-domains/> (accessed 21 July 2023).
- Caldwell, M. et al. (2020) 'AI-enabled future crime', *Crime Science*, Vol. 9, No. 1, p.4.
- CBS News (2019) *Doctored Nancy Pelosi Video Highlights Threat of 'Deepfake' Tech* [online] <https://www.cbsnews.com/news/doctored-nancy-pelosi-video-highlights-threat-of-deepfake-tech-2019-05-25/> (accessed 21 July 2023).
- Chesney, R. and Citron, D.K. (2019) 'Deep fakes: a looming challenge for privacy, democracy and national security', *California Law Review*, Vol. 107, No. 1, p.1753.

- Coldewey, D. (2020) *Facebook's 'Deepfake Detection Challenge' Yields Promising Early Results*, Tech Crunch [online] <https://techcrunch.com/2020/06/12/facebook-deepfake-detection-challenge-yields-promising-early-results/#:~:text=Facebook's%20'Deepfake%20Detection%20Challenge'%20yields%20promising%20early%20results,-Devin%20Coldewey%40techcrunch&text=The%20digitally%20face%2Dswapped%20videos,need%20to%20find%20them%20first> (accessed 21 July 2023).
- Cole, S. (2018) *Reddit Just Shut Down the Deepfakes Subreddit*, Vice [online] https://www.vice.com/en_us/article/neqb98/reddit-shuts-down-deepfakes (accessed 21 July 2023).
- Cole, S. (2020) *The Ugly Truth Behind Pornhub's 'Year in Review'*, Vice [online] https://www.vice.com/en_us/article/wxez8y/pornhub-year-in-review-deepfake (accessed 21 July 2023).
- Cruz, S. (2020) *Legal Opinion on Deepfakes and Platform Responsibility*, Digilabs [online] <https://digilabs.global/wp-content/uploads/2021/07/LegalOpiniononDeepfakesandPlatformResponsibility.pdf> (accessed 21 July 2023).
- Dasgupta, B. (2020) *BJP's Deepfake Videos Trigger New Worry Over AI Use in Political Campaigns*, Hindustan Times [online] <https://www.hindustantimes.com/india-news/bjp-s-deepfake-videos-trigger-new-worry-over-ai-use-in-political-campaigns/story-6WPIftMAOaepkwdybm8b1O.html> (accessed 21 July 2023).
- Dave, P. (2020) *Social Media Giants Warn of AI Moderation Errors as Coronavirus Empties Offices*, Thomson Reuters [online] <https://www.reuters.com/article/us-health-coronavirus-google-idUSKBN2133BM> (accessed 25 July 2023).
- Future Advocacy (2019) *Deepfakes – Press Release* [online] <https://futureadvocacy.com/deepfakes-press-release/> (accessed 21 July 2023).
- Garg, R. (2021) *Is Indian Law Equipped Enough to Deal with Deepfakes*, iPleaders [online] <https://blog.iplayers.in/is-the-indian-law-equipped-enough-to-deal-with-deepfakes/> (accessed 21 July 2023).
- Goldman, E. (2019) 'The complicated story of FOSTA and Section 230', *First Amendment Law Review*, Vol. 17, No. 1, p.279.
- Goldman, E. (2020) 'An overview of the United States Section 230 internet immunity', *Oxford Handbook of Online Intermediary Liability*, p.164, Santa Clara University, California.
- Grimmelmann, J. (2015) 'The virtues of moderation', *Yale Journal of Law and Technology*, Vol. 17, No. 1, p.44.
- Isaac, M. (2020) *Facebook Hampers Do-It-Yourself Mask Efforts*, The New York Times [online] <https://www.nytimes.com/2020/04/05/technology/coronavirus-facebook-masks.html> (accessed 21 July 2023).
- Jain, S. and Jha, P. (2020) *Deepfakes in India: Regulation and Privacy*, London School of Economics Blog [online] <https://blogs.lse.ac.uk/southasia/2020/05/21/deepfakes-in-india-regulation-and-privacy/> (accessed 21 July 2023).
- Justice KS Puttaswamy v. Union of India (2017) *AIR, SC 4161*.
- Kabbara, A.R. (2021) *Bots & Deepfakes*, The College of William & Mary, p.4 [online] https://nsiteam.com/social/wp-content/uploads/2021/08/IJO_eIntern-IP_Bots-and-Deepfakes_Kabbara_FINAL.pdf (accessed 21 July 2023).
- Kamlesh Vasvani v. Union of India (2017) *Writ Petition (Civil) No. 177 of 2013*.
- Klonick, K. (2018) 'The new governors: the people, rules and processes governing online speech', *Harvard Law Review*, Vol. 131, No. 1, p.1598.
- Kuczerawy, A. (2020) 'From 'notice and takedown' to 'notice and staydown': risks and safeguards for freedom of expression', *Oxford Handbook of Online Intermediary Liability*, p.530, Oxford University Press, Oxford.
- Laidlaw, E. (2015) *Regulating Speech in Cyberspace: Gatekeepers, Human Rights and Corporate Responsibility*, Cambridge University Press, Cambridge.

