THE LEGAL IMPLICATIONS OF DEEPFAKE TECHNOLOGYCOPYRIGHT INFRINGEMENT AND CRIMINAL LIABILITY

SANJAY.S

V BA.LLB (hons) – B section

Vels Institute of Science, Technology & Advanced Studies (VISTAS)

V.SWATI

Assistant professor of law

Vels Institute of Science, Technology & Advanced Studies (VISTAS)

ABSTRACT

Deepfake technology, grounded in generative adversarial networks (GANs) and latent diffusion models, has emerged as a powerful instrument of misinformation, identity theft, non- consensual intimate imagery (NCII), and commercial fraud. This paper conducts a doctrinal, comparative, and empirical examination of deepfake-related legal issues, focusing on two principal domains: copyright infringement and criminal liability. Drawing upon statutory frameworks, case law, and emerging legislation across the United States, United Kingdom, European Union, Australia, India, and select Asian jurisdictions, the paper argues that existing legal regimes are insufficiently equipped to address the harms generated by synthetic media.

The paper examines the technological foundations of deepfake generation, the copyright dimensions of AI training data, authorship of AI-generated outputs, performer rights, platform liability, and criminal offences ranging from NCII to fraud, electoral interference, and AI- generated child sexual abuse material (CSAM). A comparative jurisdictional analysis reveals fragmented and inconsistent regulatory responses. The paper concludes by proposing a multi- tiered regulatory framework combining technical provenance standards, copyright reform through extended collective licensing, bespoke criminal offences, reformed platform liability rules, and international treaty coordination.

Keywords*: Deepfake, Synthetic Media, Generative Adversarial Networks, Copyright Infringement, Criminal Liability, Artificial Intelligence, Regulatory Framework, Digital Identity.*

INTRODUCTION

The intersection of artificial intelligence and digital media production has generated one of the most legally and ethically significant technologies of the twenty-first century: the deepfake. The term, a portmanteau of ‘deep learning’ and ‘fake,’ gained public prominence in 2017 when a Reddit user uploaded non-consensual intimate imagery (NCII) of celebrities using AI-assisted face-swap technology. Within months, the technology proliferated into political disinformation, entertainment visual effects, corporate fraud, and impersonation campaigns of unprecedented sophistication.

OBJECT AND SCOPE OF THE STUDY

The principal object of this study is to examine the legal implications of deepfake technology with specific reference to two intersecting domains of law: copyright infringement and criminal liability. The study surveys the technological foundations of deepfake generation, the copyright dimensions of AI training data, authorship of AI-generated outputs, performers' rights, platform liability, and criminal offences ranging from non-consensual intimate imagery to fraud, electoral interference, and AI-generated child sexual abuse material.

The scope of the study is both doctrinal and comparative, encompassing the statutory and judicial frameworks of the United States, United Kingdom, European Union, Australia, India, and select Asian jurisdictions. The study further extends to an assessment of proposed and enacted legislative reforms and concludes with a normative proposal for a multi-tiered regulatory framework. The paper reflects the state of law as of early 2026.

RESEARCH PROBLEM

This paper conceptualises the central problem as the "synthesis gap": a structural and rapidly widening disparity between the sophistication, accessibility, and scale of synthetic media production, and the capacity of existing legal frameworks to provide meaningful remedies, accountability, and deterrence. This gap is not merely technological but deeply normative and institutional, manifesting across multiple, interrelated dimensions. First, core statutory concepts such as "work," "author," "originality," "likeness," and "communication to the public" were drafted in an era that assumed human creativity as the foundation of authorship. As a result, they struggle to accommodate content generated wholly or substantially by autonomous or semi-autonomous artificial intelligence systems, creating uncertainty regarding ownership, liability, and the scope of protection. The absence of clear attribution standards further complicates questions of responsibility when harm is caused by synthetic media.

Second, the defining feature of deepfakes—their increasing hyper-realism—significantly amplifies their potential for harm. Unlike earlier forms of manipulated media, advanced synthetic content can convincingly replicate voice, facial expressions, and behavioural patterns, making it difficult even for trained observers to distinguish between authentic and fabricated material. This realism undermines evidentiary reliability, complicates the burden of proof in legal proceedings, and enables new forms of deception, including reputational harm, fraud, political misinformation, and non-consensual sexual exploitation. Consequently, traditional legal doctrines premised on identifiable falsity or clear intent become harder to apply in practice.

