INTERNATIONAL LAW AND THE REGULATION OF DEEPFAKE TECHNOLOGY IN CROSS-BORDER POLITICAL MANIPULATION

Anisha Taneja*

ABSTRACT

The rise of high-tech synthetic media, often referred to as deepfakes, has created a novel form of international political interference. Deepfakes can create lifelike images, audio, and video of political figures stating and doing things that they have not said and done, and they enable state and non-state actors alike to engage in serious risks of interference in another state's political process, disparaging opponents or inflaming existing social tensions. Although it is an existing tool, human rights treaties, cybercrime conventions, regional regulatory efforts, and domestic laws collectively provide some structure, they are not sufficient to address cross-border deepfake campaigns that are strategic and specifically targeting political figures and political processes. The article proposes an international regulatory approach that has clear definitions, harmonised jurisdictional rules, and intermediary responsibility in consideration of human rights protections, standardised forensic collaboration, and targeted and actionable remedies for victims. This approach aims to balance free expression/political speech and the need to promote and protect political integrity, sovereignty, and the reputational and human rights of political actors in a growing global digital space.

Keywords: Deepfake, Artificial Media, Worldwide Law, Political Manipulation, Interference.

INTRODUCTION

Recent developments in artificial intelligence (AI) have led to easily accessible, low-cost tools to produce synthetic media in audio and video formats. Deepfakes can manufacture near-perfect representations of what they present to be authentic recordings. When leveraged to produce tailored messages about politicians, portraying them as saying inflammatory things,

*BA LLB, FIFTH YEAR, FAIRFIELD INSTITUTE OF MANAGEMENT AND TECHNOLOGY, KAPASHERA, NEW DELHI.

acknowledging wrongdoing, or committing criminal acts, deepfakes are thus vehicles of political deception and influence, potentially to the degree that they change how voters perceive candidates, disrupt political campaigns, and damage interstate relations. Unlike traditional propaganda, deepfakes can be produced at low cost, easily distributed and modified for social media, and formatted to avoid easy detection, resulting in challenges for national jurisdictions to respond or attribute the violation.[1]

This article will consider the central normative and practical question: How can international law facilitate the regulation and deter cross-border interference with domestic electoral politics without infringing on legitimate political expression? The article will consider a variety of key legal doctrines and how current domestic laws on deceptive media are serving their enforcement aims, analysing whether those approaches give context for an overall applicable legal analytical approach, and ultimately propose an agreed, rights-respecting international regulatory framework. Focusing on cross-border manipulations due to the harms to democratic processes, interstate comity, and electoral integrity heralded by deepfake manipulations in general, and, in political domains in particular, gives scope for harms and jurisdiction beyond any moulded domestic legal order.[2]

LITERATURE REVIEW

Deepfake interdisciplinary scholarship comprises the literature on technical detection, studies in political science, and legal scholarship on liability, speech rights, and privacy regulation. The technical research material demonstrates that there are quick advancements in generative models, leading to significant detection challenges for forensic clients. The political science literature highlights the asymmetrical harms: deepfakes can be deployed strategically to harm an individual's reputation in order to manipulate an important voting population, deepen pre-existing social cleavages, and deepen distrust among voters in public establishments. The legal literature has mapped the various potential legal responses, such as potential defamation, privacy, and cybercrime statutes, applicable election law, and other governance protocols applicable to Platforms; however, scholars indicate that several significant issues arise when

[1] Danielle Keats Citron and Robert Chesney, 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security' (2019) 107 California Law Review 1753.
[2] Vivek Krishnamurthy, 'Synthetic Media and International Law: Toward a Normative Framework' (2023) 117 American Journal of International Law Unbound 52.

the harmful conduct could occur in different countries, and when the violence relies upon anonymity and algorithmic dissemination.

Some central themes are emerging across the different literature. First, a variety of terms and legal definitions remain unresolved: for example, the term deepfake is employed widely to denote any type of synthetically altered media or image; however, to determine the need for regulation more in-depth categories require scrutiny, such as nonconsensual intimate seeing, voice cloning, and impersonation of political figures. Second, scholars note concerns about jurisdiction being fractured when the violent act occurs inside one state, hosted in another state, and viewed from multiple states.[3] Third, the intermediary role of platforms is acknowledged, but adversarial to some degree in regards to regulation, due to the question of their responsibilities regarding content (if they have any) moderation, and whether poorly tailored governmental rules could detain free speech.[4] At the same time, relatedly, broader protections may enable harm. In conclusion, there is consensus in the literature in favour of some form of multinational cooperation, which includes the technical aspects of forensic standards and sharing of evidence, the legal issues of mutual legal assistance and harmonised offences, and also emphasises strong procedural protections for expression and due process. This article builds upon those foundations by emphasising how international law opens the door to harmonious, rights-conscious responses.

METHODOLOGY

The study uses doctrinal and comparative legal research methods with normative proposal-making. Main legal materials analysed include international treaties and instruments related to cyber operations and the protection of human rights, regional regulatory initiatives, and representative domestic statutes focused on synthetic media, content moderation, and electoral integrity. Secondary materials include scholarly articles, policy reports, and technical literature related to the generation and detection of deepfakes. The comparative analysis focuses on a selection of jurisdictions: the European Union (as a supranational regulator); the United States (with constitutional barriers to speech regulation); India (an emerging policy response in a democratic polity with a complicated media ecology); and illustrative examples from various national approaches to establish legal design and enforcement capacity, strengths, tensions, and

[3] Michael N Schmitt (ed), Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations (2nd edn, Cambridge University Press 2017).
[4] European Commission, Proposal for a Regulation on a Single Market for Digital Services (Digital Services Act) COM (2020) 825 final.

gaps. The study maps legal authorities to the habitual scenarios of deepfake use involving cross-border political manipulation of actors and sectors (for example, a foreign-produced deepfake video geared toward election audiences, or unexpectedly coordinated campaigns to discredit opposition poses). Based on this mapping, the paper provides synthesised principles to use for an international framework, calibrated to operational realities and normative limitations.

INTERNATIONAL LEGAL FRAMEWORKS ADDRESSING DEEPFAKE POLITICAL MANIPULATION

Treaties and Principles of Human Rights: International human rights law imposes dual obligations of protecting freedom of expression and protecting individuals and public order from harm. Instruments like the International Covenant on Civil and Political Rights (ICCPR) protect expression while permitting lawful and proportionate restrictions for public order and reputation.[5] Therefore, states must carry out measures that specifically comply with manipulative deepfakes, but must be careful not to overreach and impose restrictions that will censor legitimate expression.

Cybercrime and Legal Assistance: International instruments concerning cybercrime (particularly regional instruments) and mutual legal assistance frameworks that provide mechanisms for cross-border investigation, exchange of evidence, and pursuit of prosecution for offences such as identity theft, fraud, and illicit access can also apply.[6] While these instruments were not drafted for deepfakes specifically, their procedural infrastructure can be utilised for cross-border takedown requests and attribution investigations.

Independence, Non-Interference, and Responsibility of States: If individuals, groups, or organizations carry out deepfakes at the behest of a state or with a state's tacit approval, they may violate the norm of non-interference into the sovereign affairs of states as established in international law, and states may also violate their international obligations not to interfere in the internal political affairs of others states.[7] International law on state responsibility would, in principle, provide the basis for claims in situations where deepfakes that were substantive in

[5] International Covenant on Civil and Political Rights (adopted 16 December 1966, entered into force 23 March 1976) 999 UNTS 171 (ICCPR).
[6] Convention on Cybercrime (Budapest Convention) (adopted 23 November 2001, entered into force 1 July 2004) ETS No 185.
[7] International Law Commission, Draft Articles on Responsibility of States for Internationally Wrongful Acts (2001) UN Doc A/56/10.

character, not merely simulative, constitute a wrongful act of the state that had sufficient political effect within the target state.

Considerations for Regional Regulatory Mechanisms: Contextual developments in the area of platform governance, the transparency of political advertising, and policy frameworks around AI risk can shape the environment in which deepfakes spread. These developments would/should generally favour transparency, accountability, and human rights impact assessment.[8]

NATIONAL COMPARISONS

European Union: The EU's regulatory approach has consisted of regulations for digital services and sectoral AI regulations. The transparency obligations related to political advertising and the beginnings of an EU AI governance regime imposing obligations to undertake risk assessments and guarantees of traceability for high-risk systems. Importantly, the EU model is based on compliance with rights and imposes obligations on platforms to remove harmful content, which was done while ensuring procedural safeguards.[9]

United States: Responses in the U.S. have been fragmented, wherein some states have laws that criminalise the performance of certain deepfake conduct (i.e., producing sexually explicit imagery without consent or impersonating someone for election purposes), but federal law does not, for the most part. Most federal legislation and law enforcement deals with criminal fraud and election interference, as well as targeted disinformation within the context of other offences. The First Amendment of the U.S. Constitution limits the ability of government actors to restrict what content can be used, so remedies emphasise private causes of action, platform policies, and/or some modest disclosure requirement, but do not generally include a broad criminal prohibition.

India: Recognising the context of India's legal framework on matters of online safety, it includes laws in information technology and civil and criminal remedies. In India, the scale and diversity of the information ecosystem, policy priorities emphasise platform responsibilities and digital literacy. The engagement of the judiciary and a rule-making agenda

[8] Council of Europe, Recommendation CM/Rec(2022)11 of the Committee of Ministers to Member States on Principles for Media and Communication Governance (16 November 2022).
[9] European Commission, Proposal for a Regulation on a Single Market for Digital Services (Digital Services Act) COM (2020) 825 final.

from the administrative level signals a practical effort to moderate constitutional commitments to freedom of speech in the interests of social order and true individual reputation.

Other Jurisdictions: Several jurisdictions have either enacted or are considering regulations on likeness rights, voice cloning, and election-related manipulation that deal with a range of approaches: criminal sanctions in some jurisdictions, and administrative fines and transparency obligations in others.[10]

RESULTS

The results from the comparative mapping present several recurring findings, as follows: First, domestic laws may be able to address harms which are located in a discrete location, but they become complex when there is cross-border production and hosting. Second, platform policies are the de facto gatekeepers, but each platform has its private rules, which do not apply universally and have inconsistent application across jurisdictions.[11] Third, the forensic and attribution capacity is disparate: some advanced states and private businesses have forensic tools available for detection, but there continues to be challenges to standardisation and the admissibility of digital evidence.[12] Fourth, the way human rights restrict activity--particularly in relation to political speech--is important for the shape of regulations. Contemporary high scrutiny by the courts in some states has decreased [an appetite] for general content suppression. Ultimately, these findings suggest that a sustained effort for patchwork legal responses will continue to provide marginal relief for reactively and proactively dealing with a range of systemic vulnerabilities sometimes borne of malicious actors operating transnationally.

DISCUSSION

The convergence of deepfakes and cross-border political influence presents a governance problem that is technical, legal, and geopolitical at once. The key legal challenge is to create obligations that provide some minimal deterrence against malign actors and encourage remediation, while being specific enough not to chill legitimate expression and satire. The

[10] Ganna Pogrebna and Oles Andriychuk, 'The Governance of Deepfakes in European Law' (2022) 13 European Journal of Risk Regulation 360.
[11] See Jack Goldsmith & Tim Wu, Who Controls the Internet? Illusions of a Borderless World (Oxford Univ. Press 2006).
[12] See Orin S. Kerr, Computer Crime Law 6th ed. (West Academic 2022) (discussing digital forensics, admissibility standards, and evidentiary challenges).