- Mendiratta, R. (2021) *Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021*, Stanford WILMap [online] <https://wilmap.stanford.edu/entries/information-technology-intermediary-guidelines-and-digital-media-ethics-code-rules-2021> (accessed 21 July 2023).
- Meta Transparency Center (2023) *Manipulated Media* [online] https://transparency.fb.com/en-gb/policies/community-standards/manipulated-media/?source=https%3A%2F%2Fwww.facebook.com%2Fcommunitystandards%2Fmanipulated_media%2F (accessed 25 July 2023).
- MIT Media Lab (2023) *Detect Deepfakes: How to Counteract Misinformation Created by AI* [online] <https://www.media.mit.edu/projects/detect-fakes/overview/> (accessed 21 July 2023).
- Mostert, F. (2020) 'Digital due process: a need for online justice', *Journal of Intellectual Property Law and Practice*, Vol. 15, No. 5, p.378.
- My Space Inc. v. Super Cassettes Industries Ltd. (2016) 236 DLT 478 (DB).
- Poritz, I. (2023) *States are Rushing to Regulate Deepfakes as AI Goes Mainstream*, Bloomberg [online] <https://www.bloomberg.com/news/articles/2023-06-20/deepfake-porn-political-ads-push-states-to-curb-rampant-ai-use#xj4y7vzkg> (accessed 25 July 2023).
- R. Rajagopal v. State of Tamil Nadu (1995) AIR 264.
- Rana, L. and Sharan, S. (2022) *Protection of Personality and Image Rights in India*, S.S. Rana & Co. [online] <https://ssrana.in/articles/protection-personality-image-rights-india/> (accessed 21 July 2023).
- Satariano, A. and Mozur, P. (2023) *The People Onscreen are Fake. The Disinformation is Real*, New York Times [online] <https://www.nytimes.com/2023/02/07/technology/artificial-intelligence-training-deepfake.html> (accessed 2nd January 2024).
- Simone, D. (2022) *Copyright and Collective Authorship: Locating the Authors of Collaborative Work*, p.15, Cambridge University Press, Cambridge.
- Sisario, B. and Vega, T. (2013) *Playing Whac-a-Mole with Piracy Sites*, New York Times [online] <https://www.nytimes.com/2013/01/29/business/media/playing-whac-a-mole-with-piracy-sites.html> (accessed 21 July 2023).
- Stankiewicz, K. (2019) *Perfectly Real Deepfakes Will Arrive in 6 Months to a Year*, Technology Pioneer Hao Li Says, CNBC [online] <https://www.cnbc.com/2019/09/20/hao-li-perfectly-real-deepfakes-will-arrive-in-6-months-to-a-year.html> (accessed 21 July 2023).
- Stewart, E. (2018) *The Next Big Battle Over Internet Freedom Is Here*, Vox [online] <https://www.vox.com/policy-and-politics/2018/4/23/17237640/fosta-sesta-section-230-internet-freedom> (accessed 21 July 2023).
- The Hard Copy (2021) *Hear Hrithik Calling Out Your Name* [online] <https://thehardcopy.co/personalised-celebrity-messages-using-ai/> (accessed 21 July 2023).
- The Times of India (2023) *Deepfake Porn: Amid AI Race, Another Growing Menace* [online] <https://timesofindia.indiatimes.com/world/rest-of-world/deepfake-porn-amid-ai-race-another-growing-menace/articleshow/99570570.cms> (accessed 21 July 2023).
- Tiwari, S. (2023) *Deepfake Stock Market Scams: How AI is Being Used to Trick Investors*, Business Today [online] <https://www.businesstoday.in/technology/news/story/deepfake-stock-market-scams-how-ai-is-being-used-to-trick-investors-406972-2023-11-24> (accessed 2 January 2024).
- Vicari, C. and Chadwick, A. (2020) 'Deepfakes and disinformation: exploring the impact of synthetic political video on deception, uncertainty and trust in news', *Social Media + Society*, Vol. 6, No. 1, p.2, <https://doi.org/10.1177/2056305120903408>.
- Wakefield, J. (2020) *Facebook's Supreme Court Members Announced*, BBC News [online] <https://www.bbc.com/news/technology-52558559> (accessed 21 July 2023).
- Westerlund, M. (2019) 'The emergence of deepfake technology: a review', *Technology Innovation Management Review*, Vol. 9, No. 11, p.39.
- YouTube Help (2023a) *Impersonation Policy* [online] https://support.google.com/youtube/answer/2801947?hl=en&ref_topic=9282365 (accessed 21 July 2023).
- YouTube Help (2023b) *Spam, Deceptive Practices & Scams Policies* [online] <https://support.google.com/youtube/answer/2801973?hl=en> (accessed 21 July 2023).

Notes

- 1 For the fluency of the argument here, I am assuming that these pre-existing works are copyrighted works.
- 2 This has been discussed in Part 4 of the paper.
- 3 This would only be possible in cases where the victim has a copyright on the content. See Subsection 3.4 for further discussion.
- 4 This has been discussed in detail in Section 5.
- 5 In the case of *Dharamraj Bhanushankar Dave v. WhatsApp Inc.* (2018), the Delhi High Court ruled that WhatsApp could not be held responsible for dissemination of fake news or rumours through its platform, under the old IT Rules, 2011.
- 6 Held in *Kamlesh Vasvani v. Union of India*, Writ Petition (Civil) No. 177 of 2013.
- 7 My proposal does not advocate for a complete prohibition of all uses of deepfakes, but specifically the unauthorised and harmful uses of deepfakes as discussed under Section 3, which are a menace to the society.
- 8 Since I am not a technological expert, I have not delved deep into providing solutions on how this can be achieved.
- 9 Parody and satire are an exception to this rule, as per the Facebook Community Standards.