Third, the speed, scale, and transnational nature of digital dissemination exacerbate regulatory challenges. Synthetic media can be created, uploaded, and virally circulated across multiple jurisdictions within minutes, often through anonymous or pseudonymous accounts. This diffusion outpaces the territorial limits of national legal systems, leading to enforcement gaps, jurisdictional conflicts, and difficulties in securing timely takedowns or remedies. Intermediary platforms, while central to the distribution ecosystem, operate under varying standards of liability and content moderation, further fragmenting accountability. Taken together, these dynamics reveal a profound mismatch between technological capability and legal response,

underscoring the urgent need for adaptive, harmonised, and forward-looking regulatory strategies

RESEARCH QUESTION

The study is organised around the following principal research questions:

Whether existing copyright law frameworks in major jurisdictions adequately address the reproduction of copyrighted works in AI training processes and the authorship of AI- generated deepfake outputs, and if not, what legislative or doctrinal reforms are required?

Whether existing criminal law provisions in the examined jurisdictions provide sufficient and coherent liability for deepfake-related conduct including non-consensual intimate imagery, fraud, defamation, electoral interference, and the generation of child sexual abuse material?

Whether current platform liability regimes, particularly Section 230 of the Communications Decency Act and analogous frameworks, impose adequate obligations on online platforms to detect, remove, and prevent the re-publication of harmful deepfake content?

What principles should govern a comprehensive, technology-neutral, and internationally coherent regulatory framework for deepfake synthetic media, and what specific legislative, technical, and institutional measures are required to implement such a framework?

HYPOTHESIS

The study proceeds from the hypothesis that existing legal regimes across all examined jurisdictions are insufficiently equipped to address the harms generated by synthetic media, and that this inadequacy is systemic rather than incidental.

(a) that AI training data reproduction is not adequately addressed by existing fair use and text- and-data-mining exceptions, and requires legislative resolution through a sui generis licensing framework;

(b) that criminal liability for deepfake conduct is fragmented, jurisdictionally inconsistent, and deficient in provisions addressing the creation of harmful synthetic media independently of distribution;

(c) that platform liability frameworks create inadequate incentives for proactive content moderation and require reform towards a duty-of-care standard; and (d) that effective regulation of deepfake harms ultimately requires coordinated international treaty mechanisms to prevent regulatory arbitrage and ensure cross-border enforcement.

METHODOLOGY

The methodology adopted in this paper is doctrinal and comparative. The doctrinal dimension involves systematic analysis of primary legal sources statutes, judicial decisions, legislative committee reports, regulatory guidance, and treaty instruments relevant to deepfake copyright infringement and criminal liability across the examined jurisdictions. The comparative dimension involves a structured assessment of the regulatory responses of the United States, United Kingdom, European Union, Australia, India, and China, identifying points of convergence, divergence, and lacunae.

The paper further draws upon interdisciplinary scholarly literature in law, computer science, communications, and political science to contextualise legal analysis within the technical and social dimensions of deepfake harms. Secondary empirical sources, including government reports, platform transparency data, and civil society research, are used to ground normative proposals in evidential assessments of regulatory efficacy.

The normative framework proposed in Chapter Six is developed using a principled analytical approach, identifying regulatory objectives and evaluating the fit between proposed

BACKGROUND AND CONTEXT

The intersection of artificial intelligence and digital media production has generated one of the most legally and ethically significant technologies of the twenty-first century: the deepfake. The term, a portmanteau of 'deep learning' and 'fake,' gained public prominence in 2017 when a Reddit user uploaded non-consensual intimate imagery (NCII) of celebrities using AI-assisted face-swap technology (Chesney & Citron, 2019). Within months, the technology proliferated into political disinformation, entertainment visual effects, corporate fraud, and impersonation campaigns of unprecedented sophistication.

A deepfake, for purposes of this paper, refers to any audio, visual, or audiovisual work created or substantially altered using machine learning algorithms such that the resultant content falsely represents a real person, or constitutes a substantially new work generated through unauthorised use of biometric data. This definition distinguishes deepfakes from 'cheapfakes' (traditional video editing), AI-generated fictional characters, and consensual AI- assisted creative works each of which raises distinct legal concerns.

The social consequences of democratised synthetic media are profound. Political deepfakes have been deployed to fabricate surrender announcements and inflame social divisions (Fallis, 2021). In the commercial sphere, AI voice cloning enabled fraudsters to impersonate executives, resulting in losses exceeding USD 35 million in a single reported incident (Stupp, 2019). NCII deepfakes have become a pervasive form of technology- facilitated sexual violence, disproportionately targeting women and girls (Centre for Countering Digital Hate, 2023).

GENERATIVE ADVERSARIAL NETWORKS AND LATENT DIFFUSION MODELS

The principal architecture underlying video deepfakes is the generative adversarial network (GAN), introduced by Goodfellow et al. (2014). A GAN comprises two competing neural networks a generator that synthesises realistic content and a discriminator that attempts to identify synthetic content trained iteratively on large datasets of authentic material. Critically for legal analysis, GAN training requires vast quantities of data, including potentially copyrighted images, videos, and audio recordings, raising threshold questions about whether training itself constitutes infringement.

Later-generation deepfake systems employ latent diffusion models (LDMs), which operate in a compressed latent space to generate high-resolution imagery with comparatively modest computational resources (Rombach et al., 2022). LDMs such as Stable Diffusion, typically trained on internet-scale

datasets, can be fine-tuned on a small set of reference images to reproduce the likeness of any individual with an online presence. Their legal implications are threefold: the scale of training data use raises copyright concerns; fine-tuning enables systematic misappropriation of individual likenesses; and the copyright status of LDM outputs implicates unsettled authorship doctrine.

VOICE CLONING AND DETECTION LIMITATIONS

Voice cloning technology has advanced to the point where commercially available systems such as Eleven Labs and OpenAI's Voice Engine can generate convincing synthetic voices from as little as three seconds of source audio (Khanjani et al., 2023). These tools have been deployed in CEO impersonation fraud, political disinformation, and NCII audio. The landmark Ninth Circuit decision in Midler v. Ford Motor Co. (1988), which recognised a celebrity's unique voice as protectable against commercial imitation, raises complex questions when the imitator is not a human performer but an AI system.

Deepfake detection remains technically limited. Current detection methods biological signal analysis, frequency domain analysis, and deep learning-based classifiers exhibit false positive rates of ten to forty percent depending on the technique and the sophistication of the deepfake under examination (Mirsky & Lee, 2021). This evidentiary unreliability creates substantial barriers to both criminal prosecution and copyright enforcement. A further concern is the 'liar's dividend': the availability of deepfake technology enables bad actors to dismiss authentic recordings as fabricated, undermining the epistemic function of audio-visual evidence generally (Chesney & Citron, 2019).

Copyright law in the principal jurisdictions examined provides protection to original works of authorship fixed in a tangible medium. The United States Copyright Act of 1976 vests exclusive rights of reproduction, derivation, distribution, public performance, and display in human authors. Equivalent frameworks exist under the UK Copyright, Designs and Patents Act 1988, the EU Information Society Directive (2001/29/EC), Australia's Copyright Act 1968, and India's Copyright Act 1957.

AI TRAINING AS REPRODUCTION

The most foundational copyright question in the deepfake context is whether training AI models on copyrighted works constitutes infringing reproduction. Training data ingestion involves storing transient copies of copyrighted material in computer memory and encoding expressive characteristics within model parameters processes that at least arguably constitute reproduction under copyright law. Courts in the United States are actively considering this question. In Andersen v. Stability AI (N.D. Cal., 2023), the district court declined to dismiss a reproduction claim on training data, finding that plaintiffs had adequately alleged that the training process involved storage of compressed representations of their copyrighted works. Parallel proceedings in Getty Images (US) Inc. v. Stability AI Ltd. (D. Del., 2023) involve the systematic use of a licensed photo library for AI training without authorisation.

AI developers have invoked the fair use defence under 17 U.S.C. 107, relying particularly on the transformativeness factor as elaborated in Authors Guild v. Google, Inc. (2d Cir., 2015), which upheld the digitisation of books for search indexing as a transformative use. However, the Supreme Court's decision in Andy Warhol Foundation for the Visual Arts, Inc.

v. Goldsmith (2023) substantially narrowed transformativeness analysis, requiring that an allegedly transformative use have a genuinely independent expressive purpose rather than merely commercial reuse of copyrighted material. Deepfake systems trained on celebrity images for commercial exploitation face significant difficulty satisfying this narrowed standard.

DERIVATIVE WORKS AND AUTHORSHIP

Deepfake outputs generated by overprinting copyrighted audiovisual content with synthesised faces may constitute infringing derivative works under 17 U.S.C. § 101, which defines derivative works as works 'recast, transformed or adapted' from pre-existing material. More complex questions arise where deepfakes are generated wholly by AI without direct copying from specific source frames, requiring courts to assess whether model outputs embody expressive characteristics attributable to identifiable training works.

On authorship of AI-generated deepfakes, the U.S. Copyright Office and federal courts have consistently maintained that copyright requires human authorship. In Thaler v. Perlmutter (D.D.C., 2023), the court held that autonomously AI-generated images cannot receive copyright protection absent meaningful human creative contribution. The Copyright Office's 2023 guidance on AI-generated works adopted a case-by-case approach, requiring that human creative input be 'more than de minimis' to support copyright protection. This creates a practical paradox: a highly realistic deepfake video may be ineligible for copyright protection, but its creation may nonetheless constitute actionable infringement of underlying works and give rise to criminal liability.

PERFORMERS' RIGHTS AND RIGHT OF PUBLICITY

Beyond traditional copyright, deepfakes engage performers' rights and the right of publicity. In the EU, the Term of Protection Directive (2011/77/EU) and the Information Society Directive confer economic and moral rights on performers in their performances, independently of copyright in underlying works. Deepfakes that realistically simulate a performer's gestures, vocal qualities, and expressive style may infringe these performers' rights, as highlighted by the 2023 SAG-AFTRA strike, which secured contractual protections against AI performer duplication.

The right of publicity, primarily a state law right in the United States, protects individuals' commercial interests in their name, likeness, voice, and persona. The Ninth Circuit's decisions in Midler v. Ford Motor Co. (1988) and White v. Samsung Electronics America, Inc. (1992) established that celebrity voice imitation and persona evocation for commercial purposes can give rise to right of publicity claims even without direct copyright infringement. Applied to AI-generated deepfakes, any commercial deepfake that evokes a real person's image or identity would likely satisfy these standards.

PLATFORM LIABILITY

Section 230 of the Communications Decency Act provides U.S. platforms with near- absolute immunity from civil liability for user-generated deepfake content, subject to exceptions for federal criminal prosecution and intellectual property claims. Copyright claims against deepfake-hosting platforms

are instead governed by the DMCA safe harbour framework, which requires platforms to respond expeditiously to valid takedown notices. The reactive, notice-based model has proven inadequate for deepfake copyright infringement: detection is technically challenging, 'whack-a-mole' dynamics undermine takedown efficacy, and the burden of enforcement falls disproportionately on rights holders.

The EU Digital Services Act (DSA), fully effective in 2024, adopts a more interventionist model, requiring large online platforms to conduct systemic risk assessments covering deepfake disinformation and to implement proactive mitigation measures. requirements, necessitating reconciliation as platforms operate simultaneously under both regimes.

PROPOSED COPYRIGHT REFORMS

Legislative proposals in multiple jurisdictions seek to address these doctrinal inadequacies. The proposed NO FAKES Act (2023/2025) in the United States would create a federal civil right of action for unauthorised AI-generated simulacra of voice and visual likeness, supplementing existing copyright and right of publicity remedies. In the EU, the AI Act (Regulation (EU) 2024/1689) requires general-purpose AI model providers to publish detailed summaries of training data and to comply with copyright law, while Article 4 of the Copyright in the Digital Single Market Directive (2019/790/EU) enables rights holders to opt out of AI training use, though the practical efficacy of this opt-out mechanism remains contested.

NON-CONSENSUAL INTIMATE IMAGERY

NCII deepfakes constitute the largest category of deepfake criminal conduct by content volume. Research by the Centre for Countering Digital Hate (2023) found that deepfake pornography dominates the deepfake content ecosystem, with women and minors comprising the majority of identified victims.

Documented harms include post-traumatic stress disorder, depression, occupational disruption, and severe social isolation. Despite these serious harms, criminal prosecution rates have remained low due to definitional gaps in existing statutes and challenging evidentiary requirements.

Legislative responses have been uneven but accelerating. In the United States, approximately thirty-five states have enacted specific NCII deepfake statutes, with major jurisdictions including California (A.B. 602), Texas (H.B. 4337), Virginia, and New York leading adoption. The UK Online Safety Act 2023 created a new offence of sharing intimate images without consent, extended by the Criminal Justice Bill 2024 to criminalise the act of creation itself, without requiring proof of distribution. This represents a significant doctrinal advance, recognising that the harm of NCII deepfake creation arises independently of dissemination.

FRAUD AND FINANCIAL CRIME

Deepfake audio and video are increasingly deployed in sophisticated financial fraud.

received an AI-cloned telephone call purportedly from his company's German CEO (Stupp, 2019). By 2024, a Hong Kong multinational suffered losses of approximately USD 25 million after employees were deceived in a video conference in which all other apparent participants, including the company's CFO,

were AI-generated deepfakes (Tidy, 2024). These cases demonstrate that deepfake fraud has rendered traditional voice and visual verification mechanisms inadequate.

Federal fraud statutes in the United States wire fraud (18 U.S.C. § 1343), bank fraud (18 U.S.C. § 1344), and identity fraud (18 U.S.C. § 1028) provide a general framework applicable to deepfake fraud schemes. The question of whether an AI-generated biometric imitation constitutes a 'means of identification' under the Identity Theft Enforcement and Restitution Act has not been definitively resolved by courts, though academic consensus suggests current statutory language is sufficiently broad to encompass deepfake impersonation (Ganz, 2021). Financial intelligence authorities have additionally flagged deepfake-based circumvention of KYC/AML identity verification as a growing money laundering risk, with the Basel Committee on Banking Supervision specifically addressing this threat in 2023 guidance.

DEFAMATION AND ELECTORAL INTERFERENCE

Deepfakes that falsely represent real individuals making statements or performing discrediting actions constitute prima facie defamation in most common law jurisdictions, combining elements of false statement, publication, and reputational harm. Practical challenges in deepfake defamation cases include anonymous perpetration, definitional uncertainty about whether AI-generated outputs constitute 'statements,' and characterisation difficulties in jurisdictions distinguishing libel and slander.

Electoral deepfakes pose acute threats to democratic integrity. The distribution of a deepfake audio attributed to President Biden during the January 2024 New Hampshire primary illustrates the operational deployment of this technology in electoral interference. State-level legislative responses in the United States California (A.B. 730), Texas (S.B. 751), Minnesota, and Washington impose criminal or civil liability for deepfake use in electoral communications within pre-election periods, though First Amendment challenges to these restrictions have produced divergent judicial outcomes.

CYBERSTALKING, AND NATIONAL SECURITY

AI-generated child sexual abuse material (CSAM) represents the most ethically acute deepfake criminal law issue. Following Ashcroft v. Free Speech Coalition (2002), which restricted prohibition of purely virtual CSAM in the United States, the PROTECT Act of 2003 criminalised obscene AI-generated representations of child sexual abuse. Federal circuits have upheld this provision as applied to deepfake CSAM. The UK (Protection of Children Act 1978; Coroners and Justice Act 2009), Australia, Canada, and EU member states adopt broader frameworks criminalising virtual and pseudo-photographic CSAM irrespective of real-victim involvement — an approach warranting adoption by jurisdictions that retain the Ashcroft gap.

Deepfake cyberstalking is prosecutable under 18 U.S.C. § 2261A in the United States, with analogous provisions in the UK Protection from Harassment Act 1997 and the Domestic Abuse Act 2021. Intelligence agencies across the Five Eyes alliance have identified AI- generated deepfakes as a major vector for foreign influence operations and national security disinformation, with existing national security

statutes the Espionage Act, CFAA, and Foreign Agents Registration Act providing an applicable but imperfectly fitted legal framework.

EVIDENTIARY CHALLENGES

Deepfake criminal prosecutions face three categories of evidentiary difficulty. First, authentication: the existence of sophisticated deepfakes enables defendants to contest the authenticity of genuine digital evidence, undermining the epistemic foundations of audio- visual proof. Second, expert evidence: deepfake forensic examination is technically demanding, methodologically contested, and subject to exclusion under Daubert standards or equivalent reliability requirements, necessitating the development of standardised qualification criteria for expert witnesses in this domain. Third, digital chain of custody: deepfake evidence gathered from online platforms requires rigorous documentation to satisfy continuity of evidence requirements, particularly given the risk of metadata alteration during collection.

COMPARATIVE JURISDICTIONAL ANALYSIS

UNITED STATES

The United States lacks comprehensive federal deepfake legislation, relying instead on a patchwork of adapted general statutes and state-level innovations. At the federal level, the National Defense Authorization Act for Fiscal Year 2020 mandated deepfake threat reporting and DARPA detection research. infringement and enhanced penalties for wilful conduct. State legislatures have been significantly more active: as of early 2026, approximately thirty-five states have enacted NCII deepfake statutes, and states including California, Texas, and Washington have enacted electoral deepfake restrictions. California's AB 2602 and AB 1836 (2024) specifically address AI-generated replicas of deceased performers, extending right of publicity protections posthumously.

UNITED KINGDOM

The UK's response has been characterised by incremental statutory amendment. The Online Safety Act 2023 and Criminal Justice Bill 2024 represent the most significant criminal law advances, creating creation and distribution offences for NCII deepfakes. In the copyright domain, the UK Intellectual Property Office's 2023 consultation adopted a cautious approach, maintaining the existing text and data mining exception under Section 29A of the CDPA without a commercial equivalent. The government's January 2025 AI Action Plan proposed a commercial text and data mining exception subject to rights reservation, generating substantial controversy within the creative industries and remaining unenacted as of early 2026.

EUROPEAN UNION

The EU has developed the most comprehensive deepfake regulatory architecture of any major jurisdiction. The AI Act (Regulation (EU) 2024/1689), entering into force in August 2024 with a thirty-six-month phased implementation, classifies deepfake generation systems as high-risk, subjecting providers to conformity assessments, transparency requirements, and human oversight obligations. Article 50 requires AI systems generating synthetic content to implement machine-readable disclosure of AI

generation and to prevent watermark circumvention. The GDPR further constrains deepfake training data practices: facial images and voice recordings constitute special category biometric data under Article 9, requiring explicit consent for processing a requirement routinely violated by internet-scale training data collection.

AUSTRALIA AND INDIA

Australia has adopted a proactive approach to NCII deepfakes through the Online Safety Act 2021 and the eSafety Commissioner's powers to direct platform removal of deepfake intimate content. The Australian Law Reform Commission's recommendation for a commercial text and data mining exception, analogous to U.S. fair use, remained unlegislated as of early 2026.

Ministry of Electronics and Information Technology advisories in November 2023 directed platforms to remove deepfake content within 24 hours, and the Electoral Commission issued guidelines restricting deepfake use in the 2024 General Election. Indian courts have not yet addressed deepfake copyright questions directly, and the Information Technology Act 2000 provides an imperfect fit for AI-generated content.

ASIAN JURISDICTIONS

China enacted the world's most prescriptive deepfake-specific regulation in the Cyberspace Administration of China's Provisions on the Administration of Deep Synthesis Internet Information Services (effective January 2023), requiring real-name identity verification for deepfake platform users, mandatory content watermarking, and reporting and removal mechanisms for harmful synthetic content. While this framework is technically comprehensive, its practical efficacy is difficult to assess externally, and its requirements for universal real-name verification raise freedom of expression concerns inapplicable to democratic comparators. South Korea enacted the Deepfake Sexual Crime Prevention Act in 2020, subsequently increasing penalties and extending liability to possession.

FOUNDATIONAL PRINCIPLES

The comparative analysis reveals that existing regulatory responses are fragmented, jurisdictionally inconsistent, and inadequate to address the full range of deepfake harms. This chapter proposes a multi-tiered framework grounded in five principles: technology-neutrality (applicable to all synthetic media regardless of generation architecture); proportionality (calibrating obligations to harm severity and fault); effectiveness (prioritising empirically supported deterrence and remediation); rights-compatibility (avoiding disproportionate **restriction of expression, artistic freedom, and research); and international coherence** (promoting harmonisation to prevent regulatory arbitrage).

TECHNICAL STANDARDS: PROVENANCE AND WATERMARKING

The first regulatory tier addresses synthetic media authentication infrastructure. Building on the work of the Coalition for Content Provenance and Authenticity (C2PA) and the requirements of the EU AI Act, the framework proposes mandatory embedding of standardised provenance metadata in all outputs of AI systems capable of generating realistic synthetic media.

International standardisation of metadata formats through ISO, ITU, or IETF is essential for cross-jurisdictional interoperability.

COPYRIGHT LAW REFORM

The framework recommends resolution of AI training copyright uncertainty through an extended collective licensing scheme modelled on existing cable retransmission and private copying levy frameworks. Under this model, AI developers would be required to obtain licences from collecting societies representing rights holders across relevant creative categories visual art, photography, motion pictures, sound recordings, and literary works and to pay proportional royalties based on training data type and volume. Rights holders would retain an explicit opt-out right, and uncollected royalties would fund emerging creator support. This approach avoids the unsatisfactory alternatives of categorical infringement (which would prohibit AI development) and categorical fair use (which would deny all compensation to creators).

On copyright ownership of AI-generated outputs, the framework recommends a human authorship threshold based on the degree and character of human creative contribution. Works in which humans exercise meaningful creative control through selection, arrangement, modification, or creative prompting should attract copyright vesting in the human author. Autonomously generated content should enter the public domain. This recommendation is consistent with constitutional and statutory copyright foundations across examined jurisdictions and creates incentives for maintaining meaningful human creative involvement in AI-assisted production.

CRIMINAL LAW AND PLATFORM LIABILITY REFORM

The framework proposes four categories of bespoke criminal offences. First, malicious deepfake creation: creation of a deepfake of an identifiable person with intent to cause serious harm, punishable by a substantial custodial term and not requiring proof of distribution. Second, deepfake fraud: an aggravated fraud offence attracting enhanced penalties where AI- generated synthetic media is used to impersonate individuals.

Third, deepfake electoral interference: creation or distribution of a deepfake falsely depicting candidates or electoral authorities for electoral influence purposes, subject to a disclosure defence for clearly labelled satire.

Platform liability reform is a central regulatory priority. The framework recommends amendment of Section 230 to impose a duty of care standard: platforms with actual knowledge of deepfake NCII, fraud, or electoral interference should be required to remove content expeditiously and implement reasonable technical measures to prevent re-upload. Civil liability would attach to non-compliant platforms, without extending to criminal liability. Adoption and international extension of the DSA's systemic risk mitigation framework would complement this duty-of-care approach.

INTERNATIONAL COOPERATION AND MEDIA LITERACY

Given the transborder character of deepfake creation, distribution, and harm, national regulatory responses are inherently inadequate in isolation. The framework recommends negotiation of an international treaty addressing synthetic media regulation, modelled on the Budapest Convention on Cybercrime, establishing minimum standards for national deepfake offences, mutual legal assistance procedures, extradition arrangements for serious deepfake crimes, and international technical standards for provenance and authentication. Particular attention should be directed to the inclusion of Global South countries as both active participants in regulatory design and principal targets of deepfake disinformation campaigns.

Complementary investment in public media literacy education is essential to reduce societal vulnerability to deepfake harms. Specific measures include integration of synthetic media literacy into school curricula, public awareness campaigns targeting older adults vulnerable to deepfake fraud, professional training for journalists and fact-checkers, and public provision of accessible deepfake detection tools.

CONCLUSION

This paper has established four core findings. First, the synthesis gap between deepfake technology and existing legal doctrine is real, significant, and growing. Criminal law, particularly for well-established offences such as fraud and CSAM, has been more readily adapted to deepfake conduct than copyright law, which faces foundational doctrinal uncertainties about authorship, training data reproduction, and derivative works that require legislative resolution. Second, jurisdictional fragmentation of regulatory responses creates protection gaps and regulatory arbitrage opportunities that can only be addressed through international treaty mechanisms.

Fourth, extended collective licensing offers a principled solution to AI training copyright uncertainty, respecting creators' rights while enabling continued technological development.

SUGGESTIONS

The following suggestions arise from the analysis conducted in this study. Legislatures in all examined jurisdictions should enact bespoke deepfake criminal offences addressing creation, distribution, and commercial exploitation of harmful synthetic media, independently of existing general offences. The United States Congress should enact comprehensive federal deepfake legislation consolidating the current patchwork of state statutes and providing uniform standards for NCII, fraud, and electoral deepfake conduct.

Copyright law reform through extended collective licensing should be pursued as the preferred mechanism for resolving AI training data liability, providing fair remuneration for rights holders while enabling continued AI development under clear legal rules. The human authorship requirement for copyright protection should be maintained and clarified through legislative guidance specifying the quantum of human creative contribution required to satisfy authorship standards in AI-assisted production.

Platform liability reform should introduce a duty of care standard for deepfake content, replacing the current reactive notice-and-takedown model with obligations to implement proactive detection and prevention measures proportionate to platform size and risk profile. Mandatory provenance metadata requirements should be adopted as a universal technical standard, with circumvention criminalised, to enable downstream detection, attribution, and enforcement against harmful synthetic media.

International treaty negotiations for a synthetic media governance framework should be initiated within appropriate multilateral fora, prioritising the establishment of minimum criminal law standards, mutual legal assistance mechanisms, and harmonised technical standards. Public investment in media literacy education, deepfake detection tools, and interdisciplinary research on the social impact of synthetic media should be substantially increased across all jurisdictions.

REFERENCES:

Books and Journal Articles

Basel Committee on Banking Supervision. (2023). Cryptoasset exposures: Disclosure and KYC guidance. Bank for International Settlements.

Benkler, Y., Faris, R., & Roberts, H. (2018). Network propaganda: Manipulation, disinformation, and radicalization in American politics. Oxford University Press.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., … Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

Centre for Countering Digital Hate. (2023). The deepfake danger: How AI is changing the face of online sexual abuse. CCDH.

Chesney, R., & Citron, D. K. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. California Law Review, 107(6), 1753–1820. https://doi.org/10.15779/Z38RV0D15J

Citron, D. K. (2019). Sexual privacy. Yale Law Journal, 128(7), 1870–1960.

Citron, D. K., & Chesney, R. (2021). A guide to understanding the deepfakes crisis. The Conversation.

Fallis, D. (2021). The epistemic threat of deepfakes. Philosophy & Technology, 34(4), 623– 643. https://doi.org/10.1007/s13347-020-00419-2

Ganz, A. M. (2021). Identity theft in the age of artificial intelligence. Vanderbilt Journal of Entertainment and Technology Law, 23(3), 555–602.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems, 27, 2672–2680.

Khanjani, Z., Watson, G., & Janeja, V. P. (2023). Audio deepfakes: A survey. Frontiers in Big Data, 5, 1001063. https://doi.org/10.3389/fdata.2022.1001063

Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. ACM Computing Surveys, 54(1), 1–41. https://doi.org/10.1145/3425780

Paris, B., & Donovan, J. (2019). Deepfakes and cheap fakes: The manipulation of audio and visual evidence. Data & Society.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10684–10695.

Stupp, C. (2019, August 30). Fraudsters used AI to mimic CEO's voice in unusual cybercrime case. The Wall Street Journal.

Tidy, J. (2024, February 4). Deepfake scam in Hong Kong: Finance worker tricked into paying out $25 million. BBC News.

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. Information Fusion, 64, 131–148. https://doi.org/10.1016/j.inffus.2020.06.014

U.S. Copyright Office. (2023). Copyright and artificial intelligence: Part 1 — Digital replicas. U.S. Copyright Office.

Zuboff, S. (2019). The age of surveillance capitalism: The fight for a human future at the new frontier of power. PublicAffairs.