International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Governing Synthetic Reality: Comparative Perspectives on Deepfake Regulation, Online Safety, and Democratic Speech

Shyamalal Chandra Sekhar Mishra[1], Maglin M Raja[2], Dr Swagatika Panigrahi[3], Komal Agarwal[4]

[1,2,3,4]Assistant Professor, Department of Law, Dr Ambedkar Global Law Institute, Tirupati

ABSTRACT

Generative artificial intelligence has advanced quickly enough that synthetic audio and video content now often cannot be told apart from genuine recordings. Lawmakers have struggled to keep pace with this shift, a difficulty this paper refers to as the pacing problem, where legal frameworks are drafted and debated far slower than the technology they attempt to govern. Three competing concerns pull regulatory efforts in different directions: the need to protect individuals from online harm and privacy violations, the constitutional or normative commitment to free expression, and the interest in preserving fair and informed democratic elections. To understand how different legal systems have responded to these pressures, this paper compares three regulatory paradigms. The European Union relies on a risk tiered structure established under the AI Act, paired with platform accountability duties introduced by the Digital Services Act. The United States, by contrast, has produced a patchwork of state level statutes shaped heavily by First Amendment doctrine, resulting in narrower, sector specific rules rather than a single federal framework. China has taken a security driven approach, requiring mandatory labelling of synthetic content and real name verification under rules issued by the Cyberspace Administration of China. Working through these three models, the paper argues that a blanket ban on synthetic media would raise serious constitutional concerns and would likely suppress legitimate uses of the technology alongside harmful ones. It instead makes the case for a layered system built on provenance and consent, one that combines technical measures such as the C2PA watermarking standard with binding platform moderation obligations and criminal liability reserved for clearly malicious uses. Building on this comparative groundwork, the paper sets out a Tri Pillar Synthetic Reality Governance Model meant to give policymakers a workable path between safety, individual rights, and what is technically achievable.

Keywords: Deepfake Regulation, Synthetic Media Governance, Generative Adversarial Networks, Online Safety, Democratic Speech, Comparative AI Law

1. INTRODUCTION

1.1 The Advent of Synthetic Reality

Digital alteration of images is not new. What's different is the sophistication and accessibility of the underlying tools. Back in the day manipulation was reliant on manual editing procedures that left visible traces and required professional ability. This picture changed with the development of Generative

Adversarial Networks that place a generator network against a discriminator network in a competitive training loop to produce synthetic outputs that iteratively improve until they cannot be reliably distinguished from real recordings.[1] Contemporary diffusion-based models and real-time neural rendering systems have advanced this technology, enabling realistic face swaps, voice replication, and complete body reenactment to be accomplished on consumer-grade hardware in only minutes, as opposed to the days or weeks previously required. This progression, from rudimentary manual splicing to instantaneous generative synthesis, constitutes the technological context within which contemporary regulatory structures should be evaluated. The speed of this advancement is inherently a facet of the issue this research examines, as legal frameworks are usually formed based on a fixed comprehension of a technology's potential, which frequently becomes obsolete prior to the instrument's use.

1.2 The Governance Trilemma

When dealing with synthetic media, regulators must balance three competing imperatives.

The first is on online security and privacy. Deepfake technology has been used to create non-consensual intimate pictures, impersonate someone for financial gain, and generate voices or likenesses for identity theft. These harms are borne primarily by private persons who lack a public platform from which to fight a created depiction, and women are disproportionately affected by nonconsensual synthetic imaging.[2]

Political debate and democratic speech come in second. Additionally, lawful and constitutionally permitted uses of synthetic media technologies include political satire, parody, theatrical reenactment, and the disguise of a whistleblower's name or voice to prevent reprisals. In instance, when a legislation defines synthetic content broadly rather than by reference to intent or harm, any regulatory reaction focused at damaging deepfakes runs the risk of trapping this protected expressive activity within the same net.

The third aspect is the integrity of elections and information. Synthetic media has emerged as a tool for state-sponsored disinformation efforts strategically aligned with elections, and its very presence has led to a secondary harm known as the liar's dividend. Legal experts Robert Chesney and Danielle Citron introduced this term to explain the advantage that a deceitful individual obtains not by producing a deepfake, but by suggesting that genuine and harmful video of them is actually falsified.[3] The public's exposure to deepfakes is increasing, which in turn makes it simpler to dismiss genuine recordings. This undermines trust in evidence in general, rather than in fabricated evidence specifically. This effect has already been observed in political contexts outside the jurisdictions surveyed in this paper, highlighting that the trilemma is not limited to any specific legal system. [4]

Domestic legislators often deal with the three imperatives separately but a rule designed to satisfy one will often run counter to another. A watermarking mandate, for example, aimed at protecting election integrity does little to assist a victim of non consensual intimate images who is hurt the moment such information is made and spread, whether or not it is identified as synthetic.

