Legal Challenges of Deepfakes: Liability, Harm, and Regulatory Responses

1. Aleksandra Nowak: Department of Criminal Law, University of Warsaw, Warsaw, Poland
2. Bálint Tóth*: Department of Criminal Law, Eötvös Loránd University, Budapest, Hungary
3. Andrei Ionescu: Department of Private Law, University of Bucharest, Bucharest, Romania

*Correspondence: e-mail: balint.toth@elte.hu

Abstract

Deepfake technologies have rapidly evolved from experimental artificial intelligence innovations into widely accessible tools capable of producing hyper-realistic synthetic media that blur the boundaries between truth and fabrication. Their ability to manipulate identity, falsify expressive acts, and fabricate audiovisual content presents unprecedented challenges for legal systems structured around assumptions of authenticity, traceability, and human agency. This narrative review synthesizes current scholarship and regulatory developments to examine the multifaceted harms associated with deepfakes, including reputational injury, privacy violations, political manipulation, economic fraud, and broader societal erosion of trust. These harms expose significant doctrinal gaps in defamation, privacy, tort, and criminal law, which struggle to account for the speed and anonymity of synthetic media production. The analysis further explores complexities within liability frameworks, focusing on creators, distributors, platforms, AI developers, and malicious users whose involvement complicates traditional models of responsibility. Platform governance, intermediary liability, and cross-border enforcement challenges reveal structural weaknesses in existing regulatory approaches. National, regional, and international responses—including criminal prohibitions, civil remedies, the EU Digital Services Act, and emerging AI governance initiatives—offer partial solutions, yet they remain fragmented and inconsistently enforced. Soft-law interventions, including platform policies, content labeling, and authenticity tools, provide additional layers of protection but lack uniformity and transparency. To move toward a coherent legal framework, this review highlights the need for harmonized standards integrating civil, criminal, technological, and administrative mechanisms. Proposed directions include duty-of-care obligations for AI developers and platforms, mandatory watermarking systems, cross-border cooperation tools, and experimental regulatory sandboxes. Overall, the study emphasizes that addressing deepfake risks requires an interdisciplinary approach combining technological safeguards, legal reform, and multi-stakeholder governance to preserve individual rights, societal trust, and democratic resilience in an era of increasingly sophisticated synthetic media.

Keywords: Deepfakes; Synthetic Media; Intermediary Liability; Platform Governance; AI Regulation; Privacy; Defamation; Disinformation; Digital Policy; Content Authenticity

Received: date: 10 February 2023
Revised: date: 10 March 2023
Accepted: date: 24 March 2023
Published: date: 01 April 2023

Citation: Nowak, A., Tóth, B., & Ionescu, A. (2023). Legal Challenges of Deepfakes: Liability, Harm, and Regulatory Responses. *Legal Studies in Digital Age*, 2(2), 49-60.

1. Introduction

The rapid evolution of artificial intelligence has enabled the creation of hyper-realistic synthetic media known as deepfakes, which use advanced neural networks to fabricate audio-visual content that convincingly mimics real individuals. Scholars have noted that deepfake technology emerged from breakthroughs in generative adversarial networks, which accelerated the ability of synthetic media systems to manipulate identities at scale, and these developments raise profound regulatory concerns as the capacity to generate deceptive content has outpaced traditional legal mechanisms designed for slower, more traceable forms of media manipulation (Guess et al., 2020). Regulatory theorists have emphasized that the acceleration of digital content production has occurred simultaneously with the expansion of platform governance challenges, leading to new debates about intermediary responsibility and collective rights protections as highlighted in research that examines harmful content governance globally (Gorwa, 2020) and the increasing inability of legacy regulatory frameworks to keep pace with algorithmically amplified misinformation (Reich et al., 2022).

The harms associated with deepfakes have become increasingly visible across political, economic, and social domains. Political systems face the risk of destabilization when synthetic videos are deployed to manipulate electoral processes, a concern reinforced by analyses of election-related platform governance disputes (Gorwa, 2021). Democratic resilience is threatened when false audiovisual evidence circulates online at scale, enabling sophisticated disinformation campaigns, a challenge that legal commentators have identified when advocating for legislative reform to counter digitally produced falsehoods (Ray, 2021). Reputational harms are substantial because deepfakes can compromise personal dignity and privacy, exposing individuals to identity manipulation and character attacks, especially in environments where platforms struggle to control large volumes of illegal and harmful content, as discussed in examinations of intermediary liability dynamics (Watney, 2022). Economic harms also emerge as synthetic media infiltrates digital-financial transactions and retail payment systems, contributing to vulnerability in markets that depend heavily on trust and authenticity, a risk described in studies of digital-financial consumer exposure (Paglietti & Rabitti, 2022). These harms exacerbate a broader public sphere crisis in which pervasive misinformation erodes institutional trust, as identified in work analysing the democratic implications of social media governance (Guess et al., 2020).

The regulatory debates surrounding deepfakes span multiple legal regimes, including platform accountability structures, intermediary liability frameworks, and AI governance principles. Discussions of platform governance demonstrate that governments increasingly rely on gatekeeper-style interventions to address cross-border enforcement challenges (Hörnle, 2021). Scholars examining intermediary liability frequently argue that legal systems must address the tension between innovation incentives and the duty of care expected from digital platforms (Machado & Aguiar, 2023). Comparative analyses of Internet intermediary liability in various jurisdictions reveal inconsistencies in how responsibility is assigned for user-generated synthetic content (Ali et al., 2021). Broader debates in digital trade policy highlight how emerging technologies, including deepfakes, complicate cross-border regulatory coherence (Kira et al., 2022). Ethical critiques of AI systems further argue that structural power imbalances shape the harms created by algorithmic manipulation and emphasize the need for collective safeguards against technologically mediated deception (Dyson, 2022). These overlapping debates illustrate the growing need for integrated regulatory responses that acknowledge both technical complexity and global interdependence.

This article adopts a scientific narrative review using a descriptive analysis method to synthesize legal, regulatory, and scholarly debates on deepfakes. The aim of this study is to comprehensively examine the legal challenges of deepfakes with a focus on liability, harm, and regulatory responses.

2. Conceptual Foundations of Deepfakes

Deepfakes represent a technologically advanced form of synthetic media produced through machine-learning systems capable of generating or altering visual and auditory content with exceptional realism. At the core of this technology are generative adversarial networks, a class of AI models that learn to fabricate increasingly convincing outputs by training two neural networks against one another. Legal scholarship examining broader patterns of platform governance provides important

context for understanding how such technologies emerged within digital ecosystems characterized by rapid innovation and minimal ex-ante oversight, a trend noted in analyses of harmful content regulation (Gorwa, 2020). The evolution of digital platforms as primary vectors for user-generated content also accelerated the diffusion of synthetic media tools, and researchers studying the relationship between social media, democracy, and misinformation have emphasized how algorithmic amplification facilitates the spread of manipulated audiovisual material (Guess et al., 2020). This technological environment enabled deepfakes to transition from experimental research outputs into widely accessible tools embedded in social, political, and economic communication structures.

Although deepfakes are one subset of synthetic media, the broader category encompasses a wide range of computer-generated or computer-modified content, including AI-generated text, avatars, synthetic voices, and digitally altered images. Scholars examining platform liability and regulatory models have highlighted that the legal system often struggles to distinguish between forms of AI-generated content because the underlying mechanisms blur traditional boundaries between original creation, derivative work, and automated synthesis, a tension identified in discussions of intermediary duties of care (Machado & Aguiar, 2023). While synthetic media refers generally to any AI-produced output, deepfakes specifically involve the realistic manipulation of human identity or performance. This distinction is important because AI-generated content that does not depict or impersonate real individuals may raise ethical questions but typically lacks the same legal implications for defamation, privacy, or personal autonomy as deepfakes that map one person's likeness onto another's actions. Regulatory analyses exploring the limits of platform governance during election periods have emphasized that impersonation-based manipulation creates unique risks to democratic integrity by fabricating expressive acts in the name of real individuals (Gorwa, 2021).

Deepfakes can be characterized along a spectrum ranging from benign to malicious uses. Benign applications include entertainment, satire, accessibility tools, and creative experimentation, yet even these uses operate within a landscape where platform enforcement agencies confront ongoing challenges in distinguishing harmful manipulations from culturally legitimate forms of expression, a dilemma emphasized in research that explores legal responses to digital misinformation (Ray, 2021). Malicious deepfakes, by contrast, deliberately intend to deceive, harm, or manipulate, and they often intersect with categories of illegal or harmful content that social media intermediaries struggle to detect and remove at scale, as illustrated in work analyzing intermediary liability for harmful content (Watney, 2022). The growing availability of automated generation tools has increased the risk that malicious actors can deploy personalized deepfakes for fraud, extortion, political disruption, or reputational damage.

From a legal perspective, deepfakes raise critical issues of deception, authenticity, attribution, and agency. Establishing authenticity becomes particularly complex when manipulated audiovisual artifacts closely mimic evidence traditionally relied upon in judicial and administrative contexts. Scholars studying enforcement challenges across digital borders have noted that the ability to falsify expressive acts compromises foundational assumptions about identity and presence in online environments (Hörnle, 2021). Attribution is equally problematic because deepfake production often involves anonymous actors, fragmented creation chains, and the use of open-source tools distributed across multiple jurisdictions. This problem echoes broader concerns in digital trade governance, where scholars have observed that emerging technologies complicate efforts to establish coherent regulatory frameworks across borders (Kira et al., 2022). Questions of agency also arise because deepfakes destabilize the legal meaning of consent, authorship, and representation—issues that become more intricate in environments where corporate platforms mediate user interactions and exert significant influence over content visibility, as noted in critiques of structural digital power dynamics (Dyson, 2022).

The difficulty deepfakes pose for legal categorization stems from the fact that existing frameworks rely heavily on the historical reliability of human-authored audiovisual evidence and on clear distinctions between creators, distributors, and intermediaries. Researchers examining platform regulatory politics have shown that these assumptions no longer hold in ecosystems where automated systems can generate content that appears human-authored but lacks a stable origin point or traceable intent (Reich et al., 2022). Further complications arise as financial, consumer, and communication markets increasingly integrate digital technologies that depend heavily on trust, authenticity, and identity verification, making them more vulnerable to synthetic manipulation, a risk emphasized in analyses of digital-financial consumer vulnerabilities

(Paglietti & Rabitti, 2022). Deepfakes therefore challenge not only substantive legal rules but also the procedural and evidentiary foundations upon which modern legal systems operate.

3. Types of Harms Associated with Deepfakes

The harms associated with deepfakes extend across personal, political, economic, and societal dimensions, creating a layered risk landscape that current legal frameworks struggle to accommodate. One of the most prominent forms of injury is reputational harm, which arises when manipulated audiovisual content falsely attributes statements, actions, or behaviors to individuals. The ability of deepfakes to convincingly simulate a person's likeness enables character assassination efforts that are far more persuasive than traditional defamatory speech, and scholars examining digital misinformation have emphasized how algorithmic amplification intensifies the reach and durability of reputationally damaging content (Guess et al., 2020). The ease with which false videos can circulate across platforms has been linked to broader governance concerns, particularly within legal analyses exploring intermediary responsibility for harmful or illegal content, which note that platforms often lack sufficient tools to prevent reputational injury from cascading across networks (Watney, 2022). Studies of regulatory responses to online manipulation further highlight that reputational harm is magnified in political contexts, where false portrayals can discredit public figures or misrepresent policy positions, a dynamic explored in platform governance research addressing electoral communication distortions (Gorwa, 2021).

Privacy violations present another significant category of harm, particularly in the context of non-consensual sexual deepfakes and unauthorized identity misuse. Deepfake pornography, which frequently targets women, involves a severe intrusion into personal autonomy by fabricating sexual acts using an individual's face or body without consent. Scholars analyzing digital liability regimes have observed that such forms of synthetic exploitation exploit gaps in existing privacy and dignity protections, especially where legal systems rely on intent or demonstrable harm thresholds that are difficult to meet in digital spaces (Ali et al., 2021). The difficulty of attributing deepfake creation to identifiable perpetrators complicates enforcement, a challenge mirrored in broader analyses of cross-border digital regulation, where researchers describe how fragmented jurisdictions hinder coherent privacy protections (Hörnle, 2021). The growing availability of user-friendly synthetic media tools further exacerbates the risk of identity misuse, echoing concerns raised in analyses of digital-financial vulnerabilities in which technological complexity makes consumers increasingly exposed to manipulation and deception (Paglietti & Rabitti, 2022).

Political manipulation represents one of the most destabilizing harms associated with deepfakes. Synthetic videos can simulate political candidates, public officials, or activists making inflammatory statements, endorsing fabricated positions, or participating in staged misconduct. Research on disinformation and democratic resilience underscores that deepfakes intensify the possibility of coordinated influence operations, especially during election cycles where public perception is highly sensitive to misinformation (Ray, 2021). Scholars studying platform governance during elections have argued that synthetic audiovisual deception creates regulatory tensions because states attempt to impose rapid safeguards while platforms struggle to scale monitoring systems at the speed of digital dissemination (Gorwa, 2021). Broader analyses of digital misinformation highlight that deepfakes erode confidence in authentic political communication, undermining the informational integrity upon which democratic decision-making depends (Guess et al., 2020).

Economic harms are increasingly prevalent as deepfakes are used for fraud, impersonation, and intellectual property misuse. Synthetic voice replication has been used to deceive employees into transferring funds, while manipulated videos can impersonate executives or corporate representatives. Analyses of digital-financial consumer risk indicate that such deceptive tactics thrive in markets where trust and identity verification are central to economic transactions (Paglietti & Rabitti, 2022). Intellectual property questions also emerge when deepfakes recreate or imitate proprietary performances, datasets, or brand identities, echoing concerns raised in research examining structural challenges within digital markets and the power dynamics that shape technological exploitation (Reich et al., 2022). These economic harms demonstrate how deepfakes leverage vulnerabilities within highly digitized commercial ecosystems.

Beyond individual and sectoral harms, deepfakes contribute to a broader societal erosion of trust and an authenticity crisis. When the public becomes aware that any image or video can be fabricated, even authentic evidence risks being dismissed as falsified. Scholars analyzing harmful content regulation argue that this epistemic instability complicates efforts to govern digital spaces, as regulatory institutions depend on the ability to verify and authenticate expressive acts (Gorwa, 2020). Ethical critiques of AI further highlight that widespread deepfake proliferation amplifies collective disempowerment by undermining citizens' ability to rely on shared information environments (Dyson, 2022). This authenticity crisis strains legal doctrines built on evidentiary reliability, such as rules of witness credibility, documentary integrity, and good-faith reliance on audiovisual records.

Across all categories of harm, the central challenge lies in the misalignment between deepfake-related injuries and existing legal doctrines. Defamation law struggles with synthetic content that lacks a single identifiable author. Privacy law falters when violations occur without direct physical intrusion. Electoral law is undermined by cross-platform dissemination patterns that escape national jurisdiction. Economic fraud statutes require identifiable perpetrators, yet deepfakes often originate from anonymous or foreign actors. The multidimensional harm structure therefore reveals deepfakes as a phenomenon that strains the assumptions underlying many traditional legal frameworks, illustrating the need for adapted regulatory responses capable of addressing both the speed and complexity of AI-generated manipulation.

4. Liability Frameworks: Challenges in Assigning Responsibility

The challenge of assigning legal responsibility for deepfakes emerges from the complexity of the digital ecosystems in which they are created, disseminated, and amplified. Multiple actors participate in the lifecycle of a deepfake, and this plurality of involvement complicates conventional liability models premised on a clear separation between creators, intermediaries, and end-users. Scholars analyzing intermediary liability frameworks have noted that creators of harmful content frequently operate anonymously or in decentralized digital environments, making it difficult to distinguish between those who design the underlying algorithms, those who manipulate the audiovisual material, and those who upload it onto social media networks (Ali et al., 2021). The involvement of distributors further adds complexity because individuals who share or repost deepfakes often play a critical role in harm diffusion, yet existing legal doctrines commonly differentiate between primary creators and secondary disseminators. Analyses of platform governance have highlighted that digital platforms serve as essential conduits that determine the visibility and reach of synthetic content, and they operate within governance systems where their influence over information flows has raised concerns about democratic accountability (Gorwa, 2021). AI developers also occupy a significant position in the liability landscape because they produce the generative models that enable deepfake creation, and critiques of structural digital power have emphasized that the concentration of technological expertise within a handful of corporate actors raises questions about the responsibilities of those who design systems capable of large-scale manipulation (Reich et al., 2022). Malicious users, who often weaponize deepfakes for fraud, extortion, or political disruption, exploit the accessibility of these tools within online ecosystems already strained by harmful content governance challenges (Gorwa, 2020).

Doctrinal gaps emerge when attempting to map these actors onto traditional liability frameworks. Attribution represents one of the most significant challenges because deepfakes can be created using widely available software that leaves minimal forensic signatures, and scholars examining cross-border enforcement mechanisms note that anonymity remains a persistent impediment to effective regulation (Hörnle, 2021). Strict liability models, which assign responsibility without regard to intent, are difficult to apply to deepfake ecosystems because they risk imposing disproportionate burdens on developers or platforms whose systems may be misused without their active participation. Intent-based liability models face parallel limitations because malicious users frequently conceal their identities or operate from jurisdictions with weak enforcement structures. Studies examining harmful content regulation emphasize that intermediary liability doctrines, including safe harbor protections, complicate enforcement when platforms are shielded from liability provided they remove unlawful content upon notice (Watney, 2022). Legal analyses of regulatory transitions within content moderation frameworks have shown that debates around the duty of care attempt to bridge this doctrinal gap by requiring platforms to proactively mitigate foreseeable harms, yet such approaches remain unevenly implemented across jurisdictions (Machado & Aguiar, 2023).

Comparative approaches demonstrate that the legal treatment of deepfake-related harms varies widely between common law and civil law systems. Common law jurisdictions often rely on tort-based reasoning, requiring plaintiffs to demonstrate harm, causation, and attribution—elements that become particularly complex when dealing with synthetic media. Civil law systems, by contrast, may provide broader protections through statutory privacy or personality rights, yet these frameworks also struggle to accommodate the fluidity of AI-generated manipulation. In the United States, Section 230 of the Communications Decency Act grants broad immunity to platforms for user-generated content, creating significant barriers to holding intermediaries accountable for the spread of deepfakes, a problem heightened by political misinformation and harmful content governance debates (Guess et al., 2020). In the European Union, the Digital Services Act introduces more stringent obligations for large platforms, including systemic risk assessments and enhanced transparency, aligning more closely with proposals advocating stronger oversight mechanisms within platform ecosystems (Gorwa, 2021). The United Kingdom's Online Safety Act adopts yet another model, imposing duties of care on platforms while allowing for regulatory discretion in targeting high-risk harms, reflecting policy trends in which states increasingly view intermediaries as essential gatekeepers in digital enforcement. Across all these systems, cross-border enforcement remains a pervasive challenge, particularly when malicious deepfakes originate from jurisdictions with weak regulatory infrastructures or when digital trade policies complicate coordinated legal responses, as highlighted in analyses of emerging digital governance regimes (Kira et al., 2022).

Evidentiary and due process challenges further complicate liability frameworks. Authenticity disputes arise because deepfakes undermine confidence in audiovisual evidence traditionally used in court proceedings, mirroring concerns that technologically mediated deception destabilizes foundational assumptions about the reliability of expressive acts (Ray, 2021). Forensic detection tools, while increasingly sophisticated, face limitations due to rapidly evolving generative models, and legal scholars caution that relying on automated detection may raise new risks, including over-reliance on probabilistic assessments in judicial contexts. Ethical critiques of AI power dynamics emphasize that technological opacity undermines individuals' ability to contest synthetic representations of themselves, contributing to procedural inequities (Dyson, 2022). The difficulty of verifying content authenticity also intersects with broader concerns about digital marketplace vulnerabilities, where manipulation of identity-dependent data introduces further risks to legal certainty (Paglietti & Rabitti, 2022). As regulatory frameworks struggle to reconcile these evidentiary uncertainties with established legal procedures, deepfakes continue to expose the limitations of liability doctrines rooted in assumptions about traceability, intentionality, and the stability of digital identity.

5. Regulatory Responses and Policy Interventions

Regulatory responses to deepfakes have emerged unevenly across jurisdictions as governments attempt to mitigate the rapidly escalating risks associated with synthetic media manipulation. National approaches have increasingly moved toward the criminalization of malicious deepfakes, particularly those used for deception, harassment, fraud, or political interference. Several U.S. states have enacted laws targeting non-consensual sexual deepfakes and election-related manipulations, reflecting legislative concerns aligned with broader debates about disinformation and democratic stability that scholars have documented when analyzing digital threats to institutional legitimacy (Ray, 2021). South Korea and China have also enacted criminal provisions addressing synthetic media, with China requiring explicit labeling of AI-generated content, a move that aligns with arguments found in platform governance research demonstrating the need for traceability and accountability within digital ecosystems (Gorwa, 2021). These national measures increasingly incorporate civil liability mechanisms as well, allowing victims to pursue tort-based remedies for reputational damage or privacy violations. Comparative legal scholarship examining intermediary liability underscores the importance of expanding civil avenues because criminal measures alone fail to address the wide range of individualized harms generated by deepfake misuse (Ali et al., 2021). Election-related rules have similarly expanded, with several jurisdictions requiring disclaimers or prohibiting the spread of synthetic political content during specific periods, echoing concerns raised in research on political manipulation and harmful content regulation (Gorwa, 2020).

Regional and international frameworks have attempted to harmonize responses across borders, though such efforts remain nascent. Within the European Union, the Digital Services Act introduces systemic obligations for platforms to assess and mitigate risks associated with manipulative content, a regulatory direction consistent with calls for stronger oversight of

platform governance during high-stakes political processes (Gorwa, 2021). The EU AI Act further supplements this approach by establishing risk-based classifications that impose stricter requirements on high-risk AI systems, reflecting scholarly critiques emphasizing that unchecked technological design choices can exacerbate structural harms in digital societies (Reich et al., 2022). These regional frameworks represent early attempts to integrate deepfake concerns into broader digital governance structures. The Council of Europe has also explored soft-binding recommendations addressing synthetic media transparency and human rights implications, drawing from global assessments of harmful content regulation that advocate for cross-border coherence in legal responses (Watney, 2022). Although these initiatives mark progress, they underscore the challenge of designing international frameworks capable of regulating technologies that transcend territorial boundaries.

Soft-law and non-legal governance mechanisms supplement formal regulation by establishing industry-led standards and platform-specific policies. Industry standards have emerged through multistakeholder initiatives focused on authenticity infrastructure, watermarking protocols, and AI transparency commitments. These voluntary frameworks align with scholarly recognition that platforms play a central regulatory function in digital ecosystems where state enforcement mechanisms are often insufficient (Machado & Aguiar, 2023). Leading platforms such as YouTube, Meta, and TikTok have implemented policies restricting harmful manipulated content, removing non-consensual sexual deepfakes, and labeling altered media. These efforts reflect earlier empirical observations that platform interventions can help reduce the spread of harmful content when combined with algorithmic transparency and oversight (Guess et al., 2020). Content labeling, watermarking, and verification systems represent increasingly important governance tools because they aim to restore public trust in audiovisual evidence, a concern also present in digital financial governance research demonstrating the need for authentication safeguards in markets vulnerable to technological exploitation (Paglietti & Rabitti, 2022). Although these mechanisms offer additional layers of protection, platform-driven governance remains inconsistent and often opaque, reflecting structural power imbalances that critics argue shape the digital public sphere (Dyson, 2022).

Despite this evolving regulatory ecosystem, significant gaps and limitations persist. Overregulation remains a concern because overly broad restrictions on synthetic media risk constraining legitimate expression, including satire, artistic experimentation, and political speech. Scholars examining harmful content governance caution that regulatory approaches must carefully balance security concerns with free speech protections to avoid suppressing democratic discourse (Gorwa, 2020). Jurisdictional inconsistencies also impede effective governance because national laws differ in scope, enforcement mechanisms, and definitional clarity, leading to fragmented legal landscapes that malicious actors can exploit by operating across borders. This fragmentation mirrors the challenges observed in digital trade governance, where emerging technologies have made it increasingly difficult for states to align regulatory priorities and enforcement capacities (Kira et al., 2022). Enforcement challenges further complicate these efforts, particularly when platforms enjoy safe harbor protections that limit liability for user-generated content, a tension identified in studies examining the boundaries of intermediary responsibility (Watney, 2022). Even where legal obligations exist, the evidentiary difficulties of verifying manipulated content hinder practical enforcement, reflecting broader concerns about the fragility of authentication norms in digital societies (Ray, 2021).

Overall, regulatory responses to deepfakes remain in a formative stage, shaped by a combination of national legislation, regional governance instruments, and voluntary industry frameworks. While progress continues across multiple fronts, significant structural and doctrinal gaps persist, underscoring the need for more coordinated, evidence-based, and technologically informed policy approaches.

6. Toward a Coherent Legal Framework: Synthesis and Future Directions

The emergence of deepfakes demands an integrated legal framework capable of coordinating civil, criminal, platform governance, and AI regulatory approaches. Fragmented responses have proven insufficient because deepfakes operate across interconnected technological and social environments that blur distinctions between creators, intermediaries, and victims. Scholars examining platform governance have shown that platforms exert significant structural influence over information flows, highlighting the need for regulatory models that integrate civil liability with broader systemic oversight (Gorwa, 2021). Civil remedies provide essential avenues for individuals harmed by reputational injury or privacy violations, yet they require

complementary criminal measures to deter malicious actors who rely on anonymous or transnational distribution channels, a difficulty noted in cross-border enforcement analyses (Hörnle, 2021). AI regulatory approaches, including risk-based classifications and developer accountability measures, offer an additional dimension by addressing the technological roots of synthetic manipulation, echoing critiques that emphasize how the design and deployment of digital systems can amplify structural harms (Reich et al., 2022). The challenge lies in harmonizing these diverse legal mechanisms into a coherent system that can respond effectively to both immediate and systemic risks.

Harmonized standards—both technical and legal—are essential for building this integrated framework. Technical standards such as content authenticity tools, watermarking protocols, and traceability mechanisms help restore trust in digital communication environments, aligning with observations in digital-financial vulnerability research that authