Review of Futures Markets
ISSN 1933-7116 (Print)
fra.st/rfm

Volume 34 · Issue 2 · 2026
Pages 1529–1542

Reconciling Deepfake Regulations vis-à-vis Human Rights: A Comparative Legal Analysis Across Jurisdictions

Himanshu Thakran[1], Dr. Komal[2]

[1] PhD Research Scholar, Department of Law, Gurugram University

[2] Associate Professor, Department of Law, Gurugram University

Received: 05/03/2026 | **Accepted:** 20/08/2026 | **Published:** 21/10/2026

Abstract

Synthetic media produced by generative artificial intelligence, commonly called deepfakes, have become an industrialised instrument of financial fraud, electoral manipulation, and image based sexual abuse, and they now threaten the evidentiary foundations of courts and human rights investigations. This Article assesses how four jurisdictions, the United States, the United Kingdom, the European Union, and India, had addressed the problem by mid 2026, and it measures each regime against the International Covenant on Civil and Political Rights (ICCPR), in particular the guarantees of a fair trial in Article 14, privacy in Article 17, expression in Article 19, and political participation in Article 25. The comparison reveals four distinct regulatory philosophies: rights constrained, state-led regulation in the United States; upstream criminalisation of the technology supply chain in the United Kingdom; transparency-based, risk-tiered governance in the European Union; and accelerated intermediary liability in India. Each protects some rights at the expense of others, and all depend on provenance technology whose metadata is easily stripped and whose cryptographic assurances are weaker than they appear. The Article also examines the "authentication deficit" confronting the International Criminal Court. It concludes that domestic statutes alone cannot adequately address a borderless technical problem and proposes an additional protocol under the United Nations Global Digital Compact establishing "Provenance by Design," combining mandatory preservation of provenance metadata by distribution platforms with the formal adoption of the Berkeley Protocol by international tribunals, and disciplined throughout by the necessity and proportionality requirements of Article 19(3) of the ICCPR.

Keywords: deepfakes; synthetic media; ICCPR; freedom of expression; non-consensual intimate imagery; provenance; C2PA; International Criminal Court; comparative law

I. Introduction

Courts, tribunals, and human rights bodies increasingly rely on an implicit presumption that recorded digital evidence bears a reliable relationship to the events it depicts. By mid 2026, the rapid spread of artificial intelligence (AI) generated synthetic media, commonly known as deepfakes, had made it harder to trust digital evidence. Deepfakes have evolved from experimental AI into tools that can facilitate financial scams, spread false information during elections, create non-consensual sexual imagery, and undermine the reliability of evidence presented in court.[1]

Governments have responded to the growing problem of deepfakes by introducing new legal and regulatory measures. During 2025 and early 2026, several jurisdictions moved beyond approaches that relied

How to cite: Thakran & Komal (2026). Reconciling Deepfake Regulations vis-à-vis Human Rights: A Comparative Legal Analysis Across Jurisdictions. Review of Futures Markets, 34(2), 1529–1542.
© 2026 The Author(s). Published by Review of Futures Markets.

primarily on reporting harmful content and requiring its removal. However, these developments have not produced a consistent regulatory framework. In the United States, efforts to regulate synthetic political content have faced constitutional challenges concerning freedom of expression. The United Kingdom has extended criminal liability to certain tools used to generate non-consensual intimate images. The European Union has focused on transparency requirements for AI-generated content, while India has adopted stricter obligations for online intermediaries. These different approaches reflect a broader challenge: measures designed to prevent harm can also affect freedom of expression, the responsibilities of digital platforms, and the protection of individual rights.

Deepfakes are created, shared, and experienced across national borders, making them difficult for any single country to regulate effectively. The central question for international human rights law is therefore not simply whether an individual State's legislation is effective, but whether the different national approaches, taken together, adequately protect the rights affected by synthetic media. These rights include privacy and dignity, freedom of expression, democratic participation, and the right to a fair trial.

The Article pursues three research questions. First, how do the regulatory models adopted in the United States, the United Kingdom, the European Union, and India fare when measured against the ICCPR? Second, what technical and jurisdictional limits are common to those models, and what do they imply for the evidentiary integrity of legal proceedings, including before the International Criminal Court (ICC)? Third, what form of international coordination could reduce those limits without licensing disproportionate restrictions on expression?

The Article is organised into eight further parts. Parts II to V explain the scope and method of the study, examine the nature and scale of the deepfake threat, set out the relevant human rights framework, and analyse the resulting challenges for digital evidence. Part VI compares the regulatory frameworks of the four jurisdictions, while Part VII examines provenance technology and its limits. Part VIII proposes a harmonised international standard, "Provenance by Design," and Part IX concludes the discussion. The Article argues that every jurisdiction faces a tension between responding quickly to deepfake-related harm and protecting freedom of expression. A more sustainable solution lies in preserving the authenticity of genuine digital records from the moment of their creation.

II. Scope, Method, and Sources

This Article employs a doctrinal and comparative method. It examines primary legal sources where available, including statutes, regulations, draft rules, and judicial decisions, and it relies on practitioner commentary and official guidance where primary texts were not yet reported in standard repositories as of mid-2026. Four jurisdictions are examined. They were selected because they embody four different regulatory philosophies and together cover a very large share of global platform users.

The benchmark is the ICCPR.[2] The provisions most directly engaged by deepfakes, Articles 14, 17, 19, and 25, are interpreted by the Human Rights Committee through its General Comments. The United States ratified the Covenant in 1992 subject to reservations and a declaration that its substantive provisions are not self-executing, so the Covenant operates there as an international standard rather than as a source of judicially enforceable domestic rights.[3] The United Kingdom, India, and the member states of the European Union are parties. The Article accordingly treats the Covenant as a yardstick for evaluating legislative design, rather than as a claim that each State is domestically liable under it.

Three limitations should be stated. First, the empirical data on the prevalence and cost of deepfake fraud derive largely from vendor studies, threat-intelligence reports, and market research, which are not peer reviewed, may reflect commercial incentives, and are not always methodologically transparent. This Article therefore excludes figures that circulate only through secondary aggregators, confines Part III to a small number of data points traceable to an identified originating source, and distinguishes expressly between realized losses and forecasts. The figures are used to establish orders of magnitude and direction of travel rather than precise measurements. Second, several 2026 instruments were still in flux at the time of writing. Third, the Article does not address every regulatory domain. It concentrates on non-consensual intimate imagery (NCII), electoral deception, intermediary obligations, and provenance.

III. The Scale and Typology of the Deepfake Threat

A. Market Forecasts and Their Limits

To understand the legal challenge, one must first distinguish what has been measured from what has merely been projected. Market-research forecasts project that the generative-AI market, covering both generation and detection tools, will reach $19.8 billion by 2033, a compound annual growth rate (CAGR) of 45.1%. That 45.1% rate is a forecast of commercial expansion produced by market-research modelling. It is not an observed measure of the prevalence of deepfakes or of the harm they cause, and this Article uses it only to indicate the direction in which investment in both generation and detection capacity is expected to move.

B. The Evolution of Attack Vectors

Attack methods have changed. Traditional presentation attacks, in which altered media is held before a camera or microphone, are increasingly displaced by injection attacks, in which synthetic streams are fed directly into identity-verification systems through virtual cameras or driver shims, bypassing the physical sensor altogether. A growth rate of this magnitude is measured from a small base and derives from a vendor's own detection data, so it is evidence of the direction and speed of attacker adaptation rather than of absolute scale. Its legal significance lies elsewhere: an injection attack defeats liveness checks that presuppose a genuine camera feed, which means that the authenticity of a recording cannot be inferred from the fact that it passed through a verification interface.

C. Realized Losses and Forecasts

Financial exposure must be described with equal care, because realized losses and forecasts are different kinds of evidence. The clearest realized loss is the 2024 incident in which the engineering firm Arup lost $25.6 million after a finance worker joined a video conference in which every other apparent attendee, including the Chief Financial Officer, was a real-time deepfake.[4] That figure is a realized corporate loss arising from a real-time deepfake injection attack on a live meeting. It is a single documented incident, not an estimate of aggregate harm. The $40 billion figure often quoted in this literature is of a different kind. It is a forecast: the Deloitte Centre for Financial Services projects that fraud losses in the United States enabled by generative AI could reach $40 billion by 2027.[5] It is a model-based projection for 2027, not a measurement of present losses, and it is not comparable with the Arup figure. Table 1 collects the indicators on which this Article relies and labels each as realized, forecast, vendor-reported, or experimental.

D. Human Detection Limits

Human observers cannot serve as a reliable safeguard. In a 2025 study of 2,000 consumers in the United Kingdom and the United States, only 0.1% of participants correctly identified every deepfake shown to them across a mixed set of genuine and synthetic images and videos, and participants were 36% less effective at detecting synthetic video than synthetic images.[6] Because video is the modality in which contemporary fraud and disinformation increasingly circulate, the result is especially troubling: detection performance is weakest where the stakes are highest. On these figures, unaided human perception is fundamentally incapable of serving as a safeguard against synthetic media. The study was conducted by an identity verification vendor and its participants were consumers rather than trained forensic examiners, so the finding is properly read as a limit on lay detection and not as a ceiling on what forensic tools can achieve. That limit is nonetheless decisive for legal design. Education and digital-literacy campaigns, valuable as they are, cannot by themselves address a threat of this kind. Systemic, technology-driven interventions are required, and they raise the rights questions examined next.

IV. Deepfakes and the ICCPR: A Doctrinal Framework

Synthetic media engage the Covenant in at least four ways. They interfere with privacy and dignity under Article 17, they implicate both the protection of and the permissible restriction of expression under Article 19, they threaten the integrity of elections under Article 25, and, as Part V explains, they undermine the evidentiary basis of fair trials under Article 14.

A. Privacy, Dignity, and Non-Consensual Intimate Imagery

Article 17 protects every person against arbitrary or unlawful interference with privacy, family, home, or correspondence, and against unlawful attacks on honour and reputation.[7] The most pervasive and psychologically damaging application of deepfake technology is the generation of NCII. Privacy torts and

criminal statutes historically focused on the distribution of real intimate images. Deepfake technology has forced a doctrinal expansion. Mapping a person's face onto explicit source material misappropriates biometric identity and removes from the subject any control over sexual autonomy, even though no intimate act ever occurred.

A growing body of legislation recognizes that the dignitary harm arises at the moment of creation, regardless of whether the material is later published. The shift toward regulating "nudification" applications at their source, examined in Part VI, reflects this deeper understanding of privacy and dignity. It is also the clearest example of a State discharging the positive obligation, recognized by the Human Rights Committee, to protect individuals against interference with their privacy by private actors.[8] A State that waits for publication before acting protects reputation after the injury; a State that criminalizes the tools of creation protects privacy before it.

B. Freedom of Expression, the Liar's Dividend, and Fair Trial

Article 19 guarantees the freedom to seek, receive, and impart information and ideas of all kinds, regardless of frontiers and through any media of choice.[9] The protection extends to art, satire, and parody, domains in which synthetic media have legitimate and historically recognized value. Restrictions are permissible only if they satisfy the cumulative three-part test of Article 19(3): they must be provided by law, pursue one of the enumerated legitimate aims, and be necessary and proportionate to that aim.[10] The Committee has stressed that restrictions on the operation of internet-based information systems are permissible only to the extent compatible with that test.[11] Sweeping, categorical bans on AI-generated media will therefore routinely fail the necessity and proportionality requirements, as the American and Indian experiences in Part VI show.

Yet synthetic media also create a distinctive epistemic harm. The existence of high-fidelity fabrications gives wrongdoers plausible deniability regarding genuine evidence of their own misconduct, a phenomenon scholars have termed the "liar's dividend."[12] By arguing that an authentic recording of human rights abuses or corruption may simply be an AI fabrication, perpetrators can evade accountability and force prosecutors and journalists into resource-intensive authentication contests. Regulating deepfakes is thus not only a matter of suppressing false speech. It is equally a matter of preserving society's capacity to recognize and authenticate true speech.

The liar's dividend is also a fair-trial problem. Article 14 of the Covenant guarantees every person a fair hearing and, as the Human Rights Committee has explained, the principle of equality of arms: the same procedural rights must be available to all parties unless distinctions rest on objective and reasonable grounds.[13] That principle is strained when one party can cast doubt on a genuine recording merely by invoking the possibility of fabrication, while the party that relies on the recording lacks any technical means of rebutting the doubt. Equality of arms is restored only through procedural scaffolding, meaning standardized methods for capturing, hashing, and preserving digital material, of the kind the Berkeley Protocol supplies (Part V.D). Such scaffolding gives both parties, and the court, a shared and checkable basis for deciding whether a file is what it purports to be.

C. Democratic Participation and Electoral Integrity

Article 25 protects the right of every citizen to take part in the conduct of public affairs and to vote in genuine periodic elections guaranteeing the free expression of the will of the electors.[14] Democratic choice presupposes access to an accurate factual baseline about candidates, parties, and policies. Deepfakes deployed in electoral contexts, such as fabricated audio of a candidate admitting to a crime or manipulated video of election workers tampering with voting machines, sever the link between the genuine will of the electorate and the outcome of the ballot. The difficulty is that every state remedy for electoral deception also places the State in the position of adjudicating political truth, which is the activity that Article 19 is most concerned to restrain. As Part VI.C argues, the European Union's approach of requiring transparency rather than deletion comes closest to reconciling the two Covenant interests.

