DEEPFAKES AND SHALLOW LAWS: REGULATING DISTORTED NARRATIVES IN THE POLITICAL CYBERSPACE

Shimona Mohan[*] and Sarthak Wadhwa[**]

Abstract: *With a sizable section of the global population witnessing electoral churn, protests and resistance movements in the past couple years, this is clearly a watershed moment for modern democracies vis-a-vis technological advancements. Electoral canvassing in these democracies has expanded their footprint in the digital space, be it through social media or navigating the general contours of anonymity, engagement and misinformation that govern our online worlds. In this context, the advent of deepfakes in the recent past, and the exponential evolution of the technology both for outreach and disruption, has invited acute regulatory attention and intervention. However, traditional policy design has proven to be inadequate to respond to such a novel, ever-changing problem - necessitating a more contemporaneous reimagination of technological regulation.*

In this paper, we endeavour to advance one such perspective on the regulation of deepfakes. First, we discuss the extant industry and regulatory solutions (or lack thereof) that have emerged to tackle the proliferation of altered digital media and deepfakes around the world. Next, we assess proposed legislations to combat deepfakes and identify the pitfalls of a pure regulatory solution in this space in India and beyond. Finally, we attempt to reframe deepfakes as a

[*] Shimona Mohan is an Associate Researcher on Gender & Disarmament and Security & Technology at the United Nations Institute for Disarmament Research (UNIDIR) in Geneva, Switzerland. She is also one of the 100 Brilliant Women in AI Ethics for 2024. Her areas of focus include the multifarious intersections of security, emerging technologies (in particular AI and cybersecurity), gender and disarmament. Shimona's previous work includes stints at the Observer Research Foundation (ORF), the Think20 Secretariat under India's G20 presidency in 2023, the Carnegie Endowment for International Peace, and the Ministry of External Affairs as well as NITI Aayog under the Government of India.

[**] Sarthak Wadhwa is a lawyer practicing in New Delhi, India. He holds a B. A., LL. B. (Hons.) from the National Law School of India University, Bengaluru.

The authors would like to thank Shikhar Sharma and Debditya Saha for their diligent assistance in reviewing and editing this piece. All opinions expressed and errors herein are the authors' own, and not attributable to any other organization the affiliated with.

communication-governance problem as opposed to a platform-regulation problem, to advance a hybrid co-regulatory model to addresses deepfakes in India.

I. INTRODUCTION

The advent of advanced applications of artificial intelligence (AI), machine learning and neural networks has resulted in a mix of interesting, revolutionary and disruptive effects. An example that covers this spectrum in its entirety is the widespread development and deployment of 'deepfakes.' The term first came to be in 2017 on Reddit, and initially referred to explicit videos that were shared on the platform using open-source face-swapping technology.[1] In the recent past, deepfakes have expanded in their scope to refer to any digital media that has been created or altered using AI and related technologies in such a way that they superimpose the visual and/or auditory likeness of one person or thing over another.[2]

In the past few years, the cyberspace has seen a drastic proliferation of deepfakes – from innocuous face-swapping applications[3] and somewhat morally questionable websites that animate pictures of deceased relatives,[4] to doctored TikTok videos of celebrities playing golf[5] and seemingly convincing but completely virtual social media influencers,[6] to the much more

[1] Meredith Somersn 'Deepfakes, Explained' (*MIT Management Sloan School*, 21 July 2020) <https://mitsloan.mit.edu/ideas-made-to-matter/deepfakes-explained> accessed 27 July 2024.
[2] Vejay Lalla, Adine Mitrani, Zach Harned, 'Artificial intelligence: deepfakes in the entertainment industry' (*WIPO*, June 2022) <www.wipo.int/wipo_magazine/en/2022/02/article_0003.html> accessed 27 July 2024.
[3] Joe Taylor, '10 Best Face Swap Apps for iPhone and Android in 2023' (*Expert Photography*, 2023) <expertphotography.com/face-swap-apps/> accessed 27 July 2024.
[4] MyHeritage, 'Deep Nostalgia: Animate Your Family Photos' (*MyHeritage*, 2023) <www.myheritage.com/deep-nostalgia> accessed 27 July 2024.
[5] Lee Brown, 'Deepfake' Tom Cruise goes viral on TikTok with over 11 million views' (*New York Post*, 2 March 2021) <nypost.com/2021/03/02/deepfake-tom-cruise-goes-viral-on-tiktok-with-over-11m-views/ > accessed 27 July 2024.
[6] Nader Luthera, 'The Dark Side Of Deepfake Artificial Intelligence And Virtual Influencers' (*Forbes*, 16 January 2020) <www.forbes.com/sites/forbesbusinesscouncil/2020/01/16/the-dark-side-of-deepfake-artificial-intelligence-and-virtual-influencers/> accessed 27 July 2024.

disturbing use of deepfakes in revenge porn[7] and public announcement videos of political figures relaying statements which they never actually made.[8] Apart from potentially being affronts to the personal dignity of real individuals, deepfakes have created a cyber environment of distorted narratives emerging from the inauthenticity of public domain content, or the suspicion thereof.

The latter is inherently detrimental given its ability to spur untruthful factoids, generate conspiracy theories and spark antagonistic actions. However, deepfakes become even more potent seeds of disinformation during political crises and conflict situations when tensions and the political desire to alter public opinion is high, and the prospects of the delivery of genuine information are low or slow. This could come in the form of deepfakes of political leaders encouraging or discouraging certain public actions seemingly against their interest, falsified video feeds of the conflict arena, or deepfake video claims of the initiation of the conflict which could overturn the blame of the conflict or its escalation to the other side, among others. Given that the cyberspace has become the *de facto* fifth domain[9] of warfare, all of these actions and more have increasingly been observed in conflicts across the world.[10] The use of deepfakes is not restricted to a certain governance system or political inclination anymore, and has been known to set off ripple effects of magnitudes beyond the virtual realm. This was especially highlighted during the Russian invasion of Ukraine in 2022, wherein the extensive strategic use of deepfakes, amongst other cyber operations, brought to the fore how the dynamics of modern war and conflict have effectively transformed to include an online element apart from the kinetic.

As with most other use cases of emerging technologies, the technology outpaces the law in the case of deepfakes, and could have disastrous effects if the disingenuity of online audio-visual material remains unchecked and unregulated. This is increasingly becoming apparent with the rapid development around generative AI (gen-AI) and the resulting cascade of

---

[7] Jessica White, 'Inside the disturbing rise of 'deepfake' porn" (*Dazed*, 19 April 2022) <www.dazeddigital.com/science-tech/article/55926/1/inside-the-disturbing-rise-of-deepfake-porn> accessed 27 July 2024.

[8] William Galston, "Is seeing still believing? The deepfake challenge to truth in politics' (*Brookings*, 8 January 2020) <www.brookings.edu/research/is-seeing-still-believing-the-deepfake-challenge-to-truth-in-politics/> accessed 27 July 2024.

[9] In war and security studies, the cyberspace emerged as (and is now commonly referred to as) the fifth domain of warfare following the traditional four – land, sea, air, and space – in the early 2000s. See more: Larry D. Welch, Institute for Defence Analyses Alexandria, 'Cyberspace – The Fifth Operational Domain' (*Defence Technical Information Centre*, Technical Report No. AD1124078, 01 November 2004) <https://apps.dtic.mil/sti/citations/AD1124078> accessed 27 July 2024.

[10] Daniel L. Byman, Chongyang Gao, Chris Meserole, and V.S. Subrahmanian, 'Deepfakes and international conflict' (*Brookings*, January 2023) <www.brookings.edu/research/deepfakes-and-international-conflict/> accessed 27 July 2024.

machine-generated material online in the past year.[11] It is now dangerously easy for anyone with access to the internet to create realistic audio and visuals to suit their agenda and support their claims, which makes disinformation in the cyberspace much more accessible, affordable and convenient to generate and disseminate.

This paper first explains deepfakes as a new disruptive technology and situates the distorted narratives they generate in the political cyberspace, especially as they relate to turbulent geopolitics and conflict situations, through recent examples of their misuse. It then explores the potential to formulate optimum regulatory responses to deepfakes by drawing on existing or forthcoming techno-legal frameworks along similar issue domains. The paper concludes with a future-faced overview of new challenges and regulations for deepfakes and aims to spark further research offshoots in this area.

II. A DEEP DIVE INTO DEEPFAKES

While deepfakes are a seemingly new phenomena which have gained ground (and simultaneously, notoriety) over the past couple of years, the technology behind them is almost a decade old.[12] Deepfakes run on AI models that use deep neural networks (DNNs) in the form of two competing algorithms – a 'generator' that extracts data from face-maps in the form of facial features from a source input which can be mapped on to a target image; and a 'discriminator' that extracts data from the processed image and compares it back to the source input. The two models together constitute a Generative Adversarial Network (GAN), which in turn creates a 'deepfake' when sufficient iterations of the generation and discrimination processes yield a media output that is indistinguishable from the original media used, thus appearing entirely new.[13]

Despite the complexity of their operation and the final result, unsophisticated GANs may generate uncanny or unnatural deepfakes of people or things. Such deepfakes are discernible from genuine content due to certain tells, such as awkward positioning, colouration, movement, alignment, or other mechanical failures such as non-synchronization of events in a deepfake video with its associated audio, less blinking of the eyes or movement in

[11] Rohitha Naraharisetty, 'Generative AI Puts Us on the Brink of a Deepfakes Crisis' (*The Swaddle,* 12 Decemeber 2022) <theswaddle.com/generative-ai-puts-us-on-the-brink-of-a-deepfakes-crisis/> accessed 27 July 2024.
[12] Ian J Goodfellow, *et al.* 'Generative Adversarial Networks' (*Arxiv*, 10 June 2014) <arxiv.org/pdf/1406.2661> accessed 27 July 2024.
[13] *Ibid*

prolonged shots, or inconsistencies upon zooming/slowing down the videos.[14] While deepfakes emerging from unsophisticated GANs can be detected by simply looking at them, GANs have evolved considerably and many are now able to overcome these faults, leading to much higher quality deepfakes and hyper-realistic effects that are visually indistinguishable from authentic content. Recent studies show that most people are only able to identify political deepfakes with about 50-80% accuracy depending upon the modality of the content.[15]

The ubiquity of deepfakes has gradually created an online environment where netizens are slowly becoming more suspicious of the validity and/or veracity of material, regardless of the format it is presented to them in.[16] This is exacerbated by the advent of gen-AI, which makes it possible to not just overlay images and videos to produce deepfakes, but to create entirely new audio-visual media out of mere text commands, and have the entities in this machine-generated media say or do whatever the creator wants at the input of a few simple prompts.[17]

While there is an industry around detecting and protecting against deepfakes,[18] or using them solely for constructive purposes,[19] it is becoming increasingly difficult to moderate machine-generated content online (malicious or not) as the world becomes increasingly more digitally reliant and access to tools and services becomes exponentially easier. The perpetual problem for regulators has now become the need to regulate the ever-evolving generation and dissemination of deepfakes in a manner that is not disruptive to the online space, politically or otherwise.

---

[14] Detect Deepfakes Project, 'Detect DeepFakes: How to counteract misinformation created by AI' (*MIT Media Lab*) <www.media.mit.edu/projects/detect-fakes/overview/> accessed 27 July 2024..

[15] Matther Groh, *et al*, 'Human Detection of Political Deepfakes Across Transcripts, Audio and Video' (*Arxiv*, 14 September 2022) <arxiv.org/pdf/2202.12883.pdf> accessed 27 July 2024.

[16] Galston, (n 8); Simon Kuper, 'The age of scepticism: from distrust to 'deepfake'' (*Financial Times*, 18 October 2018) <www.ft.com/content/2fc9c1fa-d1a2-11e8-a9f2-7574db66bcd5> accessed 27 July 2024; Joanne Henry, 'Disinformation and Deepfakes Fuel Growing Mistrust' (*Catalyst: International Association of Business Communicators*, 13 April 2021) <catalyst.iabc.com/Articles/disinformation-and-deepfakes-fuel-growing-mistrust> accessed 27 July 2024; Ville Manninen, 'Deepfakes do not fool people, but still undermine trust in news' (*Journalism Research News*, 2 March 2020) <journalismresearchnews.org/deepfakes-do-not-fool-people-but-still-undermine-trust-in-news/> accessed 27 July 2024.

[17] Alex Pasternack, 'GPT-powered deepfakes are a powder keg' (*Fast Company*, 22 February 2023) <www.fastcompany.com/90853542/deepfakes-getting-smarter-thanks-to-gpt> accessed 27 July 2024.

[18] Thanh Thi Nguyen, *et al.*, 'Deep Learning for Deepfakes Creation and Detection: A Survey' (*Arxiv*, 11 August 2022) <arxiv.org/pdf/1909.11573.pdf> accessed 27 July 2024.; Q5id, "6 Real Life Deepfake Examples" (*Q5id*, 9 November 2022) <q5id.com/blog/6-real-life-deepfake-examples> accessed 27 July 2024.

[19] Alex Wilkins, 'AI-generated deepfake faces could help protect privacy on social media' (*NewScientist*, 18 November 2022) <www.newscientist.com/article/2347596-ai-generated-deepfake-faces-could-help-protect-privacy-on-social-media/> accessed 27 July 2024.

III. DEEPFAKES AND DISTORTED NARRATIVES

One of the major concerns that abound deepfakes is their capacity to create a disinformation pandemic when it comes to the political cyberspace.[20] This is especially relevant given that the occurrence of higher political activity – such as elections, terrorist attacks, or internal or international conflicts – anywhere in the world is now being coupled with a proportionate influx of information and media in online spaces and on social media platforms.[21] The spread of this material is usually swift and wide thanks to the viral nature of audio-visual content online, vested political interests of the people who engage with it, and general cognitive ability of the sharers.[22] Thus, a moderation of deepfaked content often (if ever) occurs only after a sizable number of people have already interacted with it.

