A SOCIO-LEGAL INQUIRY ON DEEPFAKES

*ANURAGINI SHIRISH AND **SHOBANA KOMAL

ABSTRACT

The proliferation of fake synthetic media has grown exponentially in the past several years. The democratization of artificial intelligence tools propelled the use of deepfake technologies for both progressive and regressive purposes. Similar to fake news, the malicious use of deepfakes is a real threat to the democracy of nations because it can significantly alter the central pillar of truth. Legislators and policymakers struggle to

* Anuragini Shirish, Université Paris-Saclay, Univ Evry, IMT-BS, LITEM, 91025, Evry-Courcouronnes, France. Anuragini Shirish is a lawyer in India and a professor of management of information systems at the Institute Mines-Télécom Business School, France. She completed her accreditation to supervise research from University Strasbourg. She also holds an executive master's degree in public policy innovation from the London School of Economics and undertook her master's degree in law at the National University of Singapore. She is recognized as a distinguished member cum laude by the Association for Information Systems (AIS). She heads the SMART BIS research group within LITEM, a joint research laboratory focusing on responsible management research at the University of Paris-Saclay, France. She is ranked as 6th in the world under straight count ranking for publishing in six top journals ranked by Association of Information Systems for the period of 2021–2023. Currently, she serves as an associate editor at the European Journal of Information Systems.

** Shobana Komal is an Intellectual Property Rights [IPR] Attorney based in India with over 20 years of experience in Indian and International IP laws. She is also a Consultant for a UK based Law Firm. She has handled IP Portfolios of several internationally renowned Indian and Foreign multi-national companies. She represents IP owners across diverse industries including IT, automotive, engineering, pharmaceutical, entertainment and media, FMCG, fashion and luxury, retail and various service sectors. She is a Speaker in National IP Programs organized by Ministry of Commerce and Industry, Government of India. She regularly offers guest lectures at Indian law schools and colleges of Science and acts as a Jury in Indian moot court competitions. She has authored and contributed to many publications, including the Japanese Patent Office Newsletter. She has drafted amendments in Copyright Law for the Government of India. She has also made significant contributions for updating IPR sections in International Chamber of Commerce Roadmaps by writing Country reports on South Asian countries.

regulate the use of this double-edged technological advancement. In this context, this study traces the workings of deepfake technology and its types. Taking inspiration from technology affordance and social change literature, this study explores how deviant actors' misuse of deepfakes lead to cascading social ramifications at various institutional levels. In addition, this study analyzes the current legal and regulatory scenarios within the major economies of the United States, United Kingdom, European Union, China, and India. The study states countries must focus on a whole-of-society approach. This study offers concrete public policy innovations to prevent, detect, respond, and repair any harm inflicted upon vulnerable actors due to the malicious use of deepfake technology.

TABLE OF CONTENTS

INTRODUCTION

Deepfake technology is the greatest threat to democracy.[1] The term "deepfake" is a portmanteau of "deep learning" and "fake."[2] A legal glossary defines deepfake as "synthetic media which involve[s] the digital manipulation of original source material and which is intended to deceive."[3] In an era where AI and digital connectivity propelled fake visual content to the forefront of societal discourse, deepfake technology blurs the line between truth and deception even further. From humble beginnings, developments in machine learning and data science have helped AI handle increasingly sophisticated tasks.[4] AI is capable of creating hyper-realistic deepfakes by either manipulating or generating text, images, video, and audio which often looks frighteningly real.[5]

A moderator on the Reddit page named "Deepfakes" created the first deepfake in 2017 by posting pornographic videos where the faces of celebrities were depicted on the bodies of adult actors.[6] More recently, political figures have become the target of deepfakes.[7] Deepfake

[1] *See* Neil Sahota, *Technological Solutions for Deepfake Detection During Elections*, NEIL SAHOTA (Aug, 10, 2023), https://www.neilsahota.com/technological-solutions-for-deepfake-detection-during-elections/; Nikolas Lanum, *UN advisor says AI may have 'massive' impact on voters: 2024 will be the 'deepfake election'*, FOX NEWS (Aug. 23, 2023, 6:00 AM), https://www.foxnews.com/media/un-advisor-ai-massive-impact-voters-2024-deepfake-election.

[2] 21st Eur. Conf. on Cyber Warfare & Sec., *The U.S. Cyber Threat Landscape*, at 22 (June 8, 2022), https://papers.academic-conferences.org/index.php/eccws/article/view/197/341.

[3] Deep Fake, WESTLAW, https://us.practicallaw.thomsonreuters.com/w-039-7965 (last visited Mar. 25, 2024).

[4] These sophisticated tasks AI is capable of handling include image recognition, natural language processing and autonoumous driving. Abill Robert et al., *Explainable AI: Interpreting and Understanding Machine Learning Models*, RSCH. GATE (Jan. 25, 2024), https://www.researchgate.net/publication/377844899.

[5] Zahra Khanjani et al., *How Deep Are the Fakes? Focusing on Audio Deepfake: A Survey*, U. MARYLAND BALTIMORE, INFO. SYS. DEP'T 1 (Nov. 28, 2021), https://doi.org/10.48550/arXiv.2111.14203.

[6] Pramukh Nanjundaswamy Vasist & Satish Krishnan, *Deepfakes: An Integrative Review of the Literature and an Agenda for Future Research*, 51 COMMC'N ASS'N FOR INFO. SYS. 557, 559 (2022); *Deepfakes*, BRITANNICA, https://www.britannica.com/technology/deepfake (last updated Mar. 21, 2024) [hereinafter *Deepfakes Britannica*].

[7] *Deepfakes Britannica*, *supra* note 6.

technology threatens our society's entire information communication environment. Its impact is akin to that of fake news.[8] Deepfakes have direct implications on democratic processes, similar to fake news.[9]

The motives for utilizing deepfake technology can be progressive. These may include contributing to greater autonomy, self-expression, improving educational experiences, enhancing customer engagement, improving artistic expression, and allowing for digital reconstruction for public safety needs.[10] Deepfake technology can also be regressive because it may produce and spread disinformation, defame people, and generate deceptive content, including non-consensual sexual content.[11]

Easy access to online tools like Faceswap and other social media apps facilitated the creation and dissemination of deepfakes.[12] In 2019, Deeptrace, an AI company, found 15,000 deepfake videos on the internet.[13] In just nine months, that number nearly doubled.[14] With this trajectory an estimated ninety percent of digital content would be synthetically generated.[15] There are growing ethical, legal, and security concerns surrounding the misuse of deepfakes. Various nations are developing and enhancing detection techniques and creating policies to mitigate the risks.[16]

This study delves into deepfakes with a focus on its social and legal ramifications. This study first provides an overview of deepfakes and its types. Second, the study discusses an examination of the potential misuses of deepfake technology, including its ramifications on various institutions. Third, a critical overview of the legal frameworks from

[8] *See generally* Vasist & Krishnan, *supra* note 6.

[9] *Id.* at 564–65.

[10] *See generally* Vasist & Krishnan, *supra* note 6.

[11] *See id.*

[12] Kavyasri Naumotu, *Deepfakes are Taking Over Social Media: Can the Law Keep Up?*, 62 INTELL. PROP. L. REV., 102, 107 (2022). See FACESWAP, https://faceswap.dev/ (last visited Mar. 26, 2024) for an available download and overview of the app.

[13] Amanda Lawson, *A Look at Global Deepfake Regulation Approaches*, RESPONSIBLE AI INST. (Apr. 24, 2023), https://www.responsible.ai/a-look-at-global-deepfake-regulation-approaches/.

[14] *Id.*

[15] *Id.*

[16] *Id.* Countries working on AI detection techniques and policies include South Korea, US, UK, France, Germany, Denmark, and Canada. See *id.* for a complete list.

selected countries is offered. A conclusion reflects upon persistent governance challenges and how best to regulate deepfakes in society, offering a solution that proposes a policy framework and offers ethical and legal insights.

I. DEEPFAKES – WORKINGS AND TYPES

Various actors produce deepfakes with diverse intentions. These actors include "(1) communities of deepfake hobbyists; (2) political players, such as foreign governments and various activists; (3) malevolent actors, such as fraudsters; and (4) legitimate actors, such as television companies."[17]

A broader definition considers deepfakes as AI-synthesized content completed through face-swap, lip-sync, and puppet-master.[18] The most common form of deepfakes is face-swap.[19] Face-swap deepfakes overlay the facial features of one individual (target person) onto the image or video footage of another (source person) to make an image or video of the target person do or say things the source person does.[20] Lip-sync deepfakes alter videos to precisely match lip movements with a corresponding audio recording.[21] Puppet-master deepfakes include videos of animating a target person (puppet) based on a subject's actions by mimicking the facial expressions and motions of a different person (master) sitting in front of a camera.[22]

Some deepfakes are made using a conventional method that involves 2D or 3D computer graphics to generate or enhance images for printed artwork and visual media. However, the latest techniques for producing

[17] Mika Westerlund, *The Emergence of Deepfake Technology: A Review*, 9 TECH. INNOVATION MGMT. REV. 39, 41(2019); *Deep Fake – The Greatest Threat to the Idea of Truth*, IAS EXPRESS (Nov. 24, 2020), https://www.iasexpress.net/deep-fake-the-greatest-threat-to-the-idea-of-truth/.

[18] Vasist & Krishnan, *supra* note 6, at 559.

[19] Pranav Dixit, *What are the different types of deepfakes and how you can spot them?*, BUS. TODAY, https://www.businesstoday.in/technology/news/story/what-are-the-different-types-of-deepfakes-and-how-you-can-spot-them-407118-2023-11-25 (last updated Nov. 25, 2023, 4:54 PM).

[20] Thanh Thi Nguyen et al., *Deep learning for deepfakes creation and detection: A survey*, at 1, *printed in* 223 COMPUT. VISION & IMAGE UNDERSTANDING (2022), *accessed at* https://arxiv.org/pdf/1909.11573.pdf.

[21] *Id.*

[22] *Id.*

deepfakes involve deep learning models, including autoencoders and generative adversarial networks (GANs).[23] "These models are used to examine facial expressions and movements of a person and synthesize facial images of anywhere person making analogous expressions and movements."[24] GAN machine learning uses two types of neural networks. The first is a generator that learns and generates fake data.[25] The other is called the discriminator, which compares and distinguishes between real and fake data.[26] If the discriminator fails to detect fake content because of its hyper-realistic similarity to the original version, it will become impossible for humans to detect the difference.

The public has access to easy-to-use software tools to create convincing deepfakes because of deepfake technology democratization. As a result, deepfakes have begun sparking concerns amongst national security agencies regarding its potential misuse.[27]

A. Audio Deepfakes

Audio deepfakes generate or alter audio to produce fake sounds that seem authentic. Audio deepfakes are produced through three methods: replay attack, speech synthesis, and voice conversion. Replay attacks use a stolen audio track to manipulate a voice recognition system, playing the voice recording of the target speaker.[28] Speech synthesis is a method of artificially generating human speech using software or hardware systems which encompasses approaches like text-to-speech.[29] Speech synthesis offers the flexibility to provide various accents and voices beyond pre-recorded human voices.[30] Voice conversion takes

[23] David Miguel Santos Maia, *Optimized Detector of Manipulated Media Content*, UNIVERSIDADE DO PORTO, at 12 (2022), https://sigarra.up.pt/faup/en/pub_geral.pub_view?pi_pub_base_id=572737.

[24] *Id.* at 18–19.

[25] Nilesh Barla, *Generative Adversarial Networks and Some of Gan Applications: Everything You Need to Know*, NEPTUNE AI (Aug. 22, 2023), https://neptune.ai/blog/generative-adversarial-networks-gan-applications.

[26] *Id.*

[27] *NSA, U.S. Federal Agencies Advise on Deepfake Threat*, NAT'L SEC. AGENCY (Sept. 12, 2023), https://www.nsa.gov/Press-Room/Press-Releases-Statements/Press-Release-View/Article/3523329/nsa-us-federal-agencies-advise-on-deepfake-threats/.

[28] Zahra Khanjani et al., *supra* note 5, at 2.

[29] *Id.*

[30] *Id.*

speech from one speaker (source) and modifies it to sound like it was spoken by another speaker (target).[31] Impersonation, a subset of voice conversion, involves pretending to be another person through voice manipulation.[32] Technological advancements enable faster and more accurate voice impersonation.[33] For example, Lyrebird, an AI startup, replicates any voice with just one minute of sample audio.[34]

A journalist was able to access his bank account using an AI replica of his own voice.[35] This interaction questions the validity of voice biometric security systems.[36] This needs biometric security systems to evolve to check not just the veracity but also ensure source-authenticity of the comparison samples. Similar activities are projected to grow rapidly.[37]

B. Text Deepfakes

Text deepfakes entail the creation of fake government documents, emails, or social media posts. It mimics the style and handwriting of specific individuals.[38] One subset of textual deepfakes includes exposed fabrications. Exposed fabrication occurs when information is deceitfully reported by tabloids and sensationalist news outlets that employ eye-catching headlines to boost profit and traffic.[39] Another subset is humorous fakes, or crafting for humorous information that is misconstrued as genuine by ignorant readers.[40] Lastly, there is the large hoax subcategory. Mainstream media is deliberately fabricated to deceive audiences into believing it is authentic news.[41] Research on detection

[31] *Id.* at 3.

[32] *Id.*

[33] *Id.* at 3-4.

[34] Ed Lauder, *Lyerbird's AI can Clone Any Voice With Just One Minute of Sample Audio*, AI BUS. (Apr. 25, 2017), https://aibusiness.com/nlp/lyrebird-s-ai-can-clone-any-voice-with-just-one-minute-of-sample-audio.

[35] Sophia Khatsenkova, *Audio deepfake scams: Criminals are using AI to sound like family and people are falling for it*, EURONEWS.NEXT (Mar. 25, 2023).

[36] *Id.*

[37] Vasist & Krishnan, *supra* note 6, at 557.

[38] Eur. Parl. Doc., No. PE 690.039, II (2021).

[39] Khanjani, *supra* note 5, at 4.

[40] *Id.*

[41] *Id.*

methods is lagging behind research focusing on techniques generating deceptive content.[42]

C. Image Deepfakes

While image editing through traditional software like Photoshop has long been practiced, AI-powered tools introduced a more robust method for extensive image alteration. The mobile application Fakeapp went viral and remains one of the most widely used face-swapping applications to showcase age progression and gender change.[43] It is also possible to change geo-locational images obtained from satellites using deepfake technology.[44] A recent deepfake image showed an explosion near the Pentagon, causing alarm across the economic sector.[45] The response to the deepfake image led to a decline in the stock market.[46]

D. Video Deepfakes

As early as the late 1990s video editors began digitally altering films. In 1997 "Forrest Gump" digitally inserted archival footage of John F. Kennedy and altered his mouth movements.[47] Similarly, Tom Hanks has utilized video digital enhancements in his movie "The Polar Express" as early as 2004.[48] Subsequently, the advent of deepfake technology significantly enhanced the believability of video editing.

Video deepfakes encompass three subcategories: re-enactment, video synthesis, and face-swap.[49] Re-enactment involves manipulating an individual's identity to impersonate them and seemingly exert

[42] *Id.* at 7.

[43] *Id.* at 5.

[44] *See* Bo Zhao et al., *Deep Fake Geography? When Geospatial Data Encounter Artificial Intelligence*, 48 CARTOGRAPHY & GEOGRAPHIC INFO. SCI. 338 (2021).

[45] Davey Alba, *How Fake AI Photo of a Pentagon Blast Went Viral and Briefly Spooked Stocks*, BLOOMBERG, https://www.bloomberg.com/news/articles/2023-05-22/ fake-ai-photo-of-pentagon-blast-goes-viral-trips- stocks-briefly (last updated May 23, 2023, 8:36 AM).

[46] *Id.*

[47] Khanjani et al., *supra* note 5, at 4.

[48] Caroline Twersky, *Tom Hanks and Robin Wright Are Embracing Deep Fake Technology*, W MAG. (Jan. 31, 2023), https://www.wmagazine.com/culture/tom-hanks-robin-wright-here-robert-zemeckis-digitally-de-aged-ai-technology.

[49] *Id.*

control over their actions, expressions, body language, voice dubbing, posture, and gaze direction.[50] Facial expressions can be driven through the technique of expression re-enactment which involves dubbing or lip-syncing for mouth movements.[51] Similarly, the head position can be altered through pose re-enactment, while gaze re-enactment adjusts the direction of the eyes and eyelid position.[52] Further, to transfer the pose of a human body, techniques like human pose synthesis or pose transfer are used.[53, 54]

Moreover, video synthesis, which generates a new video entirely, is different from traditional video editing that involves modifying existing footage.[55] Through video synthesis, neural textures enable the manipulation of video content in diverse environments, even in real-time.[56]

Lastly, face-swap is a type of video deepfake where another person's face is replaced with one's own in an image or video.[57] Eventually, a hybrid approach can be used that integrates various deepfake technologies to create even more sophisticated and compelling content.

II. RAMIFICATIONS OF DEEPFAKES

Face-swap deepfake technology and other hybrid deepfakes impact individuals (including public figures), businesses, communities, society, and democracy.[58] The misuse of deepfakes increases the risk of victimization in society. The face-swapping app Facemega advertises the ease of creating deepfakes.[59] The app could be used to create deepfake porn

[50] *Id.*

[51] *Id.*

[52] Twersky, *supra* note 48.

[53] *Id.*

[54] *Id.*

[55] *Id.*

[56] *Id.*

[57] Twersky, *supra* note 48.

[58] *See generally* Bobby Chesney & Danielle Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 CALIF. L. REV. (2019).

[59] Kylie Cheung, *Disturbing App That Advertised Emma Watson Deepfake Was Removed From App Stores*, JEZEBEL (Mar. 8, 2023, 10:30 PM), https://jezebel.com/disturbing-app-that-advertised-emma-watson-deepfake-was-1850203794.

depicting famous or public-facing women.[60] AI-generated porn is fake—but highly realistic—and done without the consent of the person whose face is used.[61] Experts warn of the short-sightedness of legal regimes and opine they are unequipped to protect vulnerable people, especially those impacted by "revenge porn," which is "non-consensual, nude images of individuals shared by ex-partners or other harassers."[62] California, Virginia, and Texas in the United States (U.S.) explicitly prohibit non-consensual deepfake sexual content.[63]

In 2023, police in India actively investigated the alleged use of a pornographic video and audio deepfake that involved impersonating a police officer to threaten and extort money from a senior citizen.[64] This vulnerable older man contemplated suicide to avoid societal shame.[65]

Face-swapping based content involves relatively fewer resources. Actors can easily use face-swapping technology[66] to commit deviances with a few pictures or videos of the concerned person and access to cloud-based computing resources.[67] However, a deepfake that is difficult to detect will require more sophisticated techniques, better quality images or videos, a variety of input data, and more computing power.[68]

[60] *Id.*

[61] *Id.*

[62] Thomas Crofts & Tyrone Kirchengast, *A Ladder Approach to Criminalising Revenge Pornography*, 83 J. CRIM. L. 87, 93 (2019); Cheung, *supra* note 59.

[63] Cheung, *supra* note 59.

[64] *Man gets caught in deepfake trap, almost ends life; among first such cases in India*, ECONOMIC TIMES, https://economictimes.indiatimes.com/news/new-updates/man-gets-caught-in-deepfake-trap-almost-ends-life-among-first-such-cases-in-india/articleshow/105611955.cms?from=mdr (last updated Nov. 30, 2023, 11:05 AM).

[65] *Id.*

[66] Jan Kietzmann et al., *Deepfakes: Trick or Treat?*, 63 BUS. HORIZONS 135, 137 (2020); Masood et al., *Deepfakes Generation and Detection: State-of-the-art, open challenges, countermeasures, and way forward*, CORNELL ARXIV, https://arxiv.org/pdf/2103.00484v2.pdf (last updated Nov. 23, 2021).

[67] Westerlund, *supra* note 17, at 41, 45.

[68] Catherine Bernaciak & Dominic A. Ross, *How Easy Is It to Make and Detect a Deepfake?*, CARNEGIE MELON UNIV. (MAR. 14, 2022), https://insights.sei.cmu.edu/blog/how-easy-is-it-to-make-and-detect-a-deepfake/; Europol, *Facing reality? law enforcement and the challenge of deepfakes,* at 22 (2024), *available at* https://www.europol.europa.eu/publications-events/publications/facing-reality-law-enforcement-and-challenge-of-deepfakes.

Interestingly, the demand for deepfakes is increasing, which also means some businesses now claim to have the ability to deliver deepfakes as a product or an online service[69] Recorded Future, a popular threat intelligence platform, reported a threat actor's willingness to pay $16,000 USD for this kind of service.[70] Moreover, crime as a service (CaaS), where criminals sell access to AI tools, technologies, and knowledge to facilitate cyber-crime, is expected to rise.[71] As technology advances, CaaS will evolve and result in the automation of crimes.[72] Automated crimes include adversarial machine learning and deepfakes exploiting in near real-time because of improved communication, connectivity, and accessibility of technology.[73] This trend speaks to prior research in the context of fake news, where technology resources such as mobile connectivity and certain institutional factors, such as political freedom and human development levels enhance fake news propensity within nations.[74]

Journalists are the guardians of truth in the information environment; deepfakes have a profound impact on journalism.[75] Several journalists fear their inability to detect the integrity of their sources due to the sophistication of deepfake technology.[76] Media freedom is positively correlated to the propensity for fake news.[77] Therefore, it is essential to provide journalists and the fact-checking community with sufficient safeguards to combat against deepfakes.[78] Scholars have also

[69] For example, deepfakes were used by criminals to override facial recognition to access cryptocurrency. Europol, *supra* note 68, at 13.

[70] *Id*; Jim Nash, *Dark News From Dark Web: Deepfakers are Getting Their Act Together*, BIOMETRIC UPDATE (May 6, 2021, 4:44 PM), https://www.biometricupdate.com/202105/dark-news-from-dark-web-deepfakers-are-getting-their-act-together.

[71] Europol, *supra* note 68, at 10.

[72] *Id.*

[73] *Id.* The emergence of 5G and cloud technology are examples of the accessibility of technology. *Id.*

[74] Anuragini Shirish & Kanika Kotwal, *Impact of Human and Economic Development on Fake News Propensity During the COVID-19 Crisis: A Cross-Country Analysis*, 31 J. GLOB. INFO. MGMT. 9, 9 (2023).

[75] Karin Wahl-Jorgensen & Matt Carlson, *Conjecturing Fearful Futures: Journalistic Discourses on Deepfakes*, 15 JOURNALISM PRAC.803, 804 (2021).

[76] *Id.* at 804–806.

[77] *See* Shirish & Kotwal, *supra* note 74, at 4–5.

[78] *See* Europol, *supra* note 68.

identified how one might soon disregard accurate information; this is called liar's dividend.[79]

The use of deepfake technology raises various legal concerns that cover a wide range of areas such as election law, criminal law, evidence, and intellectual property. These concerns are related to the creation and use of the resulting media and must be taken into account.[80] Defamation, false light, and the right of publicity (personality rights) are the most impacted aspects.[81] It becomes essential to consider other forms of regulation that incite social media platforms to proactively reduce the victimization risks and social instability from such deviances. A European Union ("EU") policy document provides a comprehensive list of psychological, financial, and societal harm from reproduced deepfakes including (s)extortion, defamation, identity theft, reputational damage, fraud, manipulation of elections, erosion of trust, damage to justice system and to democracy as a whole.[82]

Deepfakes can harm various actors in society and have cascading influence. Society functions as an inter-institutional system with multiple institutional orders, including the state, profession, market, corporation, and community. A study done in 2020 by Isam Faik et al., proposed that information technology (IT) can lead to societal change because IT affordances are integral to institutional logic.[83] Per the authors, "sensegiving, translating, and decoupling, through which IT affordances become elements of societal change. [The authors] identify three corresponding carriers through which IT affordances gain scale and stability, namely objects, networks, and platforms."[84] The institutional logic's centrality and compatibility can be strengthened or altered by the scaled IT affordances.[85] The connections between IT and positive advancements in

[79] Kaylyn J. Schiff et al., *The Liar's Dividend: Can Politicians Claim Misinformation to Evade Accountability*, SOCARXIV PAPERS, at 1 (Oct. 19, 2023), https://osf.io/preprints/socarxiv/x43ph.

[80] Erik Gerstner, *Face/Off: "DeepFake" Face Swaps and Privacy Laws*, 87 DEF. COUNS. J. 2, 4 (2020).

[81] *See id.*

[82] Tackling Deepfakes in European Policy, Parl. Eur. Doc., No. PE 690.039, at IV (2021).

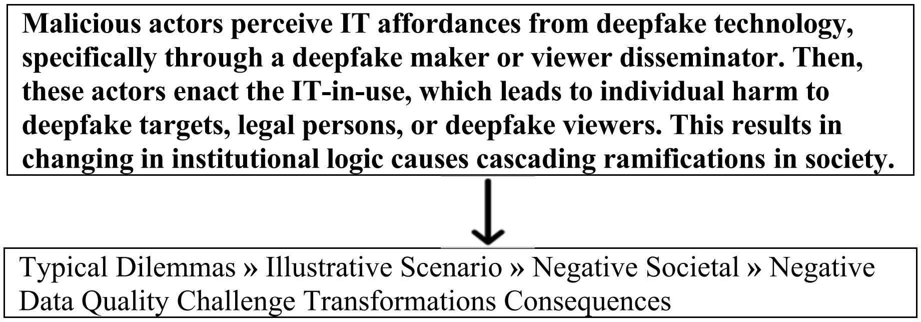
[83] Isam Faik et al., *How Information Technology Matters in Societal Change: An Affordance-Based Institutional Logics Perspective*, 44 MIS Q.2, 2 (2020).

[84] *Id.*

[85] *Id.*

society via societal change can result in enhancements in people's overall happiness and quality of life, increased opportunities for personal growth and development, and greater social and economic integration.[86] However, there are also potential downsides such as the risk of skilled labor becoming obsolete, greater threats to personal privacy, and the emergence of new forms of government control.[87]

Currently, deepfake technologies are widely available and scaled to society. It is pertinent to further examine how IT affordances with malicious motives can alter institutions and lead to negative societal transformations is more pertinent in this context. We use the above theorization to depict how deepfake technology can offer affordances to malicious users while changing the institutional logic by making one or more salient. We also rely on other recent work on deepfakes from the literature on information systems to situate the actors involved in deepfake malicious use case scenarios for the analysis (see details at Table 1).[88]



Malicious actors perceive IT affordances from deepfake technology, specifically through a deepfake maker or viewer disseminator. Then, these actors enact the IT-in-use, which leads to individual harm to deepfake targets, legal persons, or deepfake viewers. This results in changing in institutional logic causes cascading ramifications in society.

Typical Dilemmas » Illustrative Scenario » Negative Societal » Negative Data Quality Challenge Transformations Consequences

State			
Policy challenges to broadening accessibility of 5G/Cloud Technologies and	» In court, law enforcement agencies and the judiciary, cannot trace the integrity	» Reduced legal guardianship and enforcement of crime by legal institutions.	» Enforcement of the rule of law cannot be guaranteed; lessened state

[86] *Id.* at 34–35.

[87] *Id.*

[88] *See infra* Part VII; *see generally* Pramukh N. Vasist & Satish Krishnan, *Deepfakes: An Integrative Review of the Literature and an Agenda for Future Research*, 51 COMMC'NS ASS'N FOR INFO. SYS. 557 (2022).

non-traceability of deepfake technology.	of electronic media as evidence.		control on essential matters, such as justice; negatively impacts democratic processes.
Profession			
Keeping the credibility and authority of media, political and entertainment entities intact.	» Journalists' ability to share the truth is hampered without the integrity of electronic media. Thus, invalidating traditional journalistic practices. Sometimes, journalism is labelled as "deepfake" to divert attention of malicious motives.	» Professional media cannot be expected to act as a third pillar of democracy.	» Threat to democracy; information asymmetry; and public distrust. Increased administrative Burden; climate of uncertainty and chaos.
Market			
Re-evaluating the relevance of traditional media, advertising, and IP rights-based content markets; safeguarding against the	» Dark markets increasingly offer "crime as a service" using deepfake technology for no cost or nominal costs, providing a	» Dominant malicious synthetic media market rather than genuine IP market due to information asymmetry and	» Personality rights, privacy, data protection rights, human rights, and IP rights are easily infringed on easily without

emergence of new markets based on the “crime as service” model.	great incentive to deviate. This particularly threatens legal IP markets.	unethical practices.	much recourse. This can further enable other dark market creations such as pornography, human trafficking and cyber fraud.
Corporation			
Inability to standardize and control operations due to threats of deepfake attacks.	» Cybersecurity attacks or fraud using deepfake challenges companies’ current cyber security controls and procedures, jeopardizes brand image, welfare and economic goals.	» Lack of incentive to indulge in legal economic activities, reducing economic growth; possibility of economic recession leading to unethical and deviant activities by societal actors.	» Stifling economic growth; decline of social norms, causing “anomie” in society.
Community			
Distorting truth, creating public uncertainties; increasing victimization risk among women, children, vulnerable persons and minority groups.	» Difficulty in ensuring the prevention and protection of victims of non-consensual pornography, including women, children and minority groups.	» Public distrust in cyber safety among women, children and minority groups.	» Mental health crisis; community addiction; reality apathy; epistemic injustice.

TABLE 1

III. LEGAL POSITION – GLOBAL INSIGHT

As U.S. Congresswoman Yvette Clark stated, “With few laws to manage the spread of the technology, we stand at the precipice of a new era of disinformation warfare, aided by the use of new AI tools."[89] Many countries have laws regulating digital media and its processes. These laws also address fake information, such as deepfakes. The following section provides a brief overview of the current legal and regulatory situation concerning deepfakes in various jurisdictions, including the U.S., EU, UK, China, and India.

A. United States

In 2018, Senator Sasse (R-NE) of the U.S. Senate introduced the Malicious Deepfake Prohibition Act of 2018.[90] It proposed a new criminal offence for knowingly creating with the intent to distribute or to distribute fake electronic media records that appear realistic that would result in tortious conduct.[91] These are also known as deepfakes. Although the Act was meant to deter, it did not pass the legislature.[92]

In 2019, Representative Yvette D. Clark (D-NY) introduced the DEEP FAKES Accountability Act[93] to impose vast and broad restrictions on the use of deepfake technology in an attempt to "stem the potential damage of synthetic media purporting to be authentic."[94] However, the bill did not pass because it contained "enormous loopholes."[95] For instance, it would require the creator of a deepfake to watermark and identify their content as deepfake, which is highly unlikely to occur.[96]

[89] *Clarke Leads Legislation to Regulate Deepfakes*, CONGRESSWOMAN YVETTE D. CLARKE (Sept. 21, 2023), https://clarke.house.gov/clarke-leads-legislation-to-regulate-deepfakes/.

[90] *Id.*

[91] *Id.*

[92] Lindsey Wilkerson, *Still Waters Run Deep(Fakes): The Rising Concerns of "Deepfake" Technology and Its Influence on Democracy and the First Amendment*, 86 MO. L. REV. 407, 420 (2021).

[93] Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2019, H.R. 3230, 116th Cong. (1st Sess. 2019).

[94] *See* Devin Coldewey, *DEEP FAKES Accountability Act would impose unenforceable rules—but it's a start*, TECHCRUNCH (June 13, 2019, 12:25 PM), https://techcrunch.com/2019/06/13/deepfakes-accountability-act-would-impose-unenforceable-rules-but-its-a-start/.

[95] *Id.*

[96] *Id.*

In 2023, Representative Yvette D. Clark (D-NY) introduced an updated DEEP FAKES Accountability Act, which included a definition of deepfake.[97] In Section 2, subsection(n)(3), of the Bill, deepfake is defined as

> [A]ny video recording, motion-picture film, sound recording, electronic image, or photograph, or any technological representation of speech or conduct substantially derivative thereof—
> (A) which appears to authentically depict any speech or conduct of a person who did not in fact engage in such speech or conduct; and
> (B) the production of which was substantially dependent upon technical means, rather than the ability of another person to physically or verbally impersonate such person.[98]

In order to prevent the dangers this technology creates, Section 7(a) required the establishment of a task force to conduct research to "detect, or otherwise counter and combat" deepfakes.[99] Section 10 proposed mandating online platforms to have a deepfake detection system for the content distributed on such platforms.[100]

Furthermore, the bill proposed that those who intend to distribute advanced technological false personation records on any platform to ensure those records comply with certain requirements.[101] One of these requirements included specific watermarks to identify the digital provenance of the content.[102] The bill proposed that failing to comply with these standards may lead to both criminal and civil sanctions.[103] Similarly, in preparation for elections, three U.S. state governments passed laws mandating either a disclosure or temporary ban on the use of deepfakes during pre-elections that would have resulted in criminal

[97] Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2023, H.R. 5586, 118th Cong. (1st Sess. 2023).

[98] *Id.* at 14–15.

[99] *Id.* at 26.

[100] *Id.* at 30.

[101] *Id.* at 2.

[102] Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2023, H.R. 5586, 118th Cong., at 2–3 (1st Sess. 2023).

[103] *Id.* at 4–5.

and civil liability.[104] However, the federal DEEP FAKES bill and seven state bills did not pass the initial stages and were not codified.[105]

The proposed DEEP FAKES Accountability Bill offered a criminal penalty of fines, imprisonment for not more than five years, or both.[106] The civil penalty included "up to $150,000 per record or alteration and appropriate injunctive relief."[107] For altering disclosures, the proposed bill offered "a civil penalty of up to $150,000 per record or alteration, as well as appropriate injunctive relief."[108]

Despite the failing bills, President Biden under a 2022 Presidential Memorandum, created special task forces to tackle cyber-crimes.[109] These task forces target online harassment and abuse, including gender-based violence through the use of deepfakes, non-consensual distribution of intimate images, cyber stalking, and sextortion.[110]

The Presidential Memorandum outlined the approaches the U.S. should take, which involve further research and accountability.[111] The task force has concluded that it is important to study the impact of evidence-based interventions on the mental health of individuals exposed to online harassment and abuse, especially among young people.[112] This insight gained from such research may lead to the prevention of hate

[104] Adam Edelman, *States are Lagging in Tackling Political Deepfakes, Leaving Potential Threats Unchecked Heading into 2024*, NBC (Dec. 16, 2023, 4:00 AM), https://www.nbcnews.com/politics/artificial-intelligence-deepfakes-2024-election-states-rcna129525.

[105] *All Actions: H.R. 5586 – 118th Congress (2023–2024)*, CONGRESS.GOV, https://www.congress.gov/bill/118th-congress/house-bill/5586/all-actions (last visited Apr. 17, 2024); Edelman, *supra* note 104.

[106] H.R. 5586 at 5.

[107] Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2023, H.R. 5586, 118th Cong., at 7 (1st Sess. 2023).

[108] *Id.*

[109] *Executive Summary: Initial Blueprint for the White House Task Force to Address Online Harassment and Abuse*, THE WHITE HOUSE (Mar. 3, 2023), https://www.whitehouse.gov/briefing-room/statements-releases/2023/03/03/executive-summary-initial-blueprint-for-the-white-house-task-force-to-address-online-harassment-and-abuse/.

[110] *Id.*

[111] *Id.*

[112] *Id.*

crimes stemming from online misogyny and other similar rhetoric.[113] Success here is contingent on the investment in research that investigates the effects of technology and media on the cognitive, physical, and socio-emotional development of infants and adolescents.[114] To accomplish this, funds must be invested in research surrounding the repercussions of technology and media on the core cognitive, physical, and socio-emotional development of infants and adolescents.[115] The research should also investigate how online harassment and abuse influence the youth social and emotional development of youth.[116]

Furthermore, accountability measures target technological platforms, such as gaming platforms accountable under the Federal Trade Commission's authority.[117] These platforms provide access to the voice, video, and text of participants that can be used for deepfake production.[118] Consequently, the risk of victimizing young children significantly increases.[119] Additionally, cross-border collaboration and partnership are important measures the government must consider to understand the impact of deepfakes on vulnerable populations.[120]

B. European Union

The European Union tackles deepfake legislation in a more aggressive and forward-thinking manner. Apart from laws requiring unambiguous labels for synthetic material, the European Union has stepped up its investigation into the identification and mitigation of deepfakes.[121] Europe has several deepfake policies and regulatory frameworks and considers policy directives across deepfake lifecycle it includes five dimensions (technology, creation, circulation, target, and audience).[122]The policy directives include:

[113] *Id.*

[114] *Executive Summary: Initial Blueprint for the White House Task Force, supra* note 109.

[115] *Id.*

[116] *Id.*

[117] *Id.*

[118] *Id.*

[119] *Id.*

[120] *Id.*

[121] Tackling Deepfakes in European Policy, *supra* note 82, at VI.

[122] *Id.*

- The AI regulatory framework
- The General Data Protection Regulation
- Copyright Regime
- e-Commerce Directive
- Digital Services Act
- Audio Visual Media Directive
- Code of Practice on Disinformation
- Action plan on disinformation
- Democracy Action Plan[123]

The EU has put forward regulations and policies that covers the underlying technologies, creation, and circulation dimensions, requiring social media firms to remove misinformation and deepfakes from their platforms. The most recent revision to the EU's Code of Practice on Disinformation was made in June 2022, and it established sanctions for the use of deepfakes of up to six percent of worldwide turnover.[124] This code mandates that the big tech giants such as Twitter, Meta, and Google, begin to identify and flag deepfake material on their social media platforms or risk fines of millions of dollars.[125] Signatories of the code are required to commit resources, like research, to detect and prevent malicious deepfakes.[126] As a result, a permanent task force was formed to monitor the administration, technological advancement and implementation aspects of the code.[127] Furthermore, the code creates the availability for the public to access content provenance and request the removal of any malicious content.[128] Moreover, the signatories must be transparent, accountable, and respect human rights when detecting deepfakes.[129]

[123] *Id.* Direct quote from source.

[124] Amanda Lawson, *A Look at Global Deepfake Regulation Approaches*, RESPONSIBLE ARTIFICIAL INTELLIGENCE INSTITUTE (Apr. 24, 2023), https://www.responsible.ai/a-look-at-global-deepfake-regulation-approaches/.

[125] *See The Strengthened Code of Practice on Disinformation*, at 7, COM (2022) (June 16, 2022).

[126] *Id.* at 30.

[127] *Id.* at 37.

[128] *See id.* at 18.

[129] *Id.* at 1.

The code was first introduced as a means of self-regulation, but with the support of the Digital Services Act, it has now become mandatory.[130] The Act intends to regulate online intermediaries and platforms such as marketplaces, social networks, content-sharing platforms, app stores, and online travel and accommodation platforms.[131] The main goal is to prevent illegal and harmful online activities and the continuing spread of disinformation.[132]

The EU AI Act was proposed to subject deepfake providers to transparency and disclosure requirements.[133] Article 52 of the Act imposes transparency obligations for providers and users of certain AI systems.[134] It instructs AI users who generate deepfakes to disclose the content was artificially generated or manipulated.[135]

Scholars note the EU AI Act does not consider all deepfakes as high-risk AI systems.[136] However, the AI Act allows for reclassification.[137] Therefore, in order to safeguard particular individuals, it is imperative to target deepfakes that pose a major threat to them.[138] The possibility of reclassifying makes it possible to elevate the deepfake to a higher category and precisely assess the extent of its harmfulness, eliminating any possibility of harm.[139] For example, since there are significant ramifications for the creation and dissemination of deepfakes, it is important to introduce additional protections for political candidates.

[130] Lawson, *supra* note 124. *See generally* Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and Amending Directive 2000/31/EC, 2022 O.J. (L 277) 1 [hereinafter *Digital Services Act*].

[131] Digital Services Act, *supra* note 130, at 2.

[132] *Id.* at 3.

[133] Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonized Rules on Artificial Intelligence and Amending Certain Union Legislative Acts, COM (2024) 5562 final (Jan. 26, 2024) [hereinafter *Artificial Intelligence Act*].

[134] *Id.* at 164.

[135] *Id.*

[136] Mateusz Łabuz, *Regulating Deep Fakes in the Artificial Intelligence Act*, 2 APPLIED CYBERSECURITY & INTERNET GOVERNANCE 29, 29 (2023).

[137] *Id.*

[138] *Id.*

[139] *Id.*

Moreover, scholars recommend the potential harm from deepfakes should be included to a version of the AI Act.[140]

C. United Kingdom

The threat posed by deepfakes has prompted the UK government to launch a number of measures. To create systems for identifying and reacting to deepfakes, it is specifically supporting research into deepfake detection technologies and collaborating with academic and industry organizations.[141] The UK government has yet to enact any laws prohibiting the creation and dissemination of deepfake content.[142] Ironically, like the EU, the UK promotes research and development and uses its communication channels to spread awareness about the perils of revenge porn and deepfake videos.[143] It has taken a sectoral approach to AI regulation this impacts how deepfakes are regulated.[144]

Laws governing privacy, defamation, and harassment provide remedies for victims of deepfakes. For instance, Section 85 of the Copyright, Designs and Patents Act 1988 protects individuals against anyone who commissions a video or photographs they are not authorized to have.[145] Deepfakes are also included.[146] In addition, the UK's Computer Misuse Act of 1990 could apply to the production and distribution of harmful deepfakes.[147]

[140] *Id.*

[141] Lawson, *supra* note 124.

[142] *Id.*

[143] *Id.*

[144] Kir Nuthi, *An Overview of the UK's New Approach to AI,* CTR. FOR DATA INNOVATION (Apr. 19, 2023) https://datainnovation.org/2023/04/an-overview-of-the-uks-new-approach-to-ai/.

[145] Copyright, Designs and Patents Act 1988, c. 48, § 85 (UK, 1988).

[146] *UK: Considering deepfakes from a data protection prospective*, ONE TRUST DATA GUIDANCE, (May 2023) https://www.dataguidance.com/opinion/uk-considering-deepfakes-data-protection-perspective; *see* Copyright, Designs and Patents Act 1988, *supra* note 145.

[147] *See* Computer Misuse Act 1990, c. 18 (UK, 1990); Jasmine Lovell, *Deepfake Dilemma: Navigating Legal Implications and Regulatory Responses*, LINKEDIN, (June 8, 2023) https://www.linkedin.com/pulse/deepfake-dilemma-navigating-legal-implications-responses-lovell/.

In October 2023, the UK passed the long anticipated Online Safety Act,[148] including amendments to the Sexual Offences Act 2003.[149] It is believed that the catalyst for this was police released data estimating that around one in fourteen adults in England and Wales had received threats of intimate image sharing.[150] The new Online Safety Act implements a new dimension by criminalizing the non-consensual distribution of intimate deepfakes.[151] Additionally, a "“false communications offence” is committed when a person knowingly sends misinformation with the intent to cause physical or psychological harm."[152] However, these laws were not initially created for advanced AI technology, as they could only give remedies and not mitigate such malicious deepfakes.[153] While this protects potential victims of online image abuse, it fails to address non-sexual deepfakes.

D. China

It is not surprising that China is amongst the first few to adopt strict regulations, specifically on AI and deep synthesis technology, considering China consistently exercised strict control over the internet and its uses. It is currently the only country that strictly bans certain deepfakes.[154]

China has passed laws to deal with AI in its entirety. Among these bills are (1) the Algorithm Recommendation for Internet Information Services (2022),[155] (2) Deep Synthesis of Internet Information Services

[148] *See* Online Safety Act 2023, c. 50 (UK, 2023).

[149] *See* Sexual Offences Act 2023, c. 42 (UK, 2023).

[150] Lawson, *supra* note 124.

[151] Online Safety Act 2023, c. 50 § 66B (UK, 2023).

[152] Giles Parsons et al., *UK: Considering deepfakes from a data protection perspective*, DATAGUIDANCE (last visited May 2023), https://www.dataguidance.com/opinion/uk-considering-deepfakes-data-protection-perspective.

[153] *See id.*

[154] Caroline Quirk, *The High Stakes of Deepfakes: The Growing Necessity of Federal Legislation to Regulate This Rapidly Evolving Technology*, PRINCETON LEG. J., June 19, at 3 (2023).

[155] Rogier Creemers et al., *Translation: Internet Information Service Algorithmic Recommendation Management Provisions – Effective March 1, 2022*, DIGICHINA (Jan. 10, 2022), https://digichina.stanford.edu/work/translation-internet-information-service-algorithmic-recommendation-management-provisions-effective-march-1-.2022.

(2023),[156] and (3) Interim Administrative Measures for Generative Artificial Intelligence Services (2023).[157]

When utilizing "algorithmic technology of generation and synthesis," service providers are required under China's administrative regulations on the Algorithm Recommendation for Internet Information Services Act to notify consumers of the information on algorithm recommendation services.[158] The service provider is required to make the fundamental ideas, goals, and primary workings of algorithm recommendation public.[159]

The Deep Synthesis of Internet Information Services legislation requires deep synthesis service providers to take up information security and rule-making responsibilities. These responsibilities include policies and procedures such as "user registration, algorithm mechanism reviews, scientific and technological ethics reviews, information release reviews, data security, protection of personal information, anti-telecommunication fraud, and emergency response."[160] Additionally, "providers are required to add a digital "mark" that does not interfere with users' use and maintain relevant information in accordance with applicable laws."[161] When introducing new products, features, or apps that have the potential to influence public opinion or spark social mobilization, deep synthesis service providers must conduct security assessments in compliance with applicable legislation.[162]

The interim administrative procedures for generative artificial intelligence services aim to balance the advancement and use of AI

[156] *See Provisions on the Administration of Deep Synthesis Internet Information Services*, CHINA L. TRANSLATE (Nov. 15, 2022), https://www.chinalawtranslate.com/en/deep-synthesis/.

[157] *See Interim Measures for the Management of Generative Artificial Intelligence Services*, CHINA L. TRANSLATE (July 10, 2023), https://www.chinalawtranslate.com/en/generative-ai-interim/.

[158] Justina Zhang, *AI-Deep Synthesis Regulations and Legal Challenges: Recent Face Swap Fraud Cases in China*, HERBERT SMITH FREEHILLS (July 27, 2023), https://hsfnotes.com/tmt/2023/07/27/ai-deep-synthesis-regulations-and-legal-challenges-recent-face-swap-fraud-cases-in-china/.

[159] *Id.*

[160] *Id.*; *see also Provisions on the Administration of Deep Synthesis Internet Information Services*, CHINA L. TRANSLATE (Nov. 15, 2022), (https://www.chinalawtranslate.com/en/deep-synthesis/.

[161] Zhang, *supra* note 158.

[162] *Id.*

while maintaining control and protecting users.[163] However, these regulations only apply to publicly accessible generative AI services and "do not extend to the internal research and use of private organizations".[164] Service providers must sign service agreements with customers and ensure the protection of personal data.[165] They also have to carry out security evaluations for AI services that have the potential to influence public opinion or mobilize social media[166] Additionally, service providers must monitor and remove illegal information produced by their platforms[167] "These regulations may have international ramifications, giving Chinese authorities the authority to impose penalties or technological measures on foreign suppliers that disobey the regulations."[168] However, Chinese restrictions and extensive rules on AI might be deemed as censorship by critics of the authoritarian regime who also predict reduced competitiveness in the AI race.

The Cyberspace Administration of China introduced new regulations to restrict the use of deep synthesis technology and prevent the spread of misinformation. The rules apply to two entities—platform providers and end-users.[169] The policy requires that any content created using this technology is clearly marked as doctored and can be traced back to its source.[170] Additionally, the providers of deep synthesis services must comply with local laws, uphold ethical standards, and in order to prevent misleading the public, providers must maintain the

[163] *Id.*; *Regulatory and legislation: China's Interim Measures for the Management of Generative Artificial Intelligence Services officially implemented*, PwC TIANG & PARTNERS 1 (Aug. 2023), https://www.pwccn.com/en/tmt/interim-measures-for-generative-ai-services-implemented-aug2023.pdf [hereinafter *Regulatory and legislation*].

[164] Zhang, *supra* note 158; *Regulatory and legislation*, *supra* note 163, at 2.

[165] Zhang, *supra* note 158; *Regulatory and legislation*, *supra* note 163, at 2.

[166] Zhang, *supra* note 158; *Regulatory and legislation*, *supra* note 163, at 3.

[167] Zhang, *supra* note 158; *Regulatory and legislation*, *supra* note 163, at 3.

[168] Zhang, *supra* note 158.

[169] Asha Hemrajani, *China's New Legislation on Deepfakes: Should the Rest of Asia Follow Suit?*, DIPLOMATE (Mar. 8, 2023), https://thediplomat.com/2023/03/chinas-new-legislation-on-deepfakes-should-the-rest-of-asia-follow-suit/.

[170] Ben Jiang, *China's internet censors target deepfake tech to curb online disinformation*, S. CHINA MORNING POST (Dec. 12, 2022 9:00 PM), https://www.scmp.com/tech/policy/article/3203000/chinas-internet-censors-target-technology-behind-deepfakes-curb-online-disinformation

correct political direction and public opinion orientation.[171] In addition to specific laws regulating deepfakes, China has enacted various laws that concern data security, IP, and personal information protection.[172]

Despite strict internet rules, China has encountered face-swap fraud. In April 2023, Chinese influencer "CaroLailai" discovered AI generated porn with her face appearing on a porn actress.[173] In 2023, AI was used in a video chat to impersonate a friend of a legal representative of technology company Fuzhou. The deepfake video exchange had convinced him to transfer 4.3 million Yuan to a fraudster (approx. 612,000 USD). [174]

The recent cases of face-swapping in China have raised concerns regarding the legal implications of deepfake technology. In this regard, China's criminal law and civil codes have provisions that are applicable to such cases.[175] According to criminal law in China, anyone fabricating or disseminating false information to disturb the public through media platforms can face up to three years of imprisonment or fines.[176] The civil code protects individuals whose personal rights are violated, and offenders may be held liable for civil damages.[177] IT Rules offers guidance and accountability related provisions to tackle sexual offences carried out via social media in order to protect vulnerable populations, such as women[178] These rules apply to both users as well as to intermediaries of social media as it pushes for strict adherence to the guidelines, extending accountability for any misuse or abuse.[179] However, there remain challenges in identifying the creator and first sender of deepfakes.

[171] Zhang, *supra* note 158.

[172] *Id.* These include, but are not limited to, the Data Security Law, Copyrights Law, Civil Codes, and Cybersecurity Law. *Id.*

[173] *Id.*

[174] *Id.*

[175] *Id.*

[176] *Id.*; *see also Criminal law of the People's Republic of China,* CON. EXEC. COMM. ON CHINA (2016), https://www.cecc.gov/resources/legal-provisions/criminal-law-of-the-peoples-republic-of-china.

[177] Zhang, *supra* note 158.

[178] *Information Technology Rules, 2021*, DRISHTIIAS (Mar. 13, 2021), https://www.drishtiias.com/daily-news-editorials/information-technology-rules-2021.

[179] *Id.*

IV. STATE OF GOVERNANCE

According to a review of legal measures taken through 2024, the U.S. has attempted to take proactive steps to address concerns related to deepfakes, although many have failed. This includes attempting to propose laws such as the Malicious Deepfake Prohibition Act,[180] the DEEP FAKES Accountability Act of 2018,[181] and the DEEP FAKES Accountability Act of 2023.[182] Furthermore, the U.S. has allocated resources for research and development to combat disinformation.[183] Regulatory agencies like the Federal Trade Commission (FTC) have a role in monitoring and enforcing compliance with existing laws related to deceptive practices, which may include aspects of deepfake manipulation.[184] U.S. Regulatory Agencies have also provided broad plans suggesting a focus on enhancing public awareness of deceptive AI by collaborating between government, industry, and civil entities to mitigate the risks of deepfakes through upholding principles of transparency and freedom of expression.[185]

The UK initiated discussions to improve its regulations regarding deepfake technology due to the potential threats it poses. While no specific legislation targets deepfakes, existing laws related to defamation, privacy, and fraud may apply to instances involving deepfake

[180] Malicious Deep Fake Prohibition Act of 2018, S. 3805, 115th Cong. (2d Sess. 2018).

[181] Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2019, H.R. 3230, 116th Cong. (1st Sess. 2019).

[182] Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2023, H.R. 5586, 118th Cong. (1st Sess. 2023).

[183] Vivian S. Walker & Ryan E. Walsh, *Public Diplomacy and The New "Old" War: Countering State-Sponsored Disinformation*, U.S. ADVISORY COMM'N ON PUB. DIPL., at 4 (Sept. 2020), https://www.state.gov/wp-content/uploads/2020/09/Public-Diplomacy-and-the-New-Old-War-Countering-State-Sponsored-Disinformation.pdf.

[184] *A Brief Overview of the Federal Trade Commission's Investigative, Law Enforcement, and Rulemaking Authority*, FED. TRADE COMM'N, https://www.ftc.gov/about-ftc/mission/enforcement-authority (last revised May 2021).

[185] *Press Release, NSA, U.S. Federal Agencies Advise on Deepfake Threats*, NAT'L SEC. AGENCY (Sept. 12, 2023), https://www.nsa.gov/Press-Room/Press-Releases-Statements/Press-Release-View/Article/3523329/nsa-us-federal-agencies-advise-on-deepfake-threats/.

manipulation. The Online Safety Bill in the UK recognized deepfake pornography as a new criminal offense, but it does not prohibit other types of AI-generated content created without the subject's consent, which leaves victims to rely on existing laws.[186] The UK's approach includes direction from the European Union which aims to encourage teachers and educators to promote digital literacy and awareness among the public to help identify and mitigate the risks associated with deepfakes.[187] Moreover, the UK initiated the establishment of an AI safety agency tasked with assessing powerful AI models to prevent their deviation from intended objectives.[188]

China has enacted laws and regulations to combat the spread of disinformation and ensure cybersecurity. The Cyberspace Administration of China (CAC) issued guidelines requiring online platforms to implement measures to prevent the dissemination of false information, including deepfakes.[189] However, some scholars consider the implementation of the regulator to be non-transparent.[190] As a policy, the Association of Southeast Asian Nations (ASEAN) advocates strengthening deepfake detection technology and raising social awareness of deepfakes to combat the harm they cause.[191]

India has not yet implemented specific regulations or policies that directly address deepfake technology, even though existing laws may be read to apply to deepfakes. However, concerns about the spread of

[186] Parsons, *supra* note 152.

[187] *See generally* European Comm'n, Director-General for Education, Youth, Sport and Culture, *Guidelines for teachers and educators on tackling disinformation and promoting digital literacy through education and training*, EUROPEAN UNION, 2022, https://op.europa.eu/en/publication-detail/-/publication/a224c235-4843-11ed-92ed-01aa75ed71a1/language-en.

[188] Policy Paper, Introducing the AI Safety Institute, Parliament by the Secretary of State for Science, Innovation and Technology by command of His Majesty, https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute (last updated Jan. 17, 2024).

[189] *China's interim guidelines for generative AI services*, DIGWATCH (July 13, 2023), https://dig.watch/updates/chinas-interim-guidelines-for-generative-ai-services.

[190] Jamie P. Horsley, *Behind the Façade of China's Cyber Super-Regulator*, DIGICHINA (Aug. 8, 2022), https://digichina.stanford.edu/work/behind-the-facade-of-chinas-cyber-super-regulator/.

[191] *Unveiling Deepfake Dangers to the 2024 Elections, Press Release*, CYBERSECURITY ASEAN (Feb. 28, 2024), https://cybersecurityasean.com/news-press-releases/unveiling-deepfake-dangers-2024-elections.

misinformation and fake news have prompted discussions among policymakers about the need for legislative measures to address these threats.[192] The Indian government emphasized the importance of technological solutions, public awareness campaigns, and collaboration with international partners to combat the spread of deepfakes and disinformation.[193]

As of 2024, India is the third largest digitalized country in the world, following only the U.S. and China.[194] According to the State of India's 2024 Digital Economy Report, digitalization in India is better than that of countries like the UK, Germany, and Japan.[195] As a result of the rapid pace of digitalization, accelerated changes to enhance securities and surveillance have become critical to keep up with the risks associated with deepfake offenses.

Strict laws addressing cybercrimes in combination with enhanced security and surveillance systems are critical for the sake of protecting the privacy and personal rights of citizens. To foster India's economic growth, it is imperative to have dedicated legislation and comprehensive provisions added to existing laws to effectively address the complex issues posed by deepfakes.

V. PERSISTENT DEEPFAKE GOVERNANCE CHALLENGES

The governance of deepfake technology is just emerging. To expand on the aforementioned legislation seen throughout the world, this article discusses persistent issues common across jurisdictions and offers policy options as countermeasures. The challenges countries still face include (1) cross-border jurisdiction; (2) difficulty in identifying deviant actors and authenticating evidence; (3) IP rights and copyrights; (4) personality rights and legal lacuna on post-mortem privacy;

[192] *See* Nidhi Singal, *Here's what the Indian government is planning to do on deepfake tech*, BUS. TODAY, https://www.businesstoday.in/technology/news/story/heres-what-the-indian-government-is-planning-to-do-on-deepfake-tech-406874-2023-11-23 (last updated Nov. 23, 2023, 3:29 PM).

[193] *Id.*

[194] Ashutosh Mishra, *India is the third largest digitalised country in the world, says expert*, BUS. STANDARD, https://www.business-standard.com/economy/news/india-is-the-third-largest-digitalised-country-in-the-world-says-expert-124021600956_1.html (last updated Feb. 16, 2024, 11:08 PM).

[195] *Id.*

(5) the continuous need for research and development; (6) reassessing privacy and data protection concerns; and (7) cyber security resources and raising awareness.

Addressing the regulatory challenges posed by deepfake technology requires a multidisciplinary approach that involves legal, technological, and educational interventions. Collaboration between governments, industry stakeholders, and civil society is essential to developing comprehensive and effective regulatory frameworks.

A. Cross-border Jurisdiction

A significant obstacle in the digital realm is the absence of distinct boundaries, resulting in challenges when identifying the source of a Cyber incident or assessing its most severe consequences. This uncertainty may result in disputes when various jurisdictions assert control over a particular case.[196] In cross-border investigations, bilateral or multilateral agreements can set up structures for sharing information, evidence, and responsibilities.[197] Nevertheless, reaching a consensus among various jurisdictions is a challenging endeavor which rests on diplomacy and legal alignment so as to guarantee efficient governance of cyberspace.[198]

B. Difficulty in Identifying Deviant Actors and Authenticating Evidence

Since deepfake-related misuse often occurs anonymously, authorities may face difficulty in tracing the criminal activity. Showing evidence of such deviant behavior is problematic in judicial systems, which in turn further complicates the trial process.[199]

[196] Kirtika Sarangi, *Issues And Concerns Of Cyberspace Jurisdiction In India*, LINKEDIN (Nov. 8, 2023), https://www.linkedin.com/pulse/issues-concerns-cyberspace-jurisdiction-india-kirtika-sarangi-vk6qc#:~:text=One%20of%20the%20primary%20challenges,boundaries%20of%20a%20specific%20jurisdiction.

[197] *Id.*

[198] *Id.*

[199] *See generally* Marie-Helen Maras & Alex Alexandrou, *Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos*, 23 INT'L J. EVID. & PROOF 255 (Oct. 28, 2018).

On the popular American streaming website Twitch, user QTCinderella was the target of numerous unauthorized deepfake pornography videos.[200] Many lawyers thought there was no way to achieve legal remedy due to "the lack of existing regulation."[201] According to a report, an individual fraudulently transferred €200,000 from a UK energy plant by using AI technologies to imitate another company executive's voice.[202] The CEO fell victim to this scam, but the company detected the fraud quickly. Unfortunately, they were not able to catch the scammer and receive compensation for the losses.[203]

Deepfakes are synthetic media that can create confusion in court when produced as fake evidence by parties involved in civil or criminal proceedings. This can delay trials and increase the risk of wrongful convictions based on misleading evidence.[204] Deepfake technology can also manipulate surveillance footage to change the identities of people caught on camera, which can result in wrongful arrests and false charges.[205] This can seriously affect people's trust in security and surveillance systems. However, blockchain technology has been touted to accurately track deepfake content at its source.[206]

C. Intellectual Property Rights – Copyrights

Deepfake technology involves the reproduction and alteration of existing materials, which could potentially infringe on someone else's

[200] Caroline Quirk, *The High Stakes of Deepfakes: The Growing Necessity of Federal Legislation to Regulate This Rapidly Evolving Technology*, PRINCETON L. J. (June 19, 2023), https://legaljournal.princeton.edu/the-high-stakes-of-deepfakes-the-growing-necessity-of-federal-legislation-to-regulate-this-rapidly-evolving-technology/.

[201] *Id.*

[202] *Id.*

[203] *Id.*

[204] Tackling Deepfakes in European Policy, *supra* note 82, at 55.

[205] Faceadapter, *Deepfakes and Biometric Attacks: Emerging Threats in Security and Surveillance*, LINKEDIN (Apr. 17, 2023), https://www.linkedin.com/pulse/deepfakes-biometric-attacks-emerging-threats-security-surveillance/.; *see, e.g.*, Kalev Leetaru, *Deep Fakes' Greatest Threat Is Surveillance Video*, Forbes (Aug. 29, 2019), https://www.forbes.com/sites/kalevleetaru/2019/08/26/deep-fakes-greatest-threat-is-surveillance-video/.

[206] Nicholas Mesa-Cucalon, *Deepfakes: Effective Solutions for Rapidly Emerging Issues*, ANALYTICS VIDHYA (May 26, 2021), https://medium.com/analytics-vidhya/deepfakes-effective-solutions-for-rapidly-emerging-issues-8b1685feef56.

copyright. While it could have a positive outcome, using this technology without permission is socially regressive. Deepfakes can also be used to create advertisements with celebrities who never intended to endorse the products or services in the ads. Deepfakes can be used to transform the narrative and characters of a filmmaker's work or to recreate a musician's voice to create new songs.[207] In a recent case, AI generated song resembling the work of Drake and the Weeknd called "Heart on my Sleeve" went viral.[208] Universal Music Group alerted the internet platforms alleging copyright violation and consequently the unauthorized imitative track was taken down by the platforms.[209]

The advent of deepfakes has also sparked debate about potential fair use defenses, which permit limited use of copyrighted material without permission.[210] In *Campbell v. Acuff Rose*, the U.S. court held content that is transformative in nature to be protected under the fair use doctrine.[211] Thus, creators of deepfakes may not be liable for copyright infringement as deepfakes could fall within the ambit of transformative works and thus be protected by legal exceptions [212]

D. Personality Rights and Legal Lacuna on Post-Mortem Privacy

Personality rights are rights given to personalities to protect their names, images, voices, personal attributes, and mannerisms from being misused and commercially exploited.[213] These rights play an important role for celebrities and famous personalities in protecting their personality traits from being copied and displayed in public without their consent.

[207] Michael Feffer et al., *DeepDrake ft. BTS-GAN and TayloRVC: An Exploratory Analysis of Musical Deepfakes and Hosting Platforms*, HUMAN-CENTRIC MUSIC INFO. RETRIEVAL, at 1 (Nov. 10, 2023), https://ceur-ws.org/Vol-3528/.

[208] Chloe Veltman, *When you realize your favorite new song was written and performed by . . . AI*, NPR (Apr. 21, 2023, 5:00 AM), https://www.npr.org/2023/04/21/1171032649/ai-music-heart-on-my-sleeve-drake-the-weeknd.

[209] *Id.*

[210] Jonathan Alexander Fisher, *"Fair" in the Future? Long-Term Limitations of the Supreme Court's Use of Incrementalism in Fair Use Jurisprudence*, 32 FORDHAM INTELL. PROP. MEDIA & ENT. L. J. 808, 836, 844-845 (2022).

[211] Quirk, *supra* note 200, at 4.

[212] *Id.*

[213] Agnes Augustian, *Protection of Personality Rights In India: Issues And Challenges*, 1 IPR J. MAHARASHTRA NAT'L L. U., NAGPUR 44, 44 (2023).

Many countries lack explicit laws that acknowledge and address violations of personality rights, however, some countries (like the U.S.) have robust laws.[214] Violations of personality rights are likely to occur with non-consensual AI based generation of simulated likenesses of persons and their attributes, like voice and appearance.[215] The debate on post-mortem privacy has existed for decades, but the emergence of deepfakes has elevated this to a new level, with both moral and commercial implications.[216]

Grammy award-winning music composer A.R. Rahman used technology to recreate the voices of the late Shahul Hameed and Bamba Bakya for the song "ThimiriYezhuda."[217] A.R. Rahman clarified that the families gave permission to use the voices and received compensation.[218] Although many music lovers welcome creations like this, others criticize the ethical nature of exploiting deceased persons.[219]

Most countries offer legal protection for the deceased, especially their rights and interests.[220] For example, there are laws regulating in detail what can and cannot be done with a body.[221] However, there are

[214] *See* Modkova, *Comments on Issues Paper*, WIPO 1 https://www.wipo.int/export/sites/www/about-ip/en/artificial_intelligence/call_for_comments/pdf/ind_modkova.pdf.

[215] WIPO Secretariat, *Draft Issues Paper on Intellectual Property Policy and Artificial Intelligence* 6 (World Intell. Prop. Org. Draft Paper No. WIPO/IP/AI/2/GE/20/1, 2019), https://www.wipo.int/edocs/mdocs/mdocs/en/wipo_ip_ai_2_ge_20/wipo_ip_ai_2_ge_20_1.pdf

[216] Bart van der Sloot & Yvette Wagensveld, *Deepfakes: regulatory challenges for the synthetic society*, 46 COMPUT. L. & SEC. REV., at 9 (Sept. 2022). *See generally* Uta Kohl, *What post-mortem privacy may teach us about privacy*, 47 COMPUT. L. & SEC. REV. (Nov. 2022)

[217] *A.R. Rahman clarifies on using AI to recreate voices of late singers Bamba Bakya, Shahul Hammed*, BUS. INSIDER INDIA (Jan. 30, 2024, 6:14 PM), https://www.businessinsider.in/entertainment/news/a-r-rahman-clarifies-on-using-ai-to-recreate-voices-of-late-singers-bamba-bakya-shahul-hameed/articleshow/107265911.cms.

[218] *Id.*

[219] *See* Dinah Lewis Boucher, *AI can bring you the voice of dead loved ones. But is this a good thing?*, ABC NEWS, https://www.abc.net.au/news/2022-06-26/speaking-to-dead-alexa-will-bring-you-voice-of-dead-loved-ones/101183424 (last updated June 25, 2022, 5:35 PM).

[220] Sloot & Wagensveld, *supra* note 216; *see also* Kirsten R. Smolensky, *Rights of the Dead*, 37 HOFSTRA L. REV. 763, 763 (2009);

[221] Sloot & Wagensveld, *supra* note 216; *see, e.g.*, Peter F. Nemeth, *Legal Rights and Obligations to a Corpse*, 19 NOTRE DAME L. REV. 69 (1943).

yet to be any rules for the post-facto reproduction of a person's characteristics (for instance, voice, appearance), skill or talent etc.[222] Some thought provoking questions raised by Bart Van Der Sloot & Yvette Wagensveld in this regard remain unanswered:

> [T]o what extent is it desirable and permissible to have long gone historical figures teach in schools … deceased actors feature in films… deceased artists give concerts? … To what extent can family members communicate with a deceased loved one even against the will of that person; what does such communication do to trauma processing?[223]

Detailed regimes on the scope of personality rights and post-mortem privacy are required to address these questions.[224]

E. Continuous Need for Research and Development

"Experts predict that in four or five years time, more than 90% of all online content will be manipulated in whole or in part."[225] To combat the spread of fake content, Intel has developed a system that utilizes blood flow changes and eye movements to determine the authenticity of the content.[226] According to Intel, its deepfake detector, Fakecatcher, has a detection accuracy of ninety-six percent.[227] Such programs are fallible and are not a feasible option to authenticate all content via secure and trustworthy systems before they reach online media.[228] Moreover, detection systems can only guarantee about sixty-five percent accuracy in detecting fake news. In the future, as realistic deepfakes and deepfake environments can be more easily generated it is predicted that the above

[222] Sloot & Wagensveld, *supra* note 216.

[223] *Id.*

[224] *Id.*

[225] *Id.* at 11.

[226] *Intel Introduces Real-Time Deepfake Detector*, INTEL (Nov. 14, 2022), https://www.intel.com/content/www/us/en/newsroom/news/intel-introduces-real-time-deepfake-detector.html#gs.84ixfs.

[227] *Intel Introduces Real-Time Deepfake Detector*, INTEL (Nov. 14, 2022), https://www.intel.com/content/www/us/en/newsroom/news/intel-introduces-real-time-deepfake-detector.html#gs.84ixfs.

[228] Sloot & Wagensveld, *supra* note 216, at 3, 11.

accuracy levels will deteriorate.[229] Therefore, detecting them via contextual information will become very challenging.[230] From a research and development perspective, this implies persistent investment of resources in understanding detection mechanisms.

Research efforts should be more holistic, rather than solely technology focused. Therefore, research must analyze issues at a deeper humanistic level. For example, it is helpful to research what makes some people (even non-celebrities) vulnerable to victimization.[231] This requires studying the psychology of humans and their interaction in online environments, such as excessive sharing in social media or online forums.[232]

Research on why some countries are more prone to deepfake attacks is an exciting avenue to investigate. Prior research shows differences in fake news environments in different countries.[233] Research on these aspects may help governments devise policies to increase public digital literacy, therefore fostering and restoring credibility in online media content. This is vital to support a deliberative democracy.

F. Reassessing Privacy and Data Protection Concerns

Deepfakes pose a massive threat to security and surveillance. They can now distort vital personal data, such as iris scans and fingerprints, allowing attackers to mimic legitimate users and defeat security measures.[234] Currently, commercial entities such as the retail, porno-

[229] *Id.* at 11.

[230] *Id.*

[231] *See generally* Craig A. Harper et al., *Development and Validation of the Beliefs About Revenge Pornography Questionnaire*, 35 SEXUAL ABUSE 748 (Sept. 2023); Dean Fido et al., *Celebrity status, sex, and variation in psychopathy predicts judgements of and proclivity to generate and distribute deepfake pornography*, 129 COMP. IN HUM. BEHAVIOR (Apr. 2022), https://www.sciencedirect.com/science/article/abs/pii/S0747563221004647.

[232] Anuragini Shirish et al., *How Can Governments Prevent the Spread of Fake News? A Situational Deviance Prevention Analysis* (forthcoming) (2024)

[233] *See generally* Anuragini Shirish et al., *Impact of mobile connectivity and freedom on fake news propensity during the COVID-19 pandemic: a cross-country empirical examination*, 30 EUROPEAN J. INFO. SYS. 322 (2021).

[234] Zac Amos, *Can Deepfakes Beat Biometric Security*, CYBERSECURITY MAG. (Apr. 24, 2023), https://cybersecurity-magazine.com/can-deepfakes-beat-biometric-security/

graphic industry, and entertainment sectors own cutting-edge deepfake technology. Public prosecutors and data protection authorities are often focused on concerns pertaining to the processing of data from government agencies and private enterprises.[235] Increased accessibility to deepfake technology in the hands of virtually every individual means authorities must consider how to deal with citizens' data.[236] Some thought provoking questions raised by Bart Van Der Sloot & Yvette Wagensveld in this regard remain unanswered: to what extent should the state process citizens' data, and is there a legitimate way to process such data, which raises questions about the exception to the data protection and privacy regime in the context of freedom of expression?[237] Balancing freedom of expression with privacy rights is a delicate matter requiring careful consideration and the development of a new legal policy framework.[238]

These questions have been somewhat addressed by the Indian Supreme Court in its landmark judgement in KS Puttaswamy Vs Union of India (Aadhar case) where it deemed the right to privacy as a fundamental right intrinsic to human dignity and liberty, guaranteed by the Indian Constitution.[239] Every citizen has a fundamental right to their privacy which will prevail especially in those cases where deepfake technologies ascribe malicious attribution to actions, words, and any other synthetic content that is generated without their explicit agency. Those individuals, companies, and platforms who indulge in perpetration of the deepfake technologies/content are culpable for civil and criminal consequences that emerge from violation of such a fundamental right.

G. Cyber Security Resources and Raising Awareness

Experts opine that State budgets for training law enforcement and related positions in Information and Communication Technology (ICT)

[235] *Data Protection Authority & you*, EUR. DATA PROT. BD., https://www.edpb.europa.eu/sme-data-protection-guide/data-protection-authority-and-you_en (last visited Apr. 17, 2024).

[236] *See generally* Sloot & Wagensveld, *supra* note 216.

[237] *Id.*

[238] *See id.*

[239] Justice K.S. Puttaswamy v. Union of India, AIR 2018 SC (Supp) 1841 (2012) (India).

receive less attention than for broader crimes.[240] Other challenges faced in regulating deepfakes according to cyber security experts include a lack of awareness of the culture of deepfake among individuals and organizations and the lack of research and development in ICTs.[241]

Organizations also need to be aware of various technical standards set by private governance initiatives such as the Coalition for Content Provenance and Authenticity, for verifying content source, history, and origin.[242] The Zero Trust Maturity Model can also help organizations safeguard themselves from deepfake attacks.[243]

Educating the general public to identify deepfakes via literary programs is necessary to catch up with advancements in deepfake technologies effectively. Public awareness programs must teach individuals how to spot subtle nuances that may be observable in such artifacts. Examples of these include visible transitions around face edges, blurred contours, limited facial expressions, missing/additional features such as an additional finger, inconsistent lighting, metallic sounds, incorrect diction, high delays, unnatural sounds, incorrect pronunciation, and monotone speech.[244] Such public awareness programs may reduce the risk of victimization from malicious deepfakes.

CONCLUSION

Although several economies have taken steps to address deepfake related challenges, there is still a need for comprehensive and coordinated efforts at the international level to effectively mitigate the risks associated with rapidly evolving deepfake technology. There are also multi-industry organizations using a network approach to create stand-

[240] Coleman McKoy, Law Enforcement Officers' Perceptions in Combating Cybercrime at the Local Level 27–28 (2021) (Ph.D dissertation, Walden University) (on file with Walden Dissertations and Doctoral Studies Collection).

[241] Gobinda Bhattacharjee, *Issues and Challenges of Cyber Crime in India: An Ethical Perspective*, 9 INT'L J. CREATIVE RSCH. THOUGHTS 615, 618 (2021).

[242] *See Overview*, COAL. FOR CONTENT PROVENANCE & AUTHENTICITY, https://c2pa.org (last visited Apr. 17, 2024).

[243] *See Zero Trust Maturity Model*, CYBERSECURITY & INFRASTRUCTURE SEC. AGENCY, https://www.cisa.gov/zero-trust-maturity-model (last visited Apr. 17, 2024).

[244] *Deep Fakes – Threats and Countermeasures*, GER. FED. OFF. FOR INFO. SEC., https://www.bsi.bund.de/EN/Themen/Unternehmen-und-Organisationen/Informationen-und-Empfehlungen/Kuenstliche-Intelligenz/Deepfakes/deepfakes_node.html (last visited Apr. 17, 2024).

ards to combat digital fakes. For example, the DeepTrust Alliance emerged as a self-regulatory alternative.[245] Governments must use a whole-of-society approach to promote public policy innovations to prevent, detect, respond to, and repair any harm inflicted upon vulnerable actors.[246] Such an approach must also be compatible with ethical AI.

There are two broad approaches to the ethics of AI: oppositional and systematic.[247] The oppositional approach casts doubt on all AI technologies and consequently imposes guardrails, making AI more human-centric, and assumes AI operates independently.[248] The EU's approach to deepfake regulations mostly aligns with this vision.[249] The systematic approach considers AI as "a set of technologies embedded in a *system* of human agents, other artificial agents, laws, non-intelligent infrastructures, and social norms."[250] Here, "the ethics of AI can be seen to involve a *socio-technical system* ."[251] In contrast, AI is not viewed "as an isolated technical object" operating independently with its specific design features, rather it is considered "with attention to the social organization within which it will operate."[252] Technology can best be understood or regulated with the knowledge of its use context. These contexts are usually embedded within social organizations that provide various agencies between AI and other social actors within a given organization. Therefore, it is important to consider affordances of AI perceived by social actors (it can be different in various use context such as health care, public sector etc.). It is therefore essential to

[245] *See generally About*, DEEPTRUST ALLIANCE, https://www.deeptrustalliance.org/about (last visited Apr. 17, 2024).

[246] *See* Mikael Wigell et al., Best Practices in the whole-of-society approach in countering hybrid threats, Study Requested by the INGE Committee, EUR. PARLIAMENT (May 2021); Katelyn M Mason, Defending American Democracy in the Post-Truth Age: A Roadmap to a Whole-of-Society Approach (2020) (MA thesis, Naval Postgraduate School) (DTIC) (; Marianne Kjellén et al., *Governance: A 'Whole-of-Society' Approach*, *printed in* THE UNITED NATIONS WORLD WATER DEVELOPMENT REPORT (2023); J.D. Maddox et al., *Toward a Whole-of-Society Framework for Countering Disinformation*, *printed in* GREAT POWER CYBER COMPETITION (2021).

[247] THOMAS M. POWERS & JEAN-GABRIEL GANASCIA, THE ETHICS OF THE ETHICS OF AI 1 (Markus D. Dubber et al. eds., 2020).

[248] *Id.* at 19.

[249] *Id.*

[250] *Id.* (emphasis original).

[251] *Id.* (emphasis original).

[252] *Id.*

evaluate both these approaches, in depth, when designing public policy measures to tackle deepfake technology. Considering the situation or context, such as State specific factors, are equally important when designing and assessing policy options.[253]

We propose a few policy innovations to inspire policymakers grappling with the Janus-faced issue of deepfakes. On the one hand, deepfake technology contributes to innovation and other progressive outcomes. On the other hand, deepfake technology can constrain positive social transformation when misused to commit deviant cyber-crimes in society. These authors' deepfake policy framework envisions four levels: (1) Prevention, (2) Detection, (3) Response, and (4) Repair. The proposals below are specific for each level (Table 2a, 2b, 2c, 2d). These proposals are either legal or general (i.e., non-legal measures). We also explain the proposal, key objectives, policy category, owner, and target for each policy. Public policy innovation policies can use information, participation, or technology as levers to solve societal problems. In this case, prevention, detection, response, and repair are considered policy problems in the context of the growing use of deepfakes for malicious purposes.

Table 2 (a-d): Policy Framework for Malicious Deepfakes
Table 2a Prevention – General
Title: Deepfake Certification Program for Content Creation Tools **Type**: Information/Technology **Policy Owner**: Federal/Central Government **Policy Targets**: Manufactures and developers of deepfake software tools. **Proposal**: Establish industry standards and incentivize developers to prioritize security and ethical considerations in their products. **Details**: Create a government-led certification program for software and tools used in content creation, particularly those capable of generating deepfake media. Manufacturers and developers must undergo rigorous testing and adhere to specific standards to obtain certification. Standards may include transparency requirements and respect for intellectual property rights. There must be provisions to verify provenance and human consent prior to using any intellectual property via the tool. Settings that can automatically disable usage for pornographic purposes and tracking installations, thereby preventing recourse to anonymity feature. This certification will also ensure labelling or watermarking requirements are rigorously followed to verify the data quality

[253] Shirish, *supra* note 232.

of the synthetic media such tools may generate. This program aims to ensure only verified and trusted tools are available for creating synthetic media, thus reducing the potential for widespread misuse. As an example of this is similar to the "Trusted Publishers Program," run by large platforms such as Google and Apple, which verify app developers to ensure the safety and reliability the developed applications.[254]

Prevention – Legal

Title: Specific Digital Watermarking Law in the Context of Synthetic Media

Type: Technology

Policy Owner: Federal/Central Legislators

Policy Targets: Deepfakes manufacturers, developers and users.

Proposal: Improve accountability of content creation and traceability of deepfakes.

Details: Enact laws that require embedding digital watermarks in all media content generated by software capable of producing deepfakes. These watermarks serve as a form of accountability, allowing for traceability of content. Failure to include watermarks could be considered as gross negligence resulting in legal consequences for individuals or entities responsible for creating and disseminating deepfakes. For example, the cybersecurity law in China mandating internet platforms to verify the identities of users and attach digital signatures to their content is similar to this proposal.[255]

Table 2b
Detection – General

Title: Create a National Deepfake Detection Think Tank

Type: Participation

Policy Owner: Ministry in Charge of Research and Development

Policy Targets: Academia, industry and government bodies.

Proposal: Research and development funding specially designed to better detect deepfake content through sociotechnical advancements.

Details: Allocate government funding to establish a centralized research and

[254] See *Add, remove, or view a trusted publisher*, MICROSOFT, https://support.microsoft.com/en-us/office/add-remove-or-view-a-trusted-publisher-87b3d5a3-b68c-4023-87c4-7cc78a44d7ed (last visited Apr. 17, 2024) for more information about the program.

[255] There is a legal requirement to have online real names in China. Rogier Creemers et al., *Translation: Cybersecurtiyy Law of the People's Republic of China (Effective June 1, 2017)*, DIGICHINA (June 29, 2018), https://digichina.stanford.edu/work/translation-cybersecurity-law-of-the-peoples-republic-of-china-effective-june-1-2017/.

development think tank to focus exclusively on advancing deepfake detection technologies. This initiative would join experts from academia, industry, and government agencies to collaborate on developing state-of-the-art algorithms and tools for automatically identifying and flagging deepfake content across various online platforms. Additionally, it incentivizes tech companies to share data and insights to improve detection capabilities, informing the research community. As an example, the Defense Advanced Research Projects Agency previously launched programs like the Media Forensics program, which aims to develop technologies for detecting manipulated media and deepfakes.[256] Other measures focusing on deepfakes include the Content Authenticity Initiative, which aims to add a layer of tamper-proof marking to all digital content to fight against issues related to content provenance and authenticity.[257]
Detection – Legal
Title: Legal Incentives/Regulation **Type**: Technology **Policy Owner**: Legislators/Administrative Body such as Tax Authority /Regional Authority **Policy Targets**: Content-based digital platforms allowing the production or distribution of synthetic media. **Proposal**: Implement legal incentives for content-based platforms to invest in deepfake detection technology. **Details**: Introduce tax incentives or liability protections for social media platforms and technology companies that invest in and deploy advanced deepfake detection technologies on their platforms. In addition to or alternatively, introduce state-certified trusted content moderators to detect deepfakes. Additionally, establish regulations that require platforms to transparently disclose their efforts in combating deepfakes to users. Failure to implement adequate detection measures could result in fines or penalties. In the EU AI Act, very large online platforms are required to invest in the detection of deepfakes ahead of EU elections in June 2024.[258]

[256] See *Media Forensics (MediFor) (Archived)*, DEF. ADVANCED RSCH. PROJECT AGENCY, https://www.darpa.mil/program/media-forensics (last visited Apr. 17, 2024) for program details.

[257] *See How it works*, CONTENT AUTHENTICITY INITIATIVE, https://contentauthenticity.org/how-it-works (last visited Apr. 17, 2024).

[258] Clothilde Goujard, *EU turns to Big Tech to help deepfake-proof election*, POLITICO (Feb. 7, 2024, 5:57 PM), https://www.politico.eu/article/eu-big-tech-help-deepfake-proof-election-2024/.

Table 2c Response – General
Title: Establish a Rapid Response Team for Deepfake Incidents **Type**: Participation **Policy Owner**: Cyber Security Enforcement Division of the Government **Policy Targets**: Social media platforms and potential victims of Deepfakes. **Proposal**: Create a dedicated task force comprising representatives from law enforcement agencies, cybersecurity experts, and digital forensics specialists to respond swiftly to incidents involving dissemination of malicious deepfakes. **Details**: This team would be equipped with the necessary tools and resources to investigate, track, and mitigate the spread of harmful deepfake content. Additionally, further develop protocols for coordinating with social media platforms and other relevant stakeholders to expedite the removal of deepfakes. The Cybersecurity and Infrastructure Security Agency (CISA) in the United States operates the National Cyber Incident Response Plan (NCIRP), which provides a framework for responding to cyber incidents.[259]
Response – Legal
Title: Rules to Establish an Emergency Response Protocol for Deepfake Threats **Type**: Information **Policy Owner**: Law Enforcement **Policy Targets**: Platform Owner **Proposal**: Enact emergency response protocols for deepfake threats. **Details**: Introduce legislation requiring government agencies to develop and regularly update emergency response protocols tailored to address threats posed by deepfake manipulation. These protocols would outline procedures for swiftly identifying, assessing, and mitigating the spread of malicious deepfake content, with provisions for coordination between law enforcement, intelligence agencies, and relevant stakeholders. E.g. In the UK, the National Cyber Security Centre (NCSC) operates under the Cyber Security Strategy and is tasked with responding to cyber incidents involving disinformation and digital manipulation.[260] Establishing similar protocols could address deepfake threats comprehensively.

[259] See *The National Cyber Incident Response Plan (NCIRP)*, CYBERSECURITY & INFRASTRUCTURE SEC. AGENCY, https://www.cisa.gov/resources-tools/resources/national-cyber-incident-response-plan-ncirp (revised on Oct. 20, 2023) for more information on the framework consult.

[260] See *Responding to a cyber incident – a guide for CEOs*, NAT'L CYBER SEC. CENTRE, https://www.ncsc.gov.uk/guidance/ceos-responding-cyber-incidents (last visited Apr. 17, 2024) for more information on this taskforce.

Table 2d Repair – General
Title: Establish a Deepfake Victim Support Program **Type**: Information and Participation **Policy Owner**: Ministry of Health/Ministry of Security **Policy Targets**: Deepfake victims and platforms. **Proposal**: Create a government-funded program to provide comprehensive support and resources for individuals targeted or victimized by deepfake attacks**.** **Details**: This program would offer legal assistance, mental health counselling, and digital identity protection services to help victims cope with the emotional and practical consequences of deepfake exploitation. . For example, the Cyber Civil Rights Initiative operates a helpline and advocacy program for victims of non-consensual pornography.[261] A Deepfake Victim Support Program could build upon this model to address the unique challenges faced by victims of deepfake manipulation.
Repair - Legal
Title: Introduce Legislation Providing Civil Remedies for Victims of Deepfake Exploitation **Type**: Information **Policy Owner**: Department of Law and Justice or Similar body **Policy Targets**: Victims and platform owners. **Proposal**: Enact laws that make deepfake manipulation a civil liability with special provision to allow for legal aid services as a way to empower claimants. **Details**: Enact laws that allow victims to seek civil remedies, including damages and injunctive relief, against perpetrators and distributors of malicious deepfake content. Establish legal mechanisms for expedited takedown requests and content removal procedures to minimize the civil harm caused to claimants. Promote legal aid and support services should be available to assist claimants in pursuing legal action. E.g. In India, the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021, include provisions for grievance redressal mechanisms and content takedown procedures for online platforms.[262] Similar legal frameworks could be expanded to address deepfake-related harm and provide recourse for victims.

[261] See CYBER CIVIL RIGHTS, https://cybercivilrights.org (last visited Apr. 17, 2024) for more information on this initiative.

[262] Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021, *retrieved from* https://mib.gov.in/sites/default/files/IT%28Intermediary%20Guidelines%20and%20Digital%20Media%20Ethics%20Code%29%20Rules%2C%202021%20English.pdf.

These concrete policy proposals aim to address different aspects of the deepfake phenomenon, from prevention and detection to response and victim support, and to mitigate the negative impact on individuals and society. These legal innovations aim to provide concrete mechanisms for preventing, detecting, responding to, and repairing the harm caused by deepfake manipulation. We draw upon examples from different countries' legal frameworks and approaches to combat this digital threat to democracy. The socio-legal enquiry on deepfakes along with these concrete policy innovation proposals should resonate with legislators, policymakers, and other regional stakeholders currently wrestling to mitigate potential harm from the malicious use of deepfake technology.