D. An Analytical Framework for Evaluating Regulation

The remainder of the Article applies four questions drawn from these provisions. Is the restriction clearly defined, so that it is genuinely "provided by law"? Does it target a Covenant-recognized harm, such as the privacy of identifiable persons or electoral integrity, rather than mere falsity? Is it the least intrusive means,

including in the speed and manner of enforcement? And does it place the determination of unlawfulness with an independent body rather than with an executive or a platform acting under threat of liability?

V. The Evidentiary Crisis in International Criminal Law

Nowhere is the crisis of digital truth more acute than in international criminal justice. The ICC and various international investigative mechanisms increasingly rely on citizen-gathered digital material, commonly referred to as open-source intelligence, to document atrocity crimes. The right to a fair trial under Article 14 of the Covenant, and the corresponding guarantees in the Rome Statute, presuppose that the evidence on which judges decide is reliable and that the parties meet it on equal terms. Generative AI threatens that presupposition from two directions at once: fabricated material may be admitted as authentic, and authentic material may be discredited as fabricated.

A. The Free-Proof Model

The evidentiary architecture of the ICC follows a "free-proof" model. Article 69 of the Rome Statute authorizes the Court to request the submission of all evidence it considers necessary to determine the truth and to rule on the relevance or admissibility of any evidence, taking into account its probative value and any prejudice to a fair trial.[15] Chambers thus enjoy broad discretion and largely avoid the strict exclusionary rules familiar from common-law systems. Under this model, the central question is usually not whether a piece of digital evidence is admitted but how much weight it can bear. The flexibility leaves the Court without a settled technical standard for testing whether a digital file is what it purports to be.

B. The Trajectory of ICC Practice

ICC practice reveals growing reliance on digital evidence and a lag in authentication standards. In *Al Mahdi*, the Prosecution relied heavily on open-source video, satellite imagery, and interactive digital platforms to prove the deliberate destruction of cultural heritage in Timbuktu. The case advanced the Court's approach to spatial reconstruction but sidestepped the question of how raw digital files are formally authenticated.[16] In *Al-Werfalli*, the Court issued an arrest warrant based almost entirely on viral social media videos depicting extrajudicial executions in Libya. Because the videos circulated widely and appeared to show the same events from different perspectives, the Court relied on them heavily.[17]

The fragility of this reliance was exposed in the Article 70 proceedings against Jean-Pierre Bemba Gombo and others. The Prosecution tendered Facebook photographs, and the defense showed that the platform had stripped the files of their original metadata, leaving their origins, dates, and authenticity unestablished. Properly understood, the Chamber's treatment of the photographs illustrates the free-proof model rather than a rule of exclusion. Article 69 did not make the photographs inadmissible. The loss of metadata instead degraded their reliability to the point at which the Chamber could not assign them probative weight.[18] The point concerns provenance as much as forgery: missing metadata, not proof of fabrication, defeated the evidence. It also shows that the Court does not accord automatic probative weight to naked digital files absent technical verification.

C. The Yekatom and Ngaïssona Judgment

The authentication deficit reached a critical point in the judgment in *Yekatom and Ngaïssona*, delivered by Trial Chamber V on July 24, 2025. Evaluating approximately 943 items of Facebook material and 104 call data records, the Chamber produced one of the most extensive treatments of social media and telecommunications evidence in the Court's jurisprudence.[19] The case exposed a structural vulnerability. Social media platforms, in enforcing their own automated community standards, routinely remove content that may constitute vital evidence of international crimes. Where archived copies survive, they are frequently stripped of original metadata, leaving investigators and the court with naked digital files that are functionally indistinguishable from sophisticated fabrications.

The Office of the Prosecutor has sought to address part of this gap procedurally. In December 2025 it issued a dedicated Policy on Cyber-Enabled Crimes intended to strengthen the management of digital evidence.[20] Such a policy guides the Prosecution's own practice, but it does not bind Chambers or compel platforms to preserve metadata.

D. The Berkeley Protocol as Procedural Scaffolding

Practitioners have turned to soft-law instruments to navigate the problem. The Berkeley Protocol on Digital Open Source Investigations, published in 2020 by the

Office of the United Nations High Commissioner for Human Rights and the Human Rights Centre at the University of California, Berkeley, offers the first globally recognized guidelines for using publicly available digital information in international criminal and human rights investigations.[21] The Protocol prescribes methodologies for capturing, preserving, and analysing digital material. It emphasizes generating cryptographic hashes at the point of capture to establish an unbroken chain of custody, so that any later alteration, whether by AI or by a human editor, is mathematically detectable. In Article 14 terms, this is the mechanism by which equality of arms is restored: a party relying on genuine open-source material can demonstrate its integrity by method, and an opponent who invokes the liar's dividend must engage with that demonstration rather than rest on bare doubt.

Together with complementary instruments such as the Leiden Guidelines on the Use of Digitally Derived Evidence,[22] the Protocol supplies procedural scaffolding. Both are non-binding. Until the ICC and other international tribunals formally adopt the Protocol and integrate cryptographic authentication standards into their rules of procedure and evidence, the spectre of deepfakes and the liar's dividend will continue to threaten the reliability on which international criminal justice depends.

E. The UN Convention against Cybercrime

The multilateral treaty framework offers little help on this problem. The United Nations General Assembly adopted the United Nations Convention against Cybercrime on December 24, 2024.[23] The Convention is not yet in force. It remains open for signature until December 31, 2026, and it requires forty ratifications to enter into force. It also does not create bespoke deepfake offenses. Synthetic-media abuse falls within it only to the extent that it can be prosecuted under traditional cyber-fraud and computer-related frameworks. The Convention demonstrates that multilateral agreement on digital crime is attainable, but it leaves unaddressed both the provenance of digital evidence and the preservation of its metadata.

VI. Comparative Analysis of National Regulatory Frameworks

Recognizing the inadequacy of reactive measures, States enacted bespoke synthetic-media legislation at speed in 2025 and 2026. This Part examines the United States, the United Kingdom, the European Union, and India. Each is assessed using the framework in Part IV.D.

A. The United States: State Innovation and Constitutional Limits

The American posture is bifurcated and shaped almost entirely by the First Amendment. Federal law has established a nationwide baseline for intimate deepfakes through the Tools to Address Known Exploitation by Immobilizing Technological Deepfakes on Websites and Networks Act (TAKE IT DOWN Act), signed in 2025 and fully enforceable from May 19, 2026, which requires covered platforms to remove NCII within 48 hours of a valid request.[24] The regulation of political and commercial deepfakes has been left largely to the states, among which California has become the de facto national standard-setter through a package of AI legislation enacted in 2024 that took effect across 2025 and 2026.[25]

California's measures survived constitutional scrutiny where they targeted non-expressive harms. S.B. 926, effective January 1, 2025, criminalizes the creation and distribution of non-consensual deepfake pornography where the distributor knew or should have known that it would cause serious emotional distress, empowers the Attorney General to seek civil penalties, and gives victims a private right of action.[26] A.B. 621, effective January 1, 2026, raised statutory damages for non-consensual deepfake pornography to as much as $250,000 where malice is proven and added civil penalties enforceable by public prosecutors.[27] A.B. 1831 and S.B. 1381 extended child sexual abuse material laws to content generated or altered by AI, closing the argument that no real child was harmed.[28] In the commercial sphere, A.B. 2602 and A.B. 1836 protect living and deceased performers against unauthorized digital replicas, rendering such contracts unenforceable absent explicit consent or estate approval.[29]

California also moved toward technological provenance through S.B. 942, the California AI Transparency Act. Its operative date was delayed from January to August 2, 2026. It requires major AI developers to embed machine-readable watermarks in outputs and to offer labelling and detection tools, with penalties of $5,000 per violation per day.[30]

The attempt to regulate electoral deepfakes met a constitutional barrier, and its procedural history

matters. A.B. 2839 prohibited the distribution, with malice, of "materially deceptive" election communications within 120 days of an election. In *Kohls v. Bonta*, the Eastern District of California preliminarily enjoined the statute in October 2024, holding that it likely violated the First Amendment because it suppressed protected political speech and was not the least restrictive means of serving the State's interest.[31] The case proceeded alongside *The Babylon Bee, LLC v. Bonta*, which broadened the relief. On August 29, 2025, the court granted summary judgment to the plaintiffs and permanently enjoined enforcement of A.B. 2839.[32] Applying strict scrutiny, the court held that the statute constituted viewpoint-based and speaker-based discrimination, because the speech it reached, and the speakers it burdened, depended on the message conveyed. It further held that the statute unconstitutionally compelled speech by requiring disclaimers on parody and satire, a burden that fell on core expressive works and was not narrowly tailored to any interest in preventing deception. A companion measure, A.B. 2655, which would have imposed removal and labelling duties on large platforms with respect to election deepfakes, was stayed in January 2025 following a challenge by X Corp.[33]

The reasoning of the court reflects a broader strand of American doctrine. The Supreme Court has refused to treat false statements as a categorically unprotected class of speech, even where the falsity is knowing,[34] and political speech receives the most exacting scrutiny, with satire and parody expressly within its protection. *Kohls* is also a paradigm of failure under Article 19(3) of the Covenant. A prohibition on "materially deceptive" content that sweeps in parody and satire, and that compels speakers to label their own jokes, is neither necessary nor proportionate when less restrictive means, such as counter-speech and targeted disclosure, are available. The American result is accordingly not wholly dissimilar to what the Covenant would require. The United States is strong on NCII and commercial exploitation, yet structurally unable to protect its electoral infrastructure from AI manipulation through direct legislative prohibition. The practical consequence is that protection against electoral deepfakes must come from disclosure, provenance, and platform practice rather than from prohibition.

B. The United Kingdom: Criminalizing the AI Supply Chain

In contrast to the American focus on post-publication distribution, the United Kingdom has adopted a lifecycle model that targets the technological supply chain of synthetic abuse. Its culmination is the Crime and Policing Act 2026 (CPA), which received Royal Assent on April 29, 2026 and entered into force on June 29, 2026. Section 99 of the CPA amended the Sexual Offences Act 2003 by inserting new sections 66I to 66L. Section 66I creates an offense directed at the makers, adapters, and suppliers of "purported intimate image generators," colloquially known as nudification tools.[35]

Section 66I moves the point of criminal liability upstream. A person commits the offense by making, adapting, supplying, or offering to supply an AI tool where a reasonable person, having regard to all the circumstances, would consider that the technology was made or supplied for creating, or facilitating the creation of, purported intimate images of a person without consent. The test is objective: it asks what a reasonable person would conclude, not what the developer intended. The maximum penalty on indictment is three years' imprisonment and an unlimited fine. Section 66L provides for corporate liability, exposing executives to prosecution where the offense is committed with their consent or connivance. A narrow affirmative defence in section 66I(4) applies where the developer proves that it took all reasonable steps to prevent misuse. Section 66I(6) provides extraterritorial reach, but only over natural persons and bodies incorporated under UK law, even where the conduct occurred abroad.[36]

The Act also reinforces platform accountability and addresses the permanence of digital harm. Sections 100 and 101 impose a 48-hour takedown requirement on platforms for NCII following a valid report, accelerating the duties established by the Online Safety Act 2023.[37] It introduces bespoke deletion orders empowering courts to compel convicted offenders to destroy all copies of intimate images, including those in cloud storage.

Measured against the framework in Part IV.D, the UK model is the most protective of dignity and the most interventionist among Western democracies. In Covenant terms, it is the clearest instance of a State meeting its positive obligation under Article 17 at the

moment of creation, by criminalizing the instrument of the intrusion upstream rather than awaiting its publication. It scores well on legitimate aim and on judicial determination. Its vulnerabilities lie elsewhere. The objective "reasonable person" standard, applied to a dual-use technology whose lawful applications include image editing and entertainment, raises a question of legal certainty under the "provided by law" limb of Article 19(3), which the "all reasonable steps" defence mitigates but does not remove. Enforcement against foreign and open-source developers remains the principal practical limit, since extraterritorial reach extends only to UK natural persons and UK bodies and therefore leaves the most prolific sources of nudification tools outside the statute's practical grasp.

C. The European Union: Transparency by Design

The European Union's response is anchored in the precautionary, risk-based architecture of Regulation (EU) 2024/1689, the AI Act. Article 50 is binding law. It requires providers of generative systems to ensure that outputs are marked in a machine-readable format and detectable as artificially generated, and requires deployers of systems that generate deep fakes to disclose that the content has been artificially generated or manipulated.[38]

The Code of Practice on Transparency of AI-Generated Content stands on a different legal footing. On June 10, 2026, the European Commission and the newly established EU AI Office published the final Code, a co-regulatory instrument that operationalizes Article 50 and attracted approximately 190 signatories.[39] The Code is a voluntary compliance tool. Signature is a means of demonstrating adherence to the binding obligation, and it neither creates nor replaces it, so that providers and deployers who do not sign remain subject to Article 50 in full. The Code requires signatories to disclose AI generation through standardized icons and to embed persistent machine-readable watermarks in synthetic text, audio, and visual content. The deadline for full compliance with the watermarking provisions is December 2, 2026. A political agreement on the AI Omnibus in May 2026 also introduced a categorical ban on nudifier applications, taking effect in December 2026.[40]

The European model is the least restrictive of expression among the four, because it regulates disclosure rather than content. It fits comfortably within the Article 19(3) test, and it supports the Article 19 right to receive information by helping audiences distinguish authentic from synthetic material. It also best serves Article 25. Because the remedy is a label rather than a deletion, the State does not adjudicate the truth of political speech, yet voters are equipped to weigh a message knowing that it is synthetic. Of the four regimes, transparency rather than deletion offers the best balance between electoral integrity and free expression. Its weaknesses are enforcement gaps. The Code depends on voluntary adherence by open-source developers and has limited application to private, non-commercial users who run local generation models on their own hardware. Binding Article 50 is only as strong as the persistence of the disclosure itself, a problem taken up in Part VII.

D. India: Piercing the Intermediary Shield

India's evolution in 2026 shows the sharpest tension between state control, intermediary liability, and constitutional free expression. Confronted with a surge of election-related deepfakes and financial fraud, the Ministry of Electronics and Information Technology (MeitY) promulgated the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Amendment Rules, 2026 (2026 Rules), which have been in force since February 20, 2026 and amend the 2021 Rules.[41] The operative 2026 Rules must be distinguished from the later draft discussed below, which is not law.

The 2026 Rules define "synthetically generated information" (SGI) and impose timeline-centered obligations on intermediaries. They reduce the earlier 36-hour takedown window to three hours where removal is required by a government directive or court order, and to two hours where a user grievance concerns NCII, including non-consensual deepfake nudity. The two clocks therefore have different triggers: the three-hour period is order-driven, and the two-hour period is complaint-driven. Significant social media intermediaries (SSMIs) must in addition deploy automated verification tools to determine whether content is synthetic and must ensure mandatory labeling, including permanent metadata or identifiers embedded in synthetic content, a domestic version of provenance by design. Non-compliance forfeits the safe harbour of section 79 of the Information Technology Act, 2000, and may expose platform executives to liability under the Bharatiya Nyaya Sanhita, 2023.

The framework met immediate judicial scrutiny. In May 2026, the Delhi High Court (Sachin Datta, J.) examined the operational feasibility of the rules. Civil society and industry petitioners argued that a three-hour window cannot be met with human review and will compel platforms into automated, preemptive over-removal to preserve their safe harbour.[42] The constitutional objection is best framed through *Shreya Singhal v. Union of India*. There the Supreme Court read down the intermediary takedown provision so that an obligation to remove content arises only on a court order or government notification, precisely in order to prevent private over-censorship under threat of liability. It later affirmed that restrictions on internet-based expression must meet a proportionality test.[43]

Measured against that line of authority, the structure of the 2026 Rules is troubling in three respects. First, on the three-hour track the trigger is formally the kind of order that *Shreya Singhal* accepted, but compressing compliance into three hours, on pain of losing safe harbour, leaves no realistic interval in which an intermediary can assess legality, seek clarification, or contest an overbroad order. The rational response is to remove first and ask later. Second, the two-hour NCII track is set running by a user grievance rather than an order, which brings it close to the private-notice model that *Shreya Singhal* sought to restrain. The interest it serves, protection against intimate-image abuse under Article 17, is weighty, and a short NCII deadline is defensible in principle. Third, the automated verification and labeling duties on SSMIs operate ex ante and on all content, so the error rates of classifiers translate directly into the suppression or mislabeling of lawful speech, including satire. Threatening safe harbour through ultra-short deadlines thus forces preemptive, algorithmic over-removal, and it risks violating the proportionality requirement that Article 19(2) of the Constitution of India demands of any restriction on expression.[44] The same analysis applies under Article 19(3) of the Covenant. Together with *Kohls v. Bonta* in the United States, the Indian Rules show how deepfake regulation can fail the necessity and proportionality test, either by sweeping up parody and satire or by delegating censorship to automated systems.

Distinct from the operative 2026 Rules is the Draft IT (Second Amendment) Rules, released on March 31, 2026 and debated publicly through April. The draft is not in force, and its provisions should not be attributed to the operative Rules. It would effectively dissolve the user-intermediary distinction that has structured Indian internet law for a quarter century. It creates a new class, "Significant User-Publishers of News and Current Affairs," that extends platform-level compliance duties, code-of-ethics obligations, and 36-hour takedown requirements directly to high-reach individual users and content creators.[45] It also proposes an inter-ministerial Digital Content Review Panel to adjudicate disputes with user-publishers, creating an executive appellate layer that bypasses ordinary judicial review. By expanding takedown grounds to include undefined terms such as "demonstrably false claims on matters of public record," the draft approaches the limits of constitutional permissibility under Article 19(1)(a) of the Constitution of India.

These requirements sit alongside other compliance burdens, including the six-hour incident-reporting requirement imposed by CERT-In directions issued in 2022.[46] Against the Article 19(3) test, the Indian approach presents the sharpest conflict between speed of response and proportionality. Its determination of unlawfulness is largely executive, its definitions are open-textured, and its deadlines presuppose automated screening of the kind that most reliably produces over-removal.

E. Comparative Assessment

Table 2 summarizes the four regimes. The comparison reveals an inverse relationship between speed of harm reduction and strength of expressive safeguards. The United States and the European Union adopt measures least likely to chill lawful expression, but they are also least able to reduce harm quickly in the political sphere. India, and to a lesser degree the United Kingdom, prioritize rapid harm prevention at greater risk to Article 19 interests. Figure 1 makes the contrast concrete in respect of the one variable that all four regimes address in some form: the time allowed to remove unlawful content.

Two further observations follow. First, the Indian windows are roughly one-sixteenth to one-twenty-fourth of the period allowed in the United Kingdom and the United States, although the triggers differ and no regime resolves the more fundamental difficulty of establishing which content is genuine. Second, every regime relies, in whole or in part, on provenance technology as a control: California through S.B. 942,

Table 1. Comparative Summary of Regulatory Frameworks, 2026

Jurisdiction	Principal 2026 instruments	Core mechanism	Structural weakness
United States	TAKE IT DOWN Act; Cal. S.B. 926, A.B. 621, S.B. 942	Federal 48-hour NCII takedown; state criminalization and civil remedies; developer watermarking	Cannot prohibit political deepfakes consistent with the First Amendment (*Kohls v. Bonta*, permanently enjoining A.B. 2839)
United Kingdom	Crime and Policing Act 2026, § 99 (inserting Sexual Offences Act 2003, §§ 66I–66L)	Supply-chain criminalization on an objective "reasonable person" test; "all reasonable steps" defence; corporate liability; 48-hour takedown; deletion orders	Extraterritorial reach limited to natural persons and UK bodies; certainty of the "reasonable person" standard
European Union	AI Act art. 50 (binding); June 2026 Code of Practice (co-regulatory); AI Omnibus	Mandatory labeling and machine-readable marking under art. 50; Code icons and watermarks by Dec. 2, 2026; nudifier ban	Gaps for open-source projects and private, non-commercial local models; Code is voluntary
India	IT Amendment Rules, 2026 (in force Feb. 20, 2026); draft Second Amendment Rules (Mar. 2026, not in force)	3-hour takedown (government or court order); 2-hour takedown (NCII user grievances); SSMI automated verification and mandatory labeling	Proportionality risk under Art. 19(2); algorithmic over-removal; operational infeasibility; pending Delhi High Court scrutiny

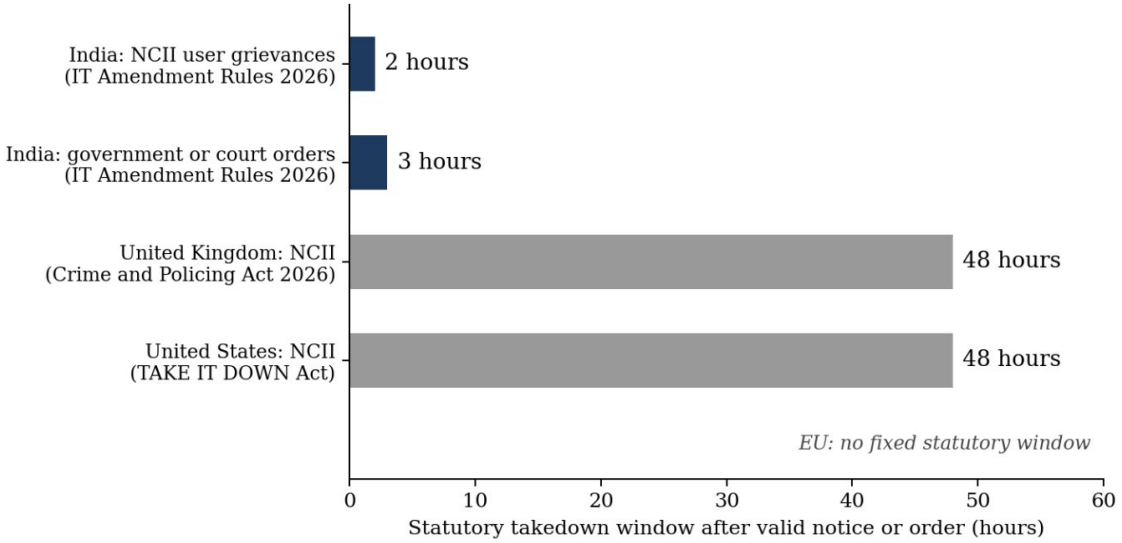


Figure 1. Statutory takedown windows for unlawful synthetic content by jurisdiction. NCII denotes non-consensual intimate imagery. The Indian three-hour window runs from a government directive or court order; the two-hour window runs from a user grievance concerning NCII. The EU framework prescribes labelling and watermarking rather than a fixed removal period. Sources: Crime and Policing Act 2026 (UK); TAKE IT DOWN Act (U.S.); Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Amendment Rules, 2026 (India).

the European Union through the Code of Practice, and India through mandatory labelling and embedded metadata. The reliability of that technology is therefore the shared point of failure, and it is the subject of the next Part.

VII. The Technological Bottleneck: Provenance, Watermarking, and C2PA

A unifying thread across the EU Code of Practice, California's S.B. 942, and India's amended Rules is the reliance on technological provenance, meaning the attachment of data to media files to declare their origin and history. The prevailing industry standard used to meet these requirements is the framework of the Coalition for Content Provenance and Authenticity (C2PA), which issues "Content Credentials." It is essential to state precisely what this technology is. C2PA is not a pixel-level watermark. A Content Credential is a cryptographically signed manifest of assertions about an asset's origin and edit history, appended to or associated with the asset as metadata. The manifest is bound to the asset by cryptographic hashing (SHA-256), and it is signed with credentials based on X.509 certificates, so that a validator can check both that the asset has not been altered since signing and that the signer is an identifiable party.[47] The design corresponds to the logic of the Berkeley Protocol: capture authenticity at the source so that later alteration is detectable.

The legal mandates of 2026 nevertheless collide with an architectural vulnerability, which is metadata fragility. Content Credentials are not fused to the pixel or audio data itself. When synthetic content is uploaded to major social media platforms, or processed through basic consumer editing software, the appended

metadata is frequently and automatically stripped during routine compression and re-encoding. Because C2PA relies on metadata rather than pixel-level fingerprinting, standard format conversion effectively erases the credentials.[48]

The vulnerability is compounded by the voluntary character of the standard. Because C2PA does not enjoy universal adoption across global distribution platforms, a bad actor can generate a nominally compliant deepfake, pass the file through a non-compliant platform or a simple format converter, and erase the cryptographic proof of AI generation. As of March 2026, no overriding international rule required audio or video files to retain Content Credentials across platforms, leaving the removal of provenance data unregulated in many jurisdictions. A deepfake generated in compliance with S.B. 942 and circulated on a foreign messaging service that strips metadata would reach an end user functionally indistinguishable from authentic media, defeating the transparency objective of the law. Without mandatory, cross-platform preservation of provenance data, domestic watermarking statutes offer a false sense of security.

A final limit is epistemic. A valid C2PA credential proves only the history of the file's processing: which device or tool produced it, who signed it, and what edits were declared. It does not prove the factual truth of the events the file depicts. A staged scene captured by a genuine camera carries a flawless credential, and a truthful image lacking one is not thereby false. Provenance can narrow the space for doubt, but it cannot replace judicial assessment of what a recording shows.

The same weakness compounds the evidentiary problem identified in Part V. The metadata that is stripped from a synthetic file and thereby conceals its origin is the same metadata that is stripped from a genuine file and thereby deprives it of credibility. A provenance regime that fails in one direction fails in both.

VIII. Toward a Harmonized International Standard: "Provenance by Design"

The comparative and technical analysis supports a central conclusion: domestic regulation, however aggressively drafted, cannot contain a borderless technical problem. Each regime meets a distinct limit: constitutional in the United States, jurisdictional in the United Kingdom, technical in the European Union, and proportionality-related in India. A solution therefore requires elevating the principle of provenance by design to a binding, harmonized international standard that ensures cross-border interoperability and preservation of provenance data. This Article proposes the negotiation of an additional protocol, under the United Nations Global Digital Compact, establishing "Provenance by Design."

The Global Digital Compact, adopted as an annex to the Pact for the Future in 2024, commits member States to cooperation on the governance of digital technologies and AI and contemplates an independent scientific panel on AI.[49] It is a political commitment rather than a treaty, so the protocol would have to be concluded as a separate binding instrument negotiated under its auspices. The United Nations Convention against Cybercrime shows that multilateral agreement on digital offenses is attainable, but its silence on deepfakes and provenance shows what remains to be done.[50] A protocol negotiated under the Compact should rest on three pillars.

1. **Provenance by design at creation.** Providers of generative systems above a defined capability threshold should be required to embed durable, interoperable provenance signals at the point of creation. The requirement should combine cryptographic metadata with robust watermarking that survives ordinary compression and format conversion, so that the system does not depend on metadata alone. Because the security analysis in Part VII shows that existing credential schemes can fail in validation, the standard should also require robust certificate revocation checking and tamper-evident timestamping, and should not assume that a valid credential is a trustworthy one.
2. **Mandatory preservation across distribution.** Distribution platforms and editing tools should be required to preserve provenance metadata, and algorithmic stripping of that metadata should be banned, subject to exceptions for privacy and safety and for lawful user redaction. The obligation should be framed as a content-neutral technical duty rather than as a takedown mandate, which keeps it consistent with the necessity and proportionality requirements of Article 19(3) by regulating infrastructure rather than speech.
3. **Evidentiary preservation and formal adoption of the Berkeley Protocol.** States and platforms

should preserve potential atrocity evidence together with original metadata and hash values. The ICC and other international tribunals should formally adopt the Berkeley Protocol and integrate its authentication methods into their rules of procedure and evidence, together with rules addressing the treatment of the liar's dividend, so that equality of arms under Article 14 does not depend on the informal practice of individual Chambers.

To remain faithful to the Covenant, the protocol should also embody four safeguards. First, it should protect satire, parody, and artistic expression expressly. Second, it should place determinations of unlawfulness with judicial rather than executive bodies, avoiding the Indian model of executive review. Third, any takedown periods should be calibrated to the realities of human review, since deadlines that can be met only by automated screening predictably induce over-removal. Fourth, it should protect authors in vulnerable settings, such as journalists, whistle blowers, and human rights defenders, who may have legitimate reasons to withhold identifying metadata. Mandatory provenance must not become an instrument of surveillance or a precondition for lawful speech.

Two objections deserve a response. The first is that a binding standard is unrealistic given the divergence among the jurisdictions studied, but that divergence is concentrated in the regulation of content. A standard confined to preserving technical metadata and authenticating evidence touches the least contested ground, and it offers every regime something it presently lacks. The second objection is that determined adversaries can circumvent provenance, and that a credential proves only the history of a file and not the truth of what it depicts. Both points are correct. But the aim is not to make deception impossible, nor to substitute a signature for judicial fact-finding. It is to restore a baseline in which authentic material can be proven authentic as to its origin and handling, thereby weakening the liar's dividend even where forgery cannot be eliminated.

IX. Conclusion

Deepfake technology is no longer an impending threat. It is a dominant, industrialized vector for digital fraud, biometric abuse, and electoral disinformation. Its commercial ecosystem is forecast to grow at a compound annual rate of 45.1% to 2033, a projection that should be read as a forecast and not as a measure of harm, and its realized costs are already visible in incidents such as the $25.6 million Arup loss. The legislative response, though swift and increasingly severe, is limited by jurisdictional isolation, constitutional constraint, and fragile technical enforcement. Each of the four regimes examined here protects some Covenant rights well and others poorly. The American regime protects expression at the cost of electoral integrity; the British regime protects privacy at the moment of creation at the cost of enforceability and, potentially, legal certainty; the European regime protects expression and electoral choice through disclosure but depends on the persistence of that disclosure; and the Indian regime pursues speed at the cost of proportionality.

The human rights consequences of this fragmentation are significant. Without the ability to trust digital records, the right to democratic participation is exposed to synthetic propaganda, the right to privacy and dignity to industrialized biometric abuse, and the right to a fair trial to the decay of evidentiary reliability and the exploitation of the liar's dividend.

The way forward lies in moving from localized, reactive platform regulation to a unified architecture of technical provenance. By embedding authenticity at the point of creation, requiring its preservation across distribution networks, formally adopting the Berkeley Protocol in international tribunals, and criminalizing the malicious supply chains that seek to defeat these controls, legal systems can rebuild the presumption on which evidence depends. The crisis of synthetic media cannot be solved by banning deception jurisdiction by jurisdiction. It can be addressed by an additional protocol establishing "Provenance by Design," which safeguards the authenticity of genuine records within the limits that the Covenant sets on every restriction of expression.

Notes

[1] *See* Robert Chesney & Danielle Keats Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 Calif. L. Rev. 1753, 1771–86 (2019).

[2] International Covenant on Civil and Political Rights, Dec. 16, 1966, 999 U.N.T.S. 171 [hereinafter ICCPR].

[3] *See* 138 Cong. Rec. S4781 (daily ed. Apr. 2, 1992) (resolution of advice and consent to ratification, with reservations, understandings, and declarations).

[4] Heather Chen & Kathleen Magramo, *Finance Worker Pays Out $25 Million After Video Call with Deepfake "Chief Financial Officer,"* CNN (Feb. 4, 2024); Kathleen Magramo, *British Engineering Giant Arup Revealed as $25 Million Deepfake Scam Victim*, CNN (May 17, 2024).

[5] Deloitte Ctr. for Fin. Servs., *Generative AI Is Expected to Magnify the Risk of Deepfakes and Other Fraud in Banking* (May 29, 2024).

[6] iProov, *iProov Study Reveals Deepfake Blindspot: Only 0.1% of People Can Accurately Detect AI-Generated Deepfakes* (Feb. 2025) (study of 2,000 consumers in the United Kingdom and the United States).

[7] ICCPR, *supra* note 2, art. 17; *see also* U.N. Hum. Rts. Comm., CCPR General Comment No. 16: Article 17 (Right to Privacy) (Apr. 8, 1988).

[8] U.N. Hum. Rts. Comm., General Comment No. 31: The Nature of the General Legal Obligation Imposed on States Parties to the Covenant, U.N. Doc. CCPR/C/21/Rev.1/Add.13, ¶ 8 (May 26, 2004).

[9] ICCPR, *supra* note 2, art. 19(2).

[10] ICCPR, *supra* note 2, art. 19(3); U.N. Hum. Rts. Comm., General Comment No. 34: Article 19: Freedoms of Opinion and Expression, U.N. Doc. CCPR/C/GC/34, ¶¶ 21–36 (Sept. 12, 2011) [hereinafter General Comment No. 34].

[11] General Comment No. 34, *supra* note 12, ¶ 43.

[12] Chesney & Citron, *supra* note 1, at 1785.

[13] ICCPR, *supra* note 2, art. 14(1), (3); U.N. Hum. Rts. Comm., General Comment No. 32: Article 14: Right to Equality Before Courts and Tribunals and to a Fair Trial, U.N. Doc. CCPR/C/GC/32, ¶ 13 (Aug. 23, 2007).

[14] ICCPR, *supra* note 2, art. 25; *see also* U.N. Hum. Rts. Comm., General Comment No. 25: The Right to Participate in Public Affairs, Voting Rights and the Right of Equal Access to Public Service (Art. 25), U.N. Doc. CCPR/C/21/Rev.1/Add.7 (Aug. 27, 1996).

[15] Rome Statute of the International Criminal Court art. 69(3)–(4), July 17, 1998, 2187 U.N.T.S. 3.

[16] *Prosecutor v. Al Mahdi*, ICC-01/12-01/15-171, Judgment and Sentence (Sept. 27, 2016).

[17] *Prosecutor v. Al-Werfalli*, ICC-01/11-01/17-2, Warrant of Arrest (Aug. 15, 2017).

[18] *Prosecutor v. Bemba Gombo*, ICC-01/05-01/13-1989-Red, Judgment pursuant to Article 74 of the Statute, ¶ 219 (Oct. 19, 2016).

[19] *Prosecutor v. Yekatom*, ICC-01/14-01/18-2784, Judgment, ¶ 129 (July 24, 2025).

[20] Off. of the Prosecutor, Int'l Crim. Ct., Policy on Cyber-Enabled Crimes (Dec. 2025).

[21] Off. of the U.N. High Comm'r for Hum. Rts. & Hum. Rts. Ctr., Univ. of Cal., Berkeley, Sch. of L., Berkeley Protocol on Digital Open Source Investigations (2020).

[22] Leiden Guidelines on the Use of Digitally Derived Evidence (Grotius Ctr. for Int'l Legal Stud., Leiden Univ.).

[23] United Nations Convention against Cybercrime, G.A. Res. 79/243, U.N. Doc. A/RES/79/243 (Dec. 24, 2024).

[24] Tools to Address Known Exploitation by Immobilizing Technological Deepfakes on Websites and Networks Act (TAKE IT DOWN Act), Pub. L. No. 119-12, 139 Stat. 55 (2025).

[25] *California Legislative Session Roundup: Which Key Privacy and AI Bills Were Enacted and Which Were Vetoed*, EPIC,https://epic.org/california-legislative-session-roundup-which-key-privacy-and-ai-bills-were-enacted-and-which-were-vetoed/ (last visited Oct. 16, 2026); *California's AI Laws Are Here—Is Your Business Ready?*, Pillsbury, https://www.pillsburylaw.com/en/news-and-insights/california-ai-laws.html (last visited Oct. 16, 2026).

[26] S.B. 926, 2023–2024 Leg., Reg. Sess. (Cal. 2024); *see California Legislative Session Roundup*, *supra* note 27.

[27] A.B. 621, 2025–2026 Leg., Reg. Sess. (Cal. 2025).

[28] A.B. 1831, 2023–2024 Leg., Reg. Sess. (Cal. 2024); S.B. 1381, 2023–2024 Leg., Reg. Sess. (Cal. 2024).

[29] A.B. 2602, 2023–2024 Leg., Reg. Sess. (Cal. 2024); A.B. 1836, 2023–2024 Leg., Reg. Sess. (Cal. 2024).

[30] S.B. 942, 2023–2024 Leg., Reg. Sess. (Cal. 2024) (codified as amended at Cal. Bus. & Prof. Code § 22757 et seq.).

[31] *Kohls v. Bonta*, 752 F. Supp. 3d 1187 (E.D. Cal. 2024).

[32] *Kohls v. Bonta*, 797 F. Supp. 3d 1177, 1192 (E.D. Cal. 2025); *see also Kohls v. Bonta*, No. 2:24-cv-02527 (E.D. Cal. Aug. 29, 2025); *Kohls v. Bonta*, EPIC, https://epic.org/documents/kohls-v-bonta/ (last visited Oct. 16, 2026).

[33] *X Corp. v. Bonta* (E.D. Cal. Jan. 2025) (order staying enforcement of Cal. A.B. 2655).

[34] *See United States v. Alvarez*, 567 U.S. 709, 718–22 (2012) (plurality opinion).

[35] Crime and Policing Act 2026, § 99 (UK) (inserting §§ 66I–66L into the Sexual Offences Act 2003, c. 42); *see also* U.K. Gov't, *Making or Supplying Purported Intimate Image Generators: Non-Statutory Guidance for the Police*, https://assets.publishing.service.gov.uk/media/6ad08b93c84c8ff1bc8b81e3/Making_or_Supplying_Purported_Intimate_Image_Generators_-_Non-Statutory_Guidance_for_the_Police_-_in_English__2_.pdf.

[36] Sexual Offences Act 2003, c. 42, §§ 66I(4), 66I(6), 66L (UK) (as inserted by Crime and Policing Act 2026, § 99).

[37] Crime and Policing Act 2026, §§ 100–101 (UK); Online Safety Act 2023, c. 50 (UK); *Deepfakes to Deletion Orders: Tackling Technology-Enabled Sexual Offending in the Crime and Policing Act 2026*, Kingsley Napley, https://www.kingsleynapley.co.uk/our-insights/articles/deepfakes-to-deletion-orders-tackling-technology-enabled-sexual-

offending-in-the-crime-and-policing-act-2026/ (last visited Oct. 16, 2026).

[38] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act), art. 50, 2024 O.J. (L 1689).

[39] *European AI Office Releases Code of Practice on Transparency of AI-Generated Content*, IPTC (June 2026), https://iptc.org/news/eu-ai-transparency-code-of-practice-june-2026/ (last visited Oct. 16, 2026).

[40] *The EU AI Act: The Current State of Play*, Travers Smith, https://www.traverssmith.com/knowledge/knowledge-container/the-eu-ai-act-the-current-state-of-play/ (last visited Oct. 16, 2026).

[41] Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Amendment Rules, 2026 (India) (in force Feb. 20, 2026) [Gazette notification number to be inserted]; Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021, G.S.R. 139(E) (Feb. 25, 2021) (India); *see also India's New Rules for AI-Generated Content and Deepfakes*, LiveLaw, https://www.livelaw.in/articles/ai-generated-content-deepfakes-524064 (last visited Oct. 16, 2026).

[42] Delhi High Court proceedings (May 2026) (Datta, J.), *as reported in India's New Rules for AI-Generated Content and Deepfakes*, *supra* note 43.

[43] *Shreya Singhal v. Union of India*, (2015) 5 SCC 1 (India); *Anuradha Bhasin v. Union of India*, (2020) 3 SCC 637 (India); *see also Justice K.S. Puttaswamy (Retd.) v. Union of India*, (2017) 10 SCC 1 (India) (adopting proportionality as a standard for rights restrictions).

[44] India Const. art. 19(2).

[45] *Draft IT (Second Amendment) Rules 2026: MeitY Expands Oversight*, Anantam IAS, https://anantamias.com/current-affairs/draft-it-amendment-rules-2026-meity/ (last visited Oct. 16, 2026).

[46] Directions Under Sub-Section (6) of Section 70B of the Information Technology Act, 2000, Ind. Computer Emergency Response Team (Apr. 28, 2022).

[47] Coal. for Content Provenance & Authenticity, C2PA Technical Specification, version 2.2 (2025).

[48] *See generally* C2PA Technical Specification, *supra* note 49 (describing manifests as metadata bound to an asset, and optional soft-binding mechanisms for re-associating stripped manifests).

[49] Global Digital Compact, G.A. Res. 79/1, annex I (Sept. 22, 2024) (Pact for the Future).

[50] United Nations Convention against Cybercrime, *supra* note 25.