Politically motivated deepfakes have often been observed in the recent past, and have set off disinformation chains and socio-political instability which remains in the public domain even if the deepfake itself is either removed, debunked or fades from its viral status. For a case in point, armed soldiers from the elite Republican Guard in Gabon seized the national radio station to rouse a failed coup against the incumbent government in 2019. The soldiers did so due to their suspicion of President Ali Bongo Ondimba being dead, given his lack of public appearances and rumours about his video for the 2018-19 New Year's address having been a deepfake. His face looked odd in the video, his eyelids were unequally open, and his facial muscles didn't move – all of which were revealed to be outcomes of a stroke that he had suffered, when he was announced to indeed be alive later on.[23]

Political deepfakes (or claims thereof) have also often been used in an attempt to tarnish the reputation of those in the limelight. In 2021, former Indian Union Minister Sadananda Gowda was seen in a video soliciting a woman over a video call, which he denounced as a malicious deepfake and filed a complaint for with the cyber-crime police in Bengaluru. He also obtained an injunction restraining the publication and circulation of the video and any

[20] Charlotte Klein, '"This Will Be Dangerous in Elections": Political Media's Next Big Challenge Is Navigating AI Deepfakes' (*Vanity Fair*, 6 March 2023) <www.vanityfair.com/news/2023/03/ai-2024-deepfake> accessed 27 July 2024.
[21] Galston (n8).
[22] Saifuddin Ahmed, 'Who inadvertently shares deepfakes? Analyzing the role of political interest, cognitive ability, and social network size' (2021) 57 *Telematics and Informatics* 101508 < www.sciencedirect.com/science/article/abs/pii/S0736585320301672> accessed 27 July 2024.
[23] Sarah Calhan, 'How misinformation helped spark an attempted coup in Gabon' (*The Washington Post*, 13 February 2020) <www.washingtonpost.com/politics/2020/02/13/how-sick-president-suspect-video-helped-sparked-an-attempted-coup-gabon/> accessed 27 July 2024.

reportage on it in the media.[24] Similar instances have been known to pick up pace and popularity especially during elections. For instance, in 2019, a manipulated video of Nancy Pelosi floated across social media, which made her appear intoxicated and slurring her words. The video, which was viewed more than 2.5 million times on Facebook in a matter of days and shared by prominent political leaders,[25] was ultimately removed from Youtube for violating its standards, while Facebook flagged it as fake and curtailed its reach, and X (formerly Twitter) refrained from taking any action.[26]

With cyberspace becoming the most recent battleground, it is also not surprising that deepfakes have been used as a strategic ploy to spread propaganda against an adversary. This was seen in Myanmar in March 2021, following a coup that transferred power to a military junta. The former chief minister of the Yangon region Phyo Min Thein appeared on a national television channel operated by the Burmese military confessing to offering bribes to ousted leader Aung San Suu Kyi. There were discrepancies in his voice and lip-syncing, leading to speculations about the video being a deepfake. However, since the video quality was too low for any conclusive forensic analysis to be done, the possibility of the video being genuine helped the military's narrative of rampant corruption in the Aung San Suu Kyi government which they claimed had prompted their coup.[27]

This practice has seen new heights in light of recent conflicts, and especially gained an international characteristic during the Russian invasion of Ukraine. In March 2022, a video of the Ukrainian President Volodymyr Zelensky showing him surrendering to the Russian invasion and calling the citizens to lay down arms and return to their homes appeared on social media. Aside from his message being against all of his previous public comments, the video was quickly called out as a deepfake since only his head seemed to be moving in the video while with the rest of the body was uncannily static.[28] The video was later put on a

[24] Akram Mihammed, ''Deep fake' video, says DV Sadananda Gowda; files complaint' (*Deccan Herald*, 19 September 2021) <www.deccanherald.com/state/deep-fake-says-dv-sadananda-gowda-on-sleaze-video-files-complaint-1032140.html> accessed 27 July 2024.

[25] Sarah Mervosh, 'Distorted Videos of Nancy Pelosi Spread on Facebook and Twitter, Helped by Trump' (*The New York Times*, 24 May 2019) <www.nytimes.com/2019/05/24/us/politics/pelosi-doctored-video.html> accessed 27 July 2024.

[26] Makena Kelly, 'Distorted Nancy Pelosi videos show platforms aren't ready to fight dirty campaign tricks: YouTube removed the video, Facebook de-ranked it, and Twitter let it stand' (*The Verge*, 24 May 2019) <www.theverge.com/2019/5/24/18637771/nancy-pelosi-congress-deepfake-video-facebook-twitter-youtube> accessed 27 July 2024.

[27] KrASIA, 'Did Myanmar's military deepfake a minister's corruption confession?' (*KrASIA*, 24 March 2021) <kr-asia.com/did-myanmars-military-deepfake-a-ministers-corruption-confession> accessed 27 July 2024.

[28] Jack Newman, 'It's a cheapfake! Experts laugh off Kremlin misinformation attempt as amateurish 'deepfake' video of Zelensky 'surrendering' is posted by hackers – and spotted almost immediately' (*Mail Online*, 18 March

Ukrainian news agency's website by hackers[29] and received a lot of social media attention[30] which died down when the President clarified the situation.[31]

On the other hand, similar deepfake videos emerged from the Ukrainian side as well, often well attributed to Ukraine-oriented sources. One portrayed Russian President Vladimir Putin announcing peace,[32] and was shared on X with a Russian caption asking Russian soldiers to leave while they were still alive.[33] The original poster of the video admitted to the video being inauthentic and satirical.[34] In another example, Ukrainian officials shared a deepfake video of Paris being attacked by Russian airstrikes,[35] captioning it with the idea that Ukraine is the first line of defence against Russian aggression and that the rest of Europe will fall if Ukrainian resistance is not supported.[36] The video clarifies that the imagery depicted is merely to raise awareness about the stakes of the Russian invasion, and it has been confirmed that it was created by French filmmakers.[37]

The political motivation for the use of deepfakes as an information tactic is now increasingly being observed, and has been amplified by the greater ease, capacity and possibilities emerging from combining deepfakes with gen-AI tools and services. China made the news in 2022 with its life-like AI influencers, which were created using gen-AI applications like Synthesia and promoted the interests of the Chinese Communist Party by commenting on topics such as the importance of China-U.S. cooperation for the global economy's recovery

---

2022) <www.dailymail.co.uk/news/article-10625935/Experts-laugh-Kremlins-amateurish-deepfake-video-Zelensky-surrendering.html> accessed 27 July 2024.

[29] Zachary Snowdon Smith, ''Hacked' Ukrainian TV Station Transmits Fake Zelensky Surrender Announcement' (*Forbes*,16 March 2022) <www.forbes.com/sites/zacharysmith/2022/03/16/hacked-ukrainian-tv-station-transmits-fake-zelensky-surrender-announcement/?sh=247b5ae5528d> accessed 27 July 2024.