limits of national sovereignty are stark: a state that has been a target of foreign deep fake interference must rely on diplomatic channels, mutual legal assistance, and/or extraterritorial regulation of platforms and intermediaries operating within its territorial boundaries.[13] These are the mechanisms for legitimate responsive action, however slow or politically inconvenient. Procedurally, platforms (viz., social media and search engines) mediate much of the response through content moderation and labelling. Yet, in the realm of anything private, this raises questions of legitimacy, due process, and consistency.

An international framework should focus on aligned definitions and process standards that allow for rapid joint action. With an international framework, we mean agreed-upon processes that ensure accountability and that guide coordination of international partner actions on deepfakes (and other manipulative media). That might mean, for example, model forensic verification protocols that tie actions to an agreed and standardised evidentiary threshold, and model processes for rapid take-down requests (or media removal) shared between states, to respond to deepfakes. Second, we need a calibrated approach to intermediary liability: conditional obligations (for example, notice-and-action; transparency reporting; targeted disclosure of political ads) that are accompanied by safe harbours for good-faith attempts to remove objectionable fraudulent media can align actions between jurisdictions, without stifling freedom of expression or unduly relying on intermediaries to make determinations, therefore leaning closer to unwanted censorship. Third, human rights-based protections should be incorporated: any (government) obligations should be tied to proportionality, contestability (meaning affected parties can argue a takedown decision), and narrowly tailored to the context of manipulation related to political processes.[14] Finally, we know that disregard for civil rights means that state actors can weaponise deepfakes in a range of ways, and therefore need diplomatic norms, and potentially state responsibility mechanisms where appropriate, to deter and redress state interference in democracy (or political processes), however the interference occurs.[15]

SUGGESTED STRUCTURE FOR MULTINATIONAL OVERSIGHT

[13] See Kristen E. Eichensehr, Cyber war & International Law: Sovereignty, Intervention, and the Attribution Problem, 95 Tex. L. Rev. 151 (2016).
[14] See International Covenant on Civil and Political Rights, Dec. 16, 1966, 999 U.N.T.S. 171 (requiring restrictions on expression to meet tests of legality, necessity, and proportionality).
[15] See Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations (Michael N. Schmitt ed., Cambridge Univ. Press 2017) (discussing state responsibility for cyber and information operations).

Fundamental Definitions and Coverage: To employ an international standardised classification scheme that distinguishes between-

- Intimate deepfakes that are non-consensual;
- Political actors impersonated for purposes of deception;
- Synthetic content utilised in political advertising that is clearly identified to be political advertising and
- satire/parody. Regulatory attention would be focused on categories (b) and (c) when the aim is to mislead electorates or influence political processes.

Obligations and Assignment of Responsibility:

- **Originators/Operators:** Making willful creation and distribution of deepfakes intended to deceive electorates a crime, where demonstrable material harm exists.
- **Platforms/Intermediaries:** Require clear moderation policies for political content, rapid notice and takedown of maliciously deepfaked content that is independently verified and public reports about political-content incidents. Exemption from liability would be contingent upon the good-faith implementation of these recommendations.
- **Model Providers/Developers:** Require risk assessments, watermarking or provenance metadata when feasible, and access controls to mitigate against misuse.[16]

Legal Domains and Collaboration:

- Develop swift pathways for mutual legal assistance in synthetic media with agreements to use standardised forms for emergency takedown and evidence preservation.
- Acknowledge an effects-based jurisdiction where manipulative content results in actual electoral or political damage in the affected state.
- Establish an international technical centre (or network) to assist in the attribution of evidence certification and capacity building for destination states with weak forensic capacity.[17]

[16] See European Commission, Proposal for an Artificial Intelligence Act, COM(2021) 206 final (laying down obligations for AI developers, including risk assessments and technical safeguards).
[17] See Budapest Convention on Cybercrime, Nov. 23, 2001, ETS No. 185 (establishing frameworks for mutual legal assistance and cross-border cooperation on digital evidence).

Remedies and Enforcement:

- Provide civil remedies (injunctive relief, damages) for targeted candidates and electoral bodies.
- Criminally sanction only organised malicious campaigns, especially state-backed interference in elections or coordinated foreign influence campaigns- with appropriate due process protections.
- Administrative fines and corrective order options for non-compliant platforms.

Federations of human rights: All measures must comply with international human rights standards: legality, necessity, proportionality, and the right to challenge actions of the state or platform before an independent adjudicator. There is a special obligation to safeguard bona fide political speech, artistic expression, and satire.

THE NEW ROLE OF DIGITAL PROVENANCE AND STANDARDS

One underappreciated but increasingly important aspect of deepfake governance is the creation of global standards regarding digital provenance. Digital provenance is a technological method of attaching verifiable metadata or cryptographic 'watermarks' to authentic media at the time of its creation. Current efforts such as the Coalition for Content Provenance and Authenticity (C2PA) and other similar open-standard initiatives are attempting to create a chain of custody for images, audio and video files, which would allow viewers, platforms and courts to verify where the file came from and what has happened to it since its creation.[18]

From the perspective of international law, these emerging standards represent two abounding opportunities: First, they are a non-regulatory, normative addition to existing legal frameworks as a mechanism for evidentiary reliability. Courts and international dispute resolution committees could rely on standardised provenance metadata when deciding if cross-border political manipulation of a target occurred.

Additionally, they may also promote cooperative law enforcement, in which states would agree as a treaty or soft-law instrument, to recognise that provenance protocols for official political speech, including campaign ads and speeches, serve as a legitimate basis against a finding of the content being a malicious deepfake. The possible applicability of these approaches also

[18] See Coalition for Content Provenance and Authenticity (C2PA), Technical Specifications v2.0 (2023), https://c2pa.org

implicates challenges, including interoperability across jurisdictions, issues of privacy for journalists and dissidents, and possible increased technological impediments to implementation for developing states.[19]

Therefore, international regulators will need to navigate between the evidentiary utility of provenance and equitable access opportunities, including legally binding data-protection guarantees. A treaty/protocol of the future on synthetic media could include specific provisions to advance open provenance standards and require cooperative capacity building to ensure that the emerging new layer of digital trust is not exclusionary to smaller, less technologically sophisticated, or less resourced states.

CONCLUSION

Deepfake technology has created a promising, economical way of engaging in political manipulation internationally that challenges the integrity of elections, political reputation, and trust amongst governments. Existing sources of international law, particularly within human rights frameworks, limit overreach; cybercrime and mutual legal assistance mechanisms provide procedural pathways; and regulatory measures within regions can assert high standards. However, the existing universe of available options is clearly inadequate to respond to a coordinated, cross-border campaign using deepfakes. A globally effective response requires the development of a definition, clearly defined obligations for originators, platforms, and developers, and an expedited cross-border cooperation mechanism to trace accountability for and removal of the offending content, all the while providing remedies that respect democratic participation and freedom of expression. A collaborative treaty or protocol, anchored in human rights norms and expanded through technical standards and capacity-building, is arguably the most realistic approach. A multi-governmental rights-based regime must be sensitive to the political realities of state interaction, as well as the always-evolving landscape of synthetic media that will shape our interactions with these technologies.[20]

[19] See U.N. Human Rights Council, Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, U.N. Doc. A/HRC/44/49 (Apr. 23, 2020) (highlighting risks of surveillance, privacy intrusions, and inequality in access to digital authentication technologies).
[20] See Michael N. Schmitt, Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations (Cambridge Univ. Press 2017) (discussing legal obligations, state responsibility, and cooperative measures in cross-border digital operations).