1.3 Research Questions

This paper is organised around three research questions.

RQ1: How do major jurisdictions balance the protection of free expression against the mitigation of harms arising from synthetic media, and what doctrinal or constitutional commitments account for the differences observed.

RQ2: What are the systemic strengths and failure modes of top down legislative models, as against co regulatory frameworks that place primary compliance obligations on platforms rather than directly on individual creators.

RQ3: How can technical provenance frameworks, particularly content credential and watermarking standards such as C2PA, complement statutory enforcement mechanisms without themselves becoming a form of compelled speech or a de facto licensing regime for expression.

2. Conceptual & Theoretical Framework

2.1 Epistemic Decay and the Liar's Dividend

The most pressing danger linked to synthetic media is the false material it produces. A more profound and lasting detriment, referred to in this study as epistemic decay, is the slow deterioration of a society's fundamental ability to reach consensus on what constitutes credible evidence. Chesney and Citron's description of the liar's dividend elucidates the mechanism accurately. Once the populace is aware that persuasive forgeries may be created, a malefactor faced with genuine and incriminating video has a credible new defence merely by claiming that the tape is artificial.[5] It is crucial to note that this dividend is accrued without the perpetrator ever having to generate a deepfake. The mere cultural cognisance that deepfakes exist is sufficient to accomplish the task. This asymmetry is significant for regulatory design, as it implies that policy instruments that are solely focused on the supply side, such as content labelling or detection tools, are unable to fully mitigate the harm. A watermark on a genuine, unaltered recording does not prevent a bad actor from falsely claiming that the watermark was faked or bypassed. The 2026 Iran war served as a documented example of this phenomenon, in which a video that served as evidence of life was met with public accusations that it was artificially produced, despite its authenticity. [6] The public discourse burden of proof shifts as epistemic degradation deepens. More often than not, the person relying on legitimate evidence must first confirm its authenticity.

2.2 The Jurisprudential Dilemma

Two doctrinal tensions recur across every jurisdiction surveyed in this paper.

Strict liability against intent based liability. Criminalising the mere creation or distribution of synthetic content portraying a real person, without requiring proof that the creator intended to deceive or harm, provides prosecutors with an easier evidential route but also criminalises conduct that causes no cognisable harm, such as clearly labelled parody or consented commercial use. By contrast, a statute that requires proof of a mens rea element, such as purpose to defraud, intent to harass or knowledge of falsity, better targets culpable activity but is correspondingly harder to show and may create a vacuum for careless or wilfully blind creation that does not reach the level of specific intent. The majority of the criminal laws analysed in the Comparative Analysis section have a mixed approach, applying strict liability for narrowly defined categories, such as non consensual intimate pictures, while needing purpose or knowledge for broader categories, such as election disinformation.

Categorical exclusions against protected political commentary. The First Amendment doctrine in the United States, along with similar free expression doctrines in other regions, identifies specific categories that are entirely excluded from constitutional protection. These include defamation, fraud, true threats, and obscenity. The Supreme Court's ruling in United States v. Alvarez determined that simply being false, without demonstrating the specific harm linked to defamation or fraud, does not strip speech of its constitutional safeguards. The Court declined to establish a separate category for "lies" on their own. [7] This is crucial for deepfake legislation, as a law merely stating "ban on producing a misleading representation of an individual" may encounter the same flaw that led to the failure of the Stolen Valour Act: it addresses falsehood in a theoretical sense rather than a tangible injury, and a court utilising Alvarez's rationale might deem it excessively broad. Consequently, proficient drafting must link

synthetic media crimes to a recognised unprotected category, such as fraud, non-consensual intimate photography, or a distinctly defined type of election trickery, instead of a broad prohibition on inauthentic representation. This is exactly the drafting principle that differentiates the EU's specific exception for clearly artistic or satirical material from more extensive real name and content regulation frameworks.

2.3 Structural Taxonomies of Synthetic Harms

Building on the harm categories introduced in the Introduction, this paper organises synthetic media harms into three tiers based on the scale of the affected interest.

Micro-level: The personal dignity and autonomy of an individual are impacted by micro-level damages. This category encompasses non-consensual intimate imagery, targeted harassment campaigns that are constructed around a fabricated depiction, and voice cloning that is employed to impersonate an individual to their family or employer. Even when a legal remedy is available, these harms are acute, immediate, and frequently irreversible in practical terms, as the reputational damage is rarely undone by the removal of content after broad distribution.

Meso-level: Rather of affecting individuals in and of themselves, harms at the meso-level effect organisations and markets. Corporate impersonation through the use of a fake executive statement, the use of synthetic audio to approve a fraudulent wire transfer, and market manipulation by the use of a fabricated announcement that is ascribed to a corporation or regulator are all examples of activities that fall under this division. The fact that these harms are typically addressed through pre-existing frameworks of fraud, securities, or corporate law rather than through deepfake-specific statutes raises the question, which will be developed further in this paper, of whether a dedicated synthetic media statute is even necessary for this tier or whether enforcement of existing law with clearer evidentiary rules would be sufficient.

Macro-level: It harms have an impact on how the political community and the state function. Election meddling, the deterioration of civic discourse caused by widespread disinformation, and the deployment of synthetic media in the context of armed conflict or national security all fall under this category. These are the harms most closely associated with the liar's dividend since they are widespread, cumulative, and unrelated to any particular created item.

In the Comparative Analysis section, it is a common finding that instruments calibrated to anxieties at the macro level, such as election integrity, are often those most heavily enforced against conduct at the micro level, and vice versa. The rest of the paper uses this three tier taxonomy to test whether a given regulatory instrument is in fact calibrated to the tier of harm it purports to address.

3. Comparative Regulatory Analysis

3.1 The European Union: Risk-Based Co-Regulation

The EU's strategy regards synthetic media mostly as an issue of transparency rather than a matter of substance. Article 50 of the AI Act does not prohibit deepfakes per se. Rather, it mandates that suppliers of AI systems producing synthetic audio, images, videos, or text label their outputs in a machine-readable format that is identifiable as artificially created, and additionally obliges users of such systems to reveal that the content has been artificially generated or altered.[8] An intentional decision to control the output of the technology, rather than its underlying motivation, is what causes the responsibility to attach, even in cases when no actual person is shown and no intent to deceive is apparent. The Act does establish a more stringent requirement for obviously creative, fictional, satirical, or artistic content: it

only requires disclosure that does not detract from the work's enjoyment. This change directly addresses the categorical exclusion problem from the Conceptual Framework, as it attempts to distinguish between expressive value and the transparency duty when drafting, instead of letting courts decide where to draw the line afterwards. [9]

Article 50 is another instance of the pacing problem in action. Its guidance documents note that content created prior to the application date of the provision on 2 August 2026 need not be labelled retroactively. Its final guidelines steer assessment towards a holistic set of factors including the level of resemblance and the audience's expectations, rather than a fixed technical test, precisely because a fixed test written today would likely be obsolete by the time enforcement began. [10]

The Digital Services Act, which is concurrent with Article 50, indirectly regulates synthetic media by imposing systemic risk obligations on Very Large Online Platforms, which are services with more than forty-five million monthly active users in the Union. Articles 34 and 35 mandate that these platforms identify and mitigate systemic risks associated with civic discourse, electoral processes, and public security. Disinformation campaigns are considered a fundamental example of such a risk, despite the fact that the DSA does not explicitly mention disinformation. [11] This structure shifts the compliance responsibility to the platform layer instead of individual creators. This is where the EU model distinctly diverges from a strict liability or intent-based criminal framework: a VLOP's obligation is assessed based on the overall reasonableness and proportionality of its risk mitigation program, rather than the legality of any single piece of synthetic content.

3.2 The United States: First Amendment Constraints and Fragmented Federalism

No broad federal deepfake statute applies to the US model. As mentioned in the Conceptual Framework, First Amendment case law, particularly United States v. Alvarez, sets the doctrinal ceiling on state legislation, which is how regulation instead occurs. Following Alvarez's line of thinking with synthetic media makes it seem like a law that forbids the mere fabrication of an inaccurate portrayal of a person is weak since it goes after abstract falsehoods instead of a specific type of unprotected speech.

This vulnerability has already played out in litigation. California's AB 2839 banned the publication of "materially deceptive" audio or video linked to an election a certain period before an election, and created a private right of action for damages, with exceptions for satire and parody. The law was challenged by a creator of a labelled parody video, and the district court in Kohls v. Bonta issued a preliminary injunction, finding that the statute was a blunt instrument that chilled humorous expression and stifled the exchange of ideas that democratic debate depends on, according to the court. [12] In 2025, the court ruled down AB 2655, which mandated big platforms to remove or label materially false election content, since it violated Section 230 of the Communications Decency Act's platform immunity. [13] Together, these rulings demonstrate the two distinct constitutional challenges that US deepfake statutes encounter: a First Amendment vagueness or overbreadth challenge, in which the statute directly targets falsity, and a Section 230 preemption challenge, in which the statute instead targets platform conduct.

Regardless of these obstacles, by 2025, 26 states had passed laws addressing political deepfakes. These laws typically followed one of two models: a prohibition model, which allows publication but requires labelling, or a disclosure model, which prohibits publication but has a pre-election window. Two examples of the former are Minnesota and Texas. [14] A distinct and comparatively more durable category of state legislation addresses non-consensual intimate imagery. This category closely mirrors

an established unprotected category (obscenity and privacy torts) than election-related falsity, and as a result, it has encountered fewer successful constitutional challenges to date.

3.3 China: State-Centric Pre-emptive Governance

China's Regulations on the Management of Deep Synthesis of Internet Information Services, collaboratively released by the Cyberspace Administration of China, the Ministry of Industry and Information Technology, and the Ministry of Public Security, have been in force since 10 January 2023 and exemplify the most centralised of the three models analysed. [15] The Provisions impose duties directly on deep synthesis service providers, rather than primarily relying on downstream accountability for individual producers. Providers should seek an individual's consent before generating a synthetic likeness of them. Providers should verify the real identity of users through compulsory real name verification. Providers should append a visible or otherwise perceptible signature or watermark to synthetic outputs to prevent public confusion. Providers should set up mechanisms to refute rumours generated on their platforms. The next year, these duties were reinforced by the Interim Measures for Generative AI Services, which extended real name verification and security assessment requirements more generally to generative AI services. [16]

The Chinese framework is structurally distinct from both the EU and US models in that it prioritises the risk of social and political instability over individual harm or electoral falsity. This concern is enforced preemptively through licensing-style obligations on providers, rather than through post-hoc civil or criminal liability against individual creators. The framework's operation has been comparatively challenging to evaluate in comparison to the due process safeguards that would typically accompany a system of the same severity in the EU or US models, as enforcement has typically consisted of content removal and administrative warnings issued during politically sensitive periods, such as elections, rather than reported court judgements. [17]

3.4 India: Judicial Improvisation without Statute

Judicial improvisation in the absence of dedicated law is the fourth model that India exhibits, which is significantly different from the three models that were explored before. Protection against the misuse of synthetic media has instead been assembled by courts from personality rights doctrine, which is itself drawn from the right to privacy under Article 21 of the Constitution, along with performer's rights under Sections 38 and 38A of the Copyright Act, 1957. India does not have a statute that specifically addresses deepfakes. [18] The Delhi High Court's ruling in Anil Kapoor v. Simply Life India & Ors. marked a significant moment in Indian jurisprudence by directly tackling the issue of deepfakes. The court issued an ex-parte, omnibus injunction that prohibits the unauthorised use of the actor's name, voice, image, and signature mannerisms and catchphrases, aiming to prevent misuse through advanced technologies, including face morphing and deepfakes, for any purpose, including commercial exploitation. [19] This was preceded by Amitabh Bachchan v. Rajat Nagi & Ors., where the same court issued an ad interim injunction operating in rem, applicable globally rather than solely to the specified defendants, a remedy significantly more expansive than those provided under the EU, US, or Chinese legal systems discussed earlier. [20] In Arijit Singh v. Codible Ventures LLP, the first Indian judgement to address generative AI speech usage, the Bombay High Court expanded this authority to voice cloning across the metaverse. [21] The judicial approach is demonstrated by more recent orders, such as the interim protection granted to Sri Sri Ravi Shankar against fabricated videos falsely depicting him endorsing health remedies and the pending matter brought by Member of Parliament Shashi Tharoor over a fabricated political statement, which extend from commercial and reputational harms to the political speech category this

paper's taxonomy treats as macro-level. [22] Indian courts can issue ex-parte injunctions in days after filing, which is faster than any of the three statutory models. However, unlike statutory schemes, this method does not include procedural safeguards, clearly defined evidentiary standards, or appellate review. As a result, courts in India effectively legislate harm categories case by case, rather than doing so through a body that can be revised by an elected legislature. This model adds a fourth stance to the typology of this study, joining judicially-driven, ex-parte, and reactive, whereas the models of the European Union, the United States, and China are all, to varying degrees, legislatively-driven and prospective.

4. Constitutional & Democratic Tensions

4.1 Political Satire versus Deceptive Disinformation

The border between protected parody and unprotected deceit is not a line the law has ever drawn with complete confidence, and synthetic media has made the subject more pressing, not resolved. The theological anchor in the United States for this problem is Hustler Magazine v. Falwell. The Supreme Court held that a public figure could not recover for intentional infliction of emotional distress based on a parody unless the parody includes a false statement of fact made with actual malice. The Court specifically noted that the jury had already determined that no reasonable person would have understood the parody as describing actual events. [23] The content's plainly stated fiction turned a harmful lie into protected criticism.

Synthetic media challenges this paradigm as the indicator of fictionality that safeguarded the Hustler parody a ludicrous and exaggerated premise that no rational reader could confuse with reality is exactly what sophisticated generative technologies aim to eradicate. The compelling influence of a deepfake is rooted in its authenticity rather than its hyperbole. The guidance of the AI Act highlights the ensuing challenge: its concluding Article 50 directives instruct evaluators to consider the degree of similarity, the essential message, and audience anticipations in a comprehensive manner, explicitly stating that content the audience does not perceive as genuine in a specific context may be excluded from the deepfake classification, despite being synthetically produced. [24] This serves as a feasible solution for a professionally made film featuring digitally de-aged performers, where the context indicates that it is fictional. It is significantly less effective for a brief video shared on social media that lacks the context of its initial posting, as a marked parody video, when re-shared without its original disclaimer, can spread through a network that is essentially the same as the misleading content the law aims to address. Kohls v. Bonta exemplifies the same fundamental conflict in the electoral arena: the creator's video was marked as parody at the time of posting, yet the statute in question would have subjected him to private lawsuits regardless, as it characterised "materially deceptive" content based on its impact on a theoretical reasonable viewer rather than the creator's initial disclosure. [25] No jurisdiction reviewed in this article has closed the doctrinal gap between content that is deceptive by purpose and content that becomes deceptive solely through downstream re-circulation.

4.2 Platform Moderation and Privatised Censorship

A subsequent tension emerges when regulations change, as seen in the EU's Digital Services Act and several US state laws, transitioning from direct liability for artists to compliance responsibilities placed on platforms. Jack Balkin's description of what he calls new school speech regulation illustrates this transition: instead of directly penalising a speaker, the government compels an intermediary, like a platform, to obstruct, restrict, or eliminate another individual's speech under the threat of legal

repercussions, a situation he refers to as collateral censorship. [26] Asymmetrical incentives are produced by this. Since the cost of a user complaint is far lower than the cost of a wrongful retention (regulatory liability or a private right of action), a platform that faces statutory liability for failing to catch deceptive synthetic content has every incentive to over-remove borderline material, including real satire, labelled parody, and contested but authentic footage. In his work on digital prior restraint, Balkin takes this worry a step further. He describes the current setup as a triangle with three sides: nation states, platforms with private governance, and individual speakers. In this triangle, the practical extent of a speaker's speech is being decided by the platform's compliance posture rather than public law alone. [27] This structure is directly demonstrated by the DSA's systemic risk framework under Articles 34 and 35, which evaluates a Very Large Online Platform's compliance based on the effectiveness of its overall risk mitigation program rather than on any particular moderation choice. This gives the platform a strong incentive to set its automated detection thresholds conservatively. [28] The same dynamic is demonstrated by California's AB 2655, which mandated that platforms either block or label materially deceptive election content. Despite the fact that the underlying statute did not survive judicial review, the platform's compliance behaviour is not automatically unwinding simply because the statute is later invalidated. This is due to the fact that risk averse moderation policies tend to persist operationally even after their legal basis is removed. This is the practical cost of collateral censorship that is often overlooked in a purely doctrinal examination of the constitutionality of the statute.

4.3 The Limits of Defamation and Right of Publicity

The same dynamic is demonstrated by California's AB 2655, which mandated that platforms either block or label materially deceptive election content. Despite the fact that the underlying statute did not survive judicial review, the platform's compliance behaviour is not automatically unwinding simply because the statute is later invalidated. This is due to the fact that risk averse moderation policies tend to persist operationally even after their legal basis is removed. This is the practical cost of collateral censorship that is often overlooked in a purely doctrinal examination of the constitutionality of the statute.

Defamation law also necessitates a false statement of fact, which does not align well with synthetic content that does not make specific factual claims but rather creates an entirely invented scene or statement. Furthermore, it demands evidence of fault (typically actual malice for public figures, as established in New York Times v. Sullivan, or negligence for private individuals), which can be challenging to prove against an anonymous or judgment-proof defendant, even when causation is evident. Claims related to the right of publicity, designed to safeguard an individual's name, voice, and likeness from unauthorised commercial use, encounter a more limited yet connected issue: numerous jurisdictions' publicity laws were formulated with a focus on commercial exploitation, like unauthorised endorsements, and fail to explicitly address non-commercial deepfakes created solely for the purpose of harassment, deception, or embarrassment. This oversight creates a gap where the meso and macro level harms highlighted in the Conceptual Framework are likely to arise. China's Deep Synthesis Provisions seek to address this issue by shifting liability to the service provider, who can be identified and regulated, even when the individual user remains anonymous, through compulsory real name verification. [29] This upstream approach avoids anonymity and jurisdiction issues that defeat tort remedies against individual creators by substituting platform-level preemptive control for the individualised fault-based liability that defamation and right of publicity doctrine were designed to deliver, which a significant normative trade is off rather than a neutral technical fix.

5. Multi-Stakeholder Governance & Technical Interventions

5.1 Watermarking, Cryptographic Provenance, and Metadata

The most advanced technical solution to the provenance issue identified in this paper is the Coalition for Content Provenance and Authenticity standard, which is commonly abbreviated as C2PA. A C2PA manifest is a structured, cryptographically signed record that is attached to a file. It records the tool used to create or edit the content, the timestamp, and each subsequent modification. The manifest is signed by a certificate from a trusted certificate authority, ensuring that any downstream viewer can verify that it has not been altered since the moment of signing. [30] OpenAI's picture and video models and Adobe's Firefly suite incorporate manifests by default, and several big news companies have editorial rules to validate them.

The fundamental constraint of the standard is structural, not incidental. The original C2PA manifest is retained as file metadata instead of being integrated within the visual or audio content, resulting in its loss when a file is processed by tools that do not maintain that metadata, including common activities like taking a screenshot, automatic re-compression by a platform during upload, or format conversion. Independent evaluation has characterised this as an operational conundrum: the material that requires verifiable origins, specifically content that becomes viral on social media, is exactly the type most prone to losing its provenance metadata throughout its viral dissemination. [31] The standard's most recent iteration, C2PA 2.1, endeavours to resolve this issue by implementing durable content credentials. This involves the pairing of the metadata manifest with an imperceptible watermark, such as Google's SynthID, that is directly embedded into the pixels, audio samples, or text tokens of the content. The theory behind this approach is that a watermark of this nature can withstand re-encoding and cropping even after the metadata manifest has been removed. [32] This represents a significant enhancement, yet it does not fully address the issue. Independent security researchers have shown that a counterfeit manifest can be created using publicly accessible C2PA tools and linked to a specific individual who did not participate in the creation of the content. Additionally, an AI-generated image can be signed by a compliant capture device to generate a manifest that certifies an edit history instead of verifying authenticity at the moment of capture. [33] To put it another way, the C2PA and the watermarking layer that is associated with it are able to reliably tell a verifier what happened to a file after it was created. However, they are not able to guarantee that what was created was true on their own, nor are they able to survive a determined adversary who intentionally routes content through non-compliant tools before it is distributed. As a result, the provenance infrastructure is treated in this work as a necessary evidence layer rather than a sufficient governance mechanism on its own. This distinction is further addressed in the Tri-Pillar model that is provided in the final section of the paper.

5.2 Platform Liability and Duty of Care

The safe harbour framework derived from Section 230 of the US Communications Decency Act and the earlier EU E-Commerce Directive was designed for a hosting paradigm where platforms primarily serve as neutral channels for third-party content, offering protection from liability as long as they do not create the illegal material themselves. This structure is discordant with a synthetic media landscape where the platform's recommendation algorithm deliberately chooses, enhances, and redistributes content, including material that originated as a designated parody but subsequently lost that designation during the amplification process. The DSA's shift from the traditional safe harbour framework, as examined in the Comparative Analysis, substitutes passive immunity with an active obligation regarding systemic risks for major platforms, whereas the UK's Online Safety Act advances this further by categorising the

creation and dissemination of non-consensual intimate imagery as a priority offence, thereby imposing a proactive responsibility on platforms to avert and promptly eliminate such content, supported by Ofcom's authority to impose civil fines of up to ten percent of a platform's worldwide revenue. [34] The UK's approach also demonstrates a duty of care model that extends further upstream than either the EU or the US position prior to 2025. This is because the Data (Use and Access) Act's amendments made it illegal to create such content and separately criminalised providing a tool intended for that purpose, which targets the tool provider rather than just the end distributor. [35]

A genuine duty of care framework for recommendation algorithms, as opposed to a content removal duty, would require platforms to account for their own amplification choices, since synthetic content that a human moderator would flag within hours can reach millions of viewers within minutes through automated recommendation, regardless of moderation delay. No jurisdiction studied in this article governs the recommendation system as such with the same specificity used to content classification, which this paper highlights as a structural gap similar to all three models rather than a failure specific to anyone.

5.3 Criminal and Civil Remedy Frameworks

The treatment of non-consensual intimate imagery is the clearest example of a well-targeted, narrowly drawn statutory response among the jurisdictions surveyed. This is solely because this harm category maps most cleanly onto an established basis for restricting speech, rather than requiring a new and constitutionally fragile category built around falsity or deception generally. Knownly publishing or threatening to publish non-consensual intimate imagery, including AI-generated depictions, is criminalised under the United States federal TAKE IT DOWN Act, which was enacted in May 2025 with nearly unanimous bipartisan support in both chambers of Congress. Additionally, the act imposes a takedown obligation on covered platforms upon notification. [36] Instead of depending on either individual criminal liability or an upstream platform obligation alone, the UK approach combines the two, creating a criminal offence for both sharing and, since recent amendments, creating such imagery on top of the Online Safety Act's platform duty structure. This achieves the same result through a different institutional route. [37]

This limited category of statute can teach us about the synthetic media challenge this study addresses. Legislators have had more success drafting durable, court-tested rules when harm is defined by an established doctrinal category, non-consent to an intimate depiction, than when they define harm by synthetic media as a technology or falsity, as in the Alvarez-vulnerable statutes. This suggests that the tailored remedy framework most likely to survive constitutional review and reach the conduct it targets is one built harm by harm along the Conceptual Framework's micro, meso, and macro taxonomy, rather than a single omnibus deepfake statute.

6. Conclusion & Policy Recommendations

6.1 Rationale

The Governance Trilemma is not resolved by any jurisdiction that has been examined. The liar's dividend is not addressed by the EU's disclosure model. The United States model is only viable when statutes address established harms, rather than falsity itself. By shifting liability upstream, China's model accomplishes durability at the expense of individualised due process. The Tri-Pillar Model integrates the resilient component of each.

6.2 Pillar I: Harm-Specific Statutory Anchoring

Never apply anchor statutes to "synthetic media" in general; instead, focus on crimes including known injury, lack of permission, fraud, or particular electoral lies. While statutes based on untruth (AB 2839, AB 2655) did not pass review, statutes based on consent (UK Sexual Offences Act, TAKE IT DOWN Act) did. For specific injuries, such as NCII, use strict liability; for more general harms, such as disinformation, demand intent. By definition, electoral rules should not apply to commentary or satire, but only to intentionally misleading claims on voting mechanisms. Existing creator-centered responsibility does not cover real-time and agentic generation; therefore, it is necessary to mandate periodic legislation review and expand coverage to these areas.

6.3 Pillar II: Layered Technical Provenance

Rather than serving as a replacement for enforcement, provenance is proof. Given the recognised hazards of forgery and stripping, it is recommended that persistent content credentials, a signed manifest, and an in-content watermark be required by default. Additionally, continuous published red-teaming should be implemented. Courts and regulatory agencies only consider credentials to be evidence that can be refuted. Provenance approaches should be extended to include real-time and streaming content, which is where manifests are now the least reliable.

6.4 Pillar III: Proportionate Platform Duties

Apply a systemic-risk approach (DSA Art. 34–35) to measure platform mitigation program compliance, not individual posts. Need faster appeals for wrongfully removed satire, disaggregated transparency reporting, and duty-of-care for recommendation systems, not only content classification. Use cross-border mutual recognition for removal requests to close the jurisdictional gap no regime addresses.

6.5 Judicial Application

The disclosure of fictitious or satirical aim in a clear and contemporaneous manner (Hustler) should be considered very persuasive evidence against liability. It is not appropriate to punish a creator who discloses information when such revelation is later removed by re-circulation by a third party.

6.6 Conclusion

This issue cannot be resolved by a single instrument in isolation. The Tri-Pillar Model is a preliminary framework that necessitates ongoing adaptation to real-time and cross-border cases. It views statute, provenance, and platform accountability as mutually compensating layers.

6.7 Limitations

The scope does not include India, Brazil, and South Korea. No empirical compliance data is currently available for the DSA or the UK Online Safety Act. No jurisdiction examined thus far governs real-time or agentic synthetic content.

Reference

1. Goodfellow et al., 'Generative Adversarial Networks' (2014) arXiv:1406.2661
2. R. Chesney and D. K. Citron, 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security' (2019) 107 California Law Review 1753
3. R. Chesney and D. K. Citron, 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security' (2018) SSRN 3213954.
4. S. A. Thompson and T. Hsu, 'Netanyahu Posts "Proof of Life" Video as A.I. Sows Doubts About What's Real', The New York Times, 17 March 2026
5. R. Chesney and D. K. Citron, 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and Nati-

nal Security' (2019) 107 California Law Review 1753, 1785
6. S. A. Thompson and T. Hsu, 'Netanyahu Posts "Proof of Life" Video as A.I. Sows Doubts About What's Real', The New York Times, 17 March 2026
7. *United States v. Alvarez*, 567 U.S. 709 (2012)
8. Regulation (EU) 2024/1689 (AI Act), Article 50(2), (4)
9. European Commission, Guidelines on Article 50 of the AI Act (2026)
10. Bird & Bird, 'European Commission adopts final Guidelines on AI Act Article 50 transparency obligations' (2026)
11. Regulation (EU) 2022/2065 (Digital Services Act), Articles 34-35
12. *Kohls v. Bonta*, No. 2:24-cv-02527 (E.D. Cal. 2024)
13. Reporting on the invalidation of California AB 2655 (E.D. Cal. 2025), addressing conflict with 47 U.S.C. sec 230
14. First Amendment Encyclopedia, 'Political Deepfakes and Elections' (2025)
15. Provisions on the Administration of Deep Synthesis of Internet Information Services (China), effective 10 January 2023
16. Interim Measures for the Management of Generative Artificial Intelligence Services (China), effective 15 August 2023
17. Library of Congress, Global Legal Monitor, 'China: Provisions on Deep Synthesis Technology Enter into Effect' (2023)
18. copyright Act, 1957, ss. 38, 38A (India)
19. *Anil Kapoor v. Simply Life India & Ors.*, 2023 SCC OnLine Del 6914
20. *Amitabh Bachchan v. Rajat Nagi & Ors.*, 2022 SCC OnLine Del 4110
21. *Arijit Singh v. Codible Ventures LLP* (Bombay HC, 2024)
22. Delhi High Court interim orders, Sri Sri Ravi Shankar (26 August 2025) and Shashi Tharoor v. Union of India & Ors. (2026, pending)
23. *Hustler Magazine, Inc. v. Falwell*, 485 U.S. 46 (1988)
24. Bird & Bird, 'European Commission adopts final Guidelines on AI Act Article 50 transparency obligations' (2026)
25. *Kohls v. Bonta*, No. 2:24-cv-02527 (E.D. Cal. 2024)
26. J. M. Balkin, 'Old-School/New-School Speech Regulation' (2014) 127 Harvard Law Review 2296
27. J. M. Balkin, 'Free Speech Is a Triangle' (2018) 118 Columbia Law Review 2011
28. Regulation (EU) 2022/2065 (Digital Services Act), Articles 34-35
29. Provisions on the Administration of Deep Synthesis of Internet Information Services (China), effective 10 January 2023
30. Coalition for Content Provenance and Authenticity, C2PA Technical Specification (2024)
31. RAND Corporation, 'Overpromising on Digital Provenance and Security' (2025)
32. Coalition for Content Provenance and Authenticity, C2PA 2.1 Specification, Durable Content Credentials (2026)
33. Hacker Factor, analysis of forged C2PA manifests (2026)
34. Online Safety Act 2023 (UK)
35. Data (Use and Access) Act 2025 (UK)
36. TAKE IT DOWN Act, Pub. L. 119-12 (2025)
37. Sexual Offences Act 2003, s. 66B (UK), as i