[30] Kieran Press-Reynolds, 'Zelenskyy denies deepfake video of him surrendering after hackers broadcast it on Ukrainian TV website' (*Insider*, 17 March 2022) <www.insider.com/volodymyr-zelenskyy-deepfake-video-surrendering-ukraine-russia-forces-2022-3> accessed 27 July 2024.

[31] Володимир Зеленський (@zelenskiy_official), 'We are at home and protect[ing] Ukraine [*translated from Ukrainian*]' (*Instagram*, 16 March 2022) <www.instagram.com/p/CbKYM_Zg2jo/> accessed 27 July 2024.

[32] The Kremlin, 'Address by the President of the Russian Federation' (*President of Russia*, 21 February 2022) <kremlin.ru/events/president/news/67828> accessed 27 July 2024.

[33] Serhii Sternenko (@sternenko), 'Президент РФ обьявил о капитуляции россии. Русский солдат, бросай оружие и иди домой, пока жив! [*The President of the Russian Federation announced the surrender of Russia. Russian soldier, drop your weapons and go home while you're alive!* (translated from Russian)]' (*Twitter*, 16 March 2022 7:13pm) <twitter.com/sternenko/status/1504090918994993160?> accessed 27 July 2024.

[34] Reuters Fact Check, 'Fact Check-Doctored video appears to show Putin announcing peace' (*Reuters*, 17 March 2022) <www.reuters.com/article/factcheck-putin-address-idUSL2N2VK1CC> accessed 27 July 2024.

[35] Kate Gill, 'Deep fake capturing Paris being hit by airstrikes shared by Ukrainian minister' (*Independent*, March 2022) <www.independent.co.uk/tv/news/ukraine-paris-bomb-deep-fake-b2034714.html> accessed 27 July 2024.

[36] Daisy Stephens, ''If we fall, you fall': Fake video shows Paris attacked in chilling warning from Ukraine' (*LBC*, 12 March 2022) <www.independent.co.uk/tv/news/ukraine-paris-bomb-deep-fake-b2034714.html> accessed 27 July 2024.

[37] Ibid.

from COVID-19.[38] In early 2023, the same entities behind this, widely considered to be Chinese state-backed parties, put out a series of videos featuring a number of AI-generated people with American accents voicing support for a military-backed coup in Burkina Faso.[39] Although they have since been suspended from using Synthesia by the company itself, there is no dearth of affordable and easy-to-use gen-AI tools that could instead be alternatives.

IV. REGULATING DEEPFAKES

While deepfakes and disinformation were already on the radar of a few regulatory frameworks around the world, recent advancements in AI and gen-AI as well as the observed effects of using deepfakes in politically strained situations should give lawmakers a push to create effective rules governing the creation and use of deepfakes as well. However, very few regulatory frameworks reflect this in a substantial capacity or are currently being developed to do so, often being outrun by the whiplash pace of technological developments around AI. Simultaneously, the industry and related stakeholders are taking actions of their own to attempt to contain the escalating deepfake threat.

A. Industry Action

While countries trudge through long policy processes to establish some semblance of regulations for deepfakes, relevant actors within the industry have come up with their own efforts to counteract deepfakes through detection systems, market standards, and industry-academia coalitions.

The dissemination of deepfakes in political contexts has often seen social media being a critical factor in their viral status. For instance, in May 2023, a deepfake image surfaced on social media, predominantly X, appearing to show a large explosion near the Pentagon building in the US, which led to a brief dip in the stock market.[40] The confusion on X was further amplified by several verified accounts sharing the fake picture. In the aftermath of this incident, X announced a pilot feature that lets users fact-check and add information about the

[38] Adam Satariano, Paul Mozur, 'The People Onscreen Are Fake. The Disinformation Is Real' (*The New York Times*, 07 February 2023) <www.nytimes.com/2023/02/07/technology/artificial-intelligence-training-deepfake.html> accessed 27 July 2024.
[39] Pasternack (n 17)
[40] Mohammed Hadad, 'Fake Pentagon explosion photo goes viral: How to spot an AI image' (*Al Jazeera*, 23 May 2023) <www.aljazeera.com/news/2023/5/23/fake-pentagon-explosion-photo-goes-viral-how-to-spot-an-ai-image> accessed 27 July 2024.

veracity of an image in its community notes, which will then appear below matching images throughout the platform.[41]

Other social media platforms have previously launched initiatives of their own, for instance, Facebook teamed up with a dozen other partners from the tech industry and academic institutions to launch a deepfake detection challenge in 2019.[42] Given that social media giants like X and Meta are not yet required by law to regulate deepfakes,[43] their methods to counter the same have mostly been fragmented knee-jerk reactions and/or passive research projects whose results, implementation, monitoring and evaluation remain murky.

A more consolidated and purposeful alternative to this has come in the form of technical standards by a coalition of industry partners. A good example of this is the specification for digital provenance[44] released by the Coalition for Content Provenance and Authenticity (C2PA), which consists of Adobe, Microsoft, Arm, Intel TruePic and the BBC. The C2PA has previously stated that their standard allows content creators and editors to create media that can't secretly be tampered with, and they are also able to disclose information about who has created or changed digital content as well as how it has been altered.[45]

Apart from efforts undertaken solely by tech organisations and related stakeholders, and due to the unique security challenges politically-inclined deepfakes in general may present, organisations like the Defense Advanced Research Projects Agency (DARPA) in the US have been working on deepfake detection and verification as well. In 2016, DARPA's Media Forensics program began work around developing detection systems that can spot photo and video manipulations and provide a truth measure of visual media on the internet.[46] In 2021,

[41] James Farrell, 'Twitter to expand crowd-sourced fact-checking tool after Pentagon explosion deep fake' (*Silicon Angle*, 30 May 2023) <siliconangle.com/2023/05/30/twitter-expand-crowd-sourced-fact-checking-tool-pentagon-explosion-deep-fake/> accessed 27 July 2024.

[42] Kyle Wiggers, 'Facebook, Microsoft, and others launch Deepfake Detection Challenge' (*Venture Beat*, 11 December 2019) <venturebeat.com/ai/facebook-microsoft-and-others-launch-deepfake-detection-challenge/> accessed 27 July 2024.

[43] Matt O'Brien, 'Meta and X questioned by lawmakers over lack of rules against AI-generated political deepfakes' (*AP News*, 6 October 2023) <https://apnews.com/article/election-deepfakes-ai-x-twitter-facebook-meta-instagram-d52e8703a9e47936061bf2c8bbc94bb5> accessed 27 July 2024.

[44] Coalition for Content Provenance and Authenticity, 'C2PA Specifications' (*C2PA*) <c2pa.org/specifications/specifications/1.3/index.html> accessed 27 July 2024.

[45] Emma Woollacott, 'New Standard Aims To Protect Against Deepfakes' (*Forbes*, 27 January 2022) <www.forbes.com/sites/emmawoollacott/2022/01/27/new-standard-aims-to-protect-against-deepfakes/?> accessed 27 July 2024.

[46] Charlotte Walsh, 'What is a deepfake? This video technology is spooking some politicians' (*USA Today*, 15 March 2019) <www.usatoday.com/story/news/politics/2019/03/15/what-deepfake-video-technology-spooking-some-politicians/3109263002/> accessed 27 July 2024.

DARPA launched dedicated algorithms to this effect, which are supplemented by a quantitative measure of integrity that enables filtering and prioritization of media at scale.[47]

While it remains to be seen how efficiently industry-driven innovations to detect and stop the spread of deepfakes will play out, these currently seem to be the fastest and most direct methods to do so as legal systems still struggle to catch up with the technology and its application.

B. Laws and Regulations

i. Regulatory Gaps: The Indian Experience

Closer home in India, deepfakes are not yet seen as a regulatory concern warranting their own specific laws, despite the multitudes of examples in the public domain about their misuse in political and conflict environments and the rapid expansion of new technologies that will further exacerbate their exclusive harms. Nonetheless, despite no explicit indication from the State evincing any intention to specifically regulate deepfakes, the Indian government is acutely aware of the problems they pose in the political space, and have recently executed knee-jerk reactions to them. For instance, in preparation for the 2024 Lok Sabha elections, the Indian government considered a law that would require WhatsApp to share details about the first originator of a message in order to combat the spread of fake videos on the platform. The basis for this are multiple deepfake videos of politicians circulating on WhatsApp, and the government was in the process of sending an order to the messaging company under the Information Technology (IT) Rules, 2021, seeking the identity of the people who first shared the videos on the platform. Evidently, the State harbours some anxiety about the damage that can be wrought using deepfakes. However, the arsenal of legal options available to the State address deepfakes remain non-focal.

While Indian regulation does not cover deepfakes as a separate entity, certain aspects of the use of deepfakes can be broadly addressed through civil action. For instance, an intuitive approach could be to deploy copyright law actions against the very datasets on which deepfakes are trained and an infringement suit can be brought for the violation of exclusive rights vesting in the constituent works or the distribution of copies.[48] Civil and criminal

[47] Sarah Sybert, 'DARPA Launches New Programs to Detect Falsified Media' (*GovCIO Media*, 16 September 2021) <governmentciomedia.com/darpa-launches-new-programs-detect-falsified-media> accessed 27 July 2024.
[48] *See*: Copyright Act, 1957, s 51 – *insofar as the data sets on which the GAN is trained are comprised of copyrighted content, the unauthorized/unlicensed use thereof for trade/profit amounts to copyright infringement, and is not excused under Section 52 of the Copyright Act, 1957 (fair dealing, research, education, etc.).*

remedies may be available under the Copyright Act, especially when deepfakes are produced in this manner for commercial purposes.[49] However, it may very easily be the case that the aggrieved individual does not hold copyright over all the works featuring their likeness which may have been used to create the deepfake.[50] As such, the aggrieved individual may not have the locus to bring an infringement suit against the originator of the deepfake – assuming that such originator can be identified in the first place. Copyright law does have some flexibility to provide a novel solution to the deepfake problem, but it is not nearly comprehensive enough to provide exhaustive relief to aggrieved individuals.[51] Other civil actions in tort law such as defamation and invasion of privacy – which are themselves underdeveloped in the Indian legal tradition[52] – fall prey to issues such as the inability to pointedly identify the perpetrator of the deepfake in question, the nebulous quantification of reputational and mental harms arising from such perpetuation, and the limited remedies available in civil law to litigate, adjudicate and execute such suits in a time and cost efficient manner.

On the criminal law front, charges such as defamation, forgery, and criminal intimidation[53] may be attracted if law enforcement is able to identify and apprehend the originator of a deepfake. However, the secondary victimization suffered by the aggrieved person during the process of criminal investigation and prosecution,[54] as well as the lack of individual remedies could prove to be a disincentive against any recourse in criminal law. Specific cyber-crime provisions punishing the publication/transmission of sexually explicit content,[55] identity theft/impersonation,[56] and other threats to State sovereignty based on the specific application of the deepfake[57] could be relied upon to enhance the punishments meted out to the perpetrators of deepfakes (as applicable) but the State leads the thrust of such criminal prosecution and aggrieved individuals may still not receive due recompense for harms suffered. In any case, these are *post facto* solutions that may only come to the fore after the

[49] Copyright Act, 1975, ss 55, 63.
[50] A O'Connell, K Bakina, "Using IP Rights to Protect Human Rights: Copyright for 'Revenge Porn' Removal" (2020) 40(3) *Legal Studies* 442, 457.
[51] Sarthak Wadhwa, "Fake it Till You Make it: Finding a Solution to the Commercial Use of Deepfakes in Copyright Law" (2024) (on file with author).
[52] Manjeri Subin Sunder Raj, Ujal Kumar Mookherjee, and Aman Deep Borthakur, 'Tort Law in India' in Mauro Bussani and Anthony J. Sebok (eds.), *Comparative Tort Law: Global Perspectives* (Elgar, 2021) ch 20.
[53] Indian Penal Code, 1860, ss 463, 468, 469, 471 (forgery amounting to causing injury, cheating, damage to reputation and dishonest use).
[54] Amit Anand Choudhary, 'Process is the punishment in our criminal justice system: CJI' (*Deccan Herald*, 17 July 2022) <timesofindia.indiatimes.com/articleshow/92928964.cms?> accessed 27 July 2024.
[55] Information Technology Act, 2000, ss 67, 67A, 67B; *See also*: Indian Penal Code, 1860, s 354C.
[56] Information Technology Act, 2000, ss 66C, 66D.
[57] Information Technology Act, 2000, s 69A.

damage is done, and proactive regulation of deepfakes – and the deeper individual harms associated therewith – remains unattainable.

With the current spate of laws though, not much can be effectively done currently to protect either direct victims or the larger society from deepfakes in India. Individual legal action lies only against adjacent but ancillary issues such as disinformation, electoral misconduct, computer offences, sexual offences, etc. (as discussed above). However, unlike disinformation frameworks, the synthesis of deepfakes for political purposes comes with specific exclusive harms for the individual victims. Deepfakes of individuals are undoubtedly unjust as per the spirit of law, since deepfake-generation softwares use the likeness(es) of the person without their consent. All other use case-specific implications aside, this is an important facet of the breach of an individual's personal data.

Ideally, this should have been covered under the recent Digital Personal Data Protection (DPDP) Act of India, passed in August 2023.[58] Founded on the notions of individual 'dignity' and 'autonomy',[59] the Right to Privacy in India has been operationalized through the DPDP Act to confer and safeguard individual privacy against private actors. However, the Act does not make any mention of deepfakes or related technologies and applications, and victims thus have no substantive options for recourse – either against perpetrators directly or the intermediaries through which deepfakes may be shared – apart from relying on tangentially related provisions under other laws, which are unlikely to be sufficient. A substantive conversation around developing holistic regulations specific to machine-generated material is the need of the hour for the Indian legal space, and the foundation for this would have to be an acknowledgement of the deepfake challenge and possible remedial and/or punitive recourses to it in the DPDP Act. For this purpose, deepfakes must be comprehensively theorized to be aberrations on personal dignity and autonomy and brought within the purview of the Right to Privacy. In this sense, the silence of the DPDP Act on the increasingly concerning proliferation of deepfakes appears to be a lost opportunity to advance Indian law into a novel and fruitful direction.

On a separate note, nascent and politically motivated wide-scale solutions proposed by the State do not inspire much confidence at present. For instance, in 2023, the Ministry of Electronics and Information Technology (MeitY) of India amended the Information

[58]Ministry of Law and Justice (Legislative Department), Digital Personal Data Protection Act, 2023 [No. 22 of 2023] (11 August 2023) <www.meity.gov.in/writereaddata/files/Digital%20Personal%20Data%20Protection%20Act%202023.pdf> accessed 27 July 2024..

[59] *(Retd.) Justice KS Puttaswamy v Union of India*, (2017) 10 SCC 1.

Technology Rules (Intermediary Guidelines and Digital Media Ethics Code) to account for fake news online, and allowed for a fact check unit to monitor the same.[60] While the amendment does not explicitly refer to deepfakes and machine-generated visual media, legal communities are already pointing out the potential for gross violations of fundamental freedoms that may result out of this amendment based on the unbridled power that the fact check unit holds.[61] If the same unit was to be employed for reining in political deepfakes, the overseeing body would especially need to be politically neutral and technologically astute to prevent the scope of biased rulings. This concern, amongst others around the potential clash of the IT rules with freedom of speech, was specifically raised by media personalities like political satirist Kunal Kamra.[62]

The discussion hereinabove can be condensed into three broad issues with deepfakes regulation the Indian regulatory space: *first*, the lack of proactive detection, control and individualized remedy for deepfake-based offences; *second*, the lack of a coherent theorization of deepfakes as adverse to the ideals of 'autonomy' and 'dignity' on which Indian privacy law is founded; and, *third*, especially concerning political deepfakes, an unsettling trust deficit in the State's ability to impartially combat deepfakes. Any comprehensive regulatory solution for deepfakes must perform on these three fronts to be useful in the Indian context. Inasmuch as deepfakes have also disrupted other jurisdictions, the Indian regulatory experience may be able to draw meaningful insights from the diverse array of regulatory solutions that have already emerged elsewhere in the world in response.

ii. Pure Regulatory Solutions

Regulatory developments by countries around deepfakes are, to varying extents, underway. Deepfakes have made occasional appearances in US state laws. For instance, Virginia became the first state in the US in 2019 to criminalize non-consensual sexually explicit deepfakes, punishable as a misdemeanour.[63] [64] Texas followed soon after to become the first

[60] Sohini Chowdhury, 'IT Rules Amendment Empowers Centre To Identify 'Fake News' In Social Media About Central Govt' (*LiveLaw*, 07 April 2023) <www.livelaw.in/news-updates/it-rules-amendment-empowers-centre-to-identify-fake-news-in-social-media-about-central-govt-225787> accessed 27 July 2024.

[61] Tejasi Panjiar, Prateek Waghre, 'Statement on the notification of the IT Amendment Rules, 2023' (*Internet Freedom Foundation*, 6 April 2023) <internetfreedom.in/statement-on-the-notification-of-the-it-amendment-rules-2023/> accessed 27 July 2024.

[62] Advay Vora, 'Challenge to the IT Rules 2023' (*Supreme Court Observer*, 26 June 2023) <https://www.scobserver.in/journal/challenge-to-the-it-rules-2023/> accessed 27 July 2024.

[63] General Assembly of Virginia, *A BILL to amend and re-enact § 18.2-386.2 of the Code of Virginia, relating to unlawful dissemination or sale of images of another; falsely created videographic or still image; penalty*, House Bill No. 2678 (11 February 2019) <lis.virginia.gov/cgi-bin/legp604.exe?191+ful+HB2678> accessed 27 July 2024.

state in the country to criminalize deepfakes in an electoral context, including the creation and dissemination of deepfake videos created with the intent to deceive, [65] and punishable by a year in county jail and fines up to USD 4,000.[66] California, which is known for its particular attention to internet regulations such as the California Consumer Privacy Act (CCPA), signed into law AB 730, which makes it a crime to distribute audio or video that gives a false, damaging impression of a politician's words or actions. [67] However, the law does not explicitly cover deepfakes and is set to sunset in 2023.[68] Such a time-bound law could be of some use in the forthcoming Lok Sabha elections in India as well – addressing general concerns about electoral malpractice through deepfakes, without compromising the privacy-preserving architecture of intermediaries.

At the federal level, the DEEP FAKES Accountability Bill was tabled before the US Congress in 2019, which establishes requirements for advanced technological false personation records (in simpler words, deepfakes) and establishes criminal penalties for related violations.[69] It establishes new criminal offenses related to – *first*, the production of deepfakes which do not comply with related watermark or disclosure requirements; and, *second*, the alteration of deepfakes to remove or meaningfully obscure such required disclosures. A differentiating factor of this Bill from the state laws is that it places the onus of ensuring that deepfakes are identified as such at source by the creators and manufacturers, and proposes to regulate this interplay across sectors, life cycles and use-cases. The Bill also accounts for victims of deepfakes to move to court in order to establish the falsity of the

---

[64] 'Virginia bans 'deepfakes' and 'deepnudes' pornography' (BBC, 2 July 2019) <www.bbc.com/news/technology-48839758> accessed 27 July 2024.

[65] Texas State Legislature, *Relating to the creation of a criminal offense for fabricating a deceptive video with intent to influence the outcome of an election*, Senate Bill No. 751 (adopted 14 June 2019, enforced 1 September 2019) <legiscan.com/TX/text/SB751/id/1902830> accessed 27 July 2024.

[66] Kenneth Artz, 'Texas Outlaws 'Deepfakes'—but the Legal System May Not Be Able to Stop Them' (*Law.com*, 11 October 2019) <www.law.com/texaslawyer/2019/10/11/texas-outlaws-deepfakes-but-the-legal-system-may-not-be-able-to-stop-them/?slreturn=20230430023423> accessed 27 July 2024.

[67] California State Legislature, *An act to amend, repeal, and add Section 35 of the Code of Civil Procedure, and to amend, add, and repeal Section 20010 of the Elections Code, relating to elections*, Assembly Bill No. 730 (3 October 2019) <leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB730> accessed 27 July 2024.

[68] Colin Lecher, 'California has banned political deepfakes during election season / The bill has raised questions about speech protections' (*The Verge*, 7 October 2019) <www.theverge.com/2019/10/7/20902884/california-deepfake-political-ban-election-2020> accessed 27 July 2024.

[69] 116th Congress, *Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2019 or the DEEP FAKES Accountability Act*, H. R. 3230 (introduced 12 June 2019) (reintroduced as H. R. 2395 before the 117th Congress 08 April 2021) [referred to Sub-Committee on Crime, Terrorism and Homeland Security] <www.congress.gov/bill/116th-congress/house-bill/3230> accessed 27 July 2024.

material and seek compensation in civil court.[70] However, while the Bill was tabled again in 2021, it has yet to be formalised in any manner.[71] Limited insight can also be drawn from the DEEP FAKES Act in the Indian context for the commercial or satirical use of deepfakes where such self-declaration norms could alleviate some concerns. However, there is not much clarity on how malicious deepfakes originating from functionally untraceable sources (as is likely when it comes to political deepfakes or pornography) can be addressed. There is some theorization of individual reputation and locus in addressing deepfakes but justice in these cases becomes subject to the State's capacity and willingness to pursue them.

Parallelly, the EU AI Act, which has been in discussion since 2021, provides for 'transparency obligations,' which includes obliging users to disclose if they have generated any deepfakes or artificial material.[72] The EU, perhaps, is furthest along in the comprehensive theorization of deepfakes as being oppositional to individual privacy and reputation – resulting in the development of these 'obligations' to realize the right to privacy conferred by the General Data Protection Regulation (GDPR). However, the AI Act assigns deepfakes a limited risk rating, and thus does not propose strong measures or punitive actions to non-compliance yet. As such, there is not much clarity on how such obligations will be enforced and how the right to privacy may be safeguarded. At the same time, the EU is able to deploy formidable State resources to ensure compliance with these obligations that may not be as readily available in India (with its nascent data protection capacity and overstretched cyber-policing apparatus).[73] The European Parliamentary Research Service has even gone a step further and commissioned a study on policy designs to tackle deepfakes, including text synthesis technology. [74] Meanwhile, the Indian experience is entirely captivated by criminalization and prosecution.

---

[70] Daniel Lipkowitz, 'Manipulated Reality, Menaced Democracy: An Assessment of the DEEP FAKES Accountability Act of 2019' (*NYU Journal of Legislation & Public Policy Quorum*, 5 March 2020) <nyujlpp.org/quorum/lipkowitz-manipulated-reality-menaced-democracy-deepfakes-accountability-act/> accessed 27 July 2024.

[71] 'All Information (Except Text) for H.R.2395 - DEEP FAKES Accountability Act' (*Congress.gov*, 08 April 2021) <www.congress.gov/bill/117th-congress/house-bill/2395/all-info> accessed 27 July 2024.

[72] Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, COM/2021/206: SEC(2021) 167 <eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206> accessed 27 July 2024.

[73] Nitasha Natu, 'Pendency of cybercrime cases high at 97% in courts' (*The Times of India*, 5 December 2023) <timesofindia.indiatimes.com/articleshow/105740096.cms?> accessed 27 July 2024.

[74] European Parliamentary Research Service, 'Tackling Deepfakes in European Policy' (*Europa*, July 2021) <www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf> accessed 27 July 2024.

A welcome addition to regulatory frameworks around deepfakes in early 2023 is China with its provisions on the Administration of Deep Synthesis Internet Information Services (Draft for Solicitation of Comments), better known as the Cyberspace Administration of China.[75] The provisions are meant to address generative tech across all forms of media, and specifically refer to how 'deep synthesis technologies' have been used 'unscrupulously' in a bid to disinform, slander and steal the identities of people.[76] Chinese platform providers that dispense content generation services and end-users who use these services must clearly label any material generated by AI with a watermark i.e., text or image visually superimposed on the video indicating that the content is machine-generated, in what is one of the most direct and comprehensive regulations for deepfakes in the world so far.[77]

iii. Co-Regulatory Solutions

On the other hand, the European Union (EU), known for its ambitious General Data Protection Regulation (GDPR), has also taken the lead on several related frameworks around internet security and emerging technologies, and specifically mentions deepfakes in a couple of them. For instance, in 2018, the European Commission brought out the Code of Practice on Disinformation which set out a 'European' approach to tackle disinformation. The Code was premised on buttressing popular resilience to disinformation and did not put forth simplistic solutions to the problem. In the words of the High-Level Group of Experts, whose report illuminated the ethos of the Code –

> *Any form of censorship either public or private should clearly be avoided. The HLEG's recommendations aim instead to provide short-term responses to the most pressing problems, longer-term responses to increase societal resilience to disinformation, and a framework for ensuring that the effectiveness of these responses is continuously evaluated, while new evidence-based responses are developed.*[78]

---

[75] 'Provisions on the Administration of Deep Synthesis Internet Information Services (Draft for solicitation of comments)' (*China Law Translate*, 28 January 2022) <www.chinalawtranslate.com/en/deep-synthesis-draft/> accessed 27 July 2024.

[76] Cyberspace Administration of China Ministry of Industry and Information Technology Ministry of Public Security, 'Provisions on the Administration of Deep Synthesis of Internet Information Services' (*CAC*, 25 November 2022) <www.gov.cn/zhengce/zhengceku/2022-12/12/content_5731431.htm> accessed 27 July 2024.

[77] Asha Hemrajani, 'China's New Legislation on Deepfakes: Should the Rest of Asia Follow Suit?' (*The Diplomat*, 8 March 2023) <thediplomat.com/2023/03/chinas-new-legislation-on-deepfakes-should-the-rest-of-asia-follow-suit/> accessed 27 July 2024.

[78] European Commission, 'Final report of the High Level Expert Group on Fake News and Online Disinformation' (*Europa*, 12 March 2018) <https://wayback.archive-it.org/12090/20210303153856/

Players in the digital space welcomed the Code as a co-regulatory solution and assented to the vigilance and disclosure mechanisms prescribed thereby *viz*. strategic communication between intermediaries and public authorities, rebalancing the relationship between social media and news media to promote quality journalism, and safeguarding electoral processes to make them resilient against the march of technological advancement.[79] The European Commission carried out targeted monitoring of Meta (then Facebook), Google, and Twitter on a pilot basis – and received commitment-compliance self-assessment reports from these platforms.[80] Thereafter the Code was assented to in 2020 and has already been instrumental in curbing disinformation during the COVID-19 pandemic.[81]

In June 2022, the European Commission amended the Code of Practice on Disinformation in light of the Russia-Ukraine War to hold social media platforms liable for the dissemination of 'manipulated media,' an umbrella term that implicitly includes deepfakes, by enhancing potential fines to up to 6% of global revenue.[82] It elaborated on the safety and accountability mechanisms prescribed in the 2018 Code by crystalizing the disinformation identification process (viz. labelling deepfakes as 'manipulated media', user self-declaration of AI generated content, fact-checking and enhancing media literacy) [83] and encouraging participation in civil society discourse over the subject.[84] The revised Code substantiates the self-regulatory ethos of the 2018 Code, and evolves the State's role from regulation to governance. In effect, the institutional design envisaged thereby responds to the European Commission's regulatory rationale.[85]

---

https://ec.europa.eu/digital-single-market/en/news/final-report-high-level-expert-group-fake-news-and-online-disinformation> accessed 27 July 2024.

[79] European Economic and Social Committee, Committee of the Regions (European Commission), '*Tackling online disinformation: A European Approach*' [24 April 2018] COM/2018/236 <eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52018DC0236> accessed 27 July 2024.

[80] EU Code of Practice on Disinformation, 2018, III. Measuring and Monitoring the Code's Effectiveness <ec.europa.eu/newsroom/dae/redirection/document/87534> accessed 27 July 2024.

[81] European Commission, '2018 Code of Practice on Disinformation' (16 June 2022, *Europa*) <https://digital-strategy.ec.europa.eu/en/library/2018-code-practice-disinformation> accessed 27 July 2024.; *See also*: Julie Posetti, Kalina Bontcheva, "Disinfodemic: dissecting responses to COVID-19 disinformation" (*UNESCO*, 2020) < https://unesdoc.unesco.org/ark:/48223/pf0000374417> accessed 27 July 2024. - the UNESCO policy brief on tackling COVID disinformation materially reproduces the accountability and reporting obligations created by the 2018 Code and recasts them in the context of the pandemic.

[82] 'The 2022 Code of Practice on Disinformation' (*Europa*, 16 June 2022) <digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation> accessed 27 July 2024.

[83] The Strengthened Code of Practice on Disinformation, 2022, Commitments 14-17; ch VII 'Empowering the Fact-Checking Community'.

[84] The Strengthened Code of Practice on Disinformation, 2022, Commitment 12 *read with* ch VI 'Empowering the Research Community: Disclosure of and Access to Signatories' Data for Research Organizations'.

[85] Bronwen Morgan, Karen Yeung, *An Introduction to Law and Regulation* (Cambridge, 2007) 1-14.

iv. Indian Deepfakes Regulation: The Way Ahead

In a world where manipulated media toes the line between banal entertainment, incisive satire, political malevolence and criminal behaviour – no one-size-fits-all approach could address every use-case of deepfakes. While self-regulation has been shown to be a competent middle ground between overbearing censorship and anarchy, it may not be a viable suggestion for India. India has previously seen self-regulation mechanisms in broadcasting (the Broadcasting Content Complaints Council), advertising (the Advertising Standards Council of India), and over-the-top (OTT) digital media (Digital Curated Content Complaints Council). However, public broadcasting and advertising entities had a greater financial incentive to self-regulate sensationalism and shock-value content in the conservative environment of India. Digital content creators/platforms, supplementary as they were to traditional broadcasting, on the other hand had a converse financial incentive to offer more sensational content to attract new viewers. Further, digital content was also observed to be more critical of the establishment, since it was outside the regulatory censorship norms that traditional films were subject to. Consequently, the lack of stringency in self-regulation was cracked down on and digital content was brought within the fold of government oversight, to the din and chagrin of creators.[86] Social media platforms – with revenue models even more steeply dependent on user/viewer-retention, and an even greater potential to foment critique against the State – got caught in the cross-fire as well.[87]

As such, it does not appear that self-regulation of manipulated media is something that India would be amenable to anytime soon. However, in our opinion, focusing on the 'self-regulatory' character of the EU model is uncharitable. In our understanding, the most salient import of the Code is not its regulatory design but the prescription of a policy solution to the fundamental problems with deepfakes. The EU model, in our understanding, is built on three important premises: *first*, that deepfakes exploit the discrepancy between creator-intention and viewer-reception of content where the veracity and authenticity of all media (whether manipulated or not) becomes ambiguous; *second*, that the problem with manipulated media, therefore, is not one of proliferation but of trust (the liar's dividend) emanating from the lack of effective means to disambiguate the authenticity of suspicious media; and, *third*, enabling such disambiguation through proactive labelling of content as 'manipulated media' (at the

[86] Deepali Bhandari & Abhigyan Tripathi, 'Censorship of OTT Media Services: Restraining Freedom of Expression?' (*Law School Policy Review*, 23 December 2020) <lawschoolpolicyreview.com/2020/12/23/censorship-of-ott-media-services-restraining-freedom-of-expression/> accessed 27 July 2024.

[87] Information Technology (Intermediary Guidelines & Digital Media Ethics Code) Rules, 2021.

platform-level, the civil society (fact-checker) level, and the user-level) promotes media literacy and alleviates the societal risks associated with deepfakes.

By addressing deepfakes as a communication-governance problem as opposed to a platform-regulation problem – the EU advances the foundation of some reliable first principles on which particular regulatory solutions (in the electoral and criminal contexts) can be built. How such regulatory solutions are envisaged and designed remains the prerogative of the legislature. What is important, however, is focusing on the media-literacy and empowerment of users which can ameliorate the most acute harms of deepfakes and manipulated media. It is impertinent, at this juncture, to argue the merits of State-regulation *versus* self-regulation. It is imperative instead to imbibe depth to the laws (whatever may be their form) by clearly delineating the problems sought to be solved thereby. On this front, we believe, the EU model performs relatively well compared to other, shallower alternatives to deepfakes-regulation.

V. CONCLUSION

It is clear that deepfakes are quickly evolving into a security concern in their own right given their capacity to distort narratives within as well as beyond the cyberspace, a process that will likely be boosted exponentially by gen-AI and the release of other new technologies in the AI space such as GPT-4.[88] While deepfakes currently in circulation may have chinks in their armour that give away their status of machine-generated material, future technologies will be much better at ensuring that deepfakes are eerily life-like, without production faults that may make them suspicious. This will worsen the ills we currently associate with deepfakes, especially with the marked lack of awareness amongst people about spotting deepfakes and the negligent amount and efficacy of regulations around them so far.

The release of ChatGPT at the end of 2022 sparked an onslaught of discussions around the advanced capacity of AI to create independently of humans,[89] and the epistemic communities around technology policy are now focussed on the new challenges that will emerge as a result. It is essential that these new technologies and challenges are contextualised as per the pros and cons we have observed with their predecessors in the past, and that effective techno-legal

[88] 'OpenAI's GPT-4 to bring multimodal capabilities with AI-generated videos and faster responses, say reports' (*The Economic Times*, 12 March 2023) <economictimes.indiatimes.com/news/new-updates/openais-gpt-4-to-bring-multimodal-capabilities-with-ai-generated-videos-and-faster-responses-say-reports/articleshow/98579150.cms> accessed 27 July 2024.

[89] Shimona Mohan, Varya Srivastava, 'Aggregative, derivative, creative: Generative AI and human creativity' (*ORF*, 06 January 2023) <www.orfonline.org/expert-speak/aggregative-derivative-creative/> accessed 27 July 2024.

regulations around their use (and, inevitably, abuse and misuse) are drawn up to provide future-proof guidelines. For this, we need to ascertain the values we wish to instil in emerging technologies like deepfakes, since technology is not value neutral, and ensure that general as well as sectoral frameworks reflect them. As we reach the point of no return, or 'singularity,'[90] with technological development, it is critical to ensure that our regulatory frameworks are aligned to safeguard our human, democratic, and constitutional values.

[90] Gary Grossman, 'Generative AI may only be a foreshock to AI singularity' (*Venture Beat*, 11 February 2023) <venturebeat.com/ai/generative-ai-may-only-be-a-foreshock-to-ai-singularity/> accessed 27 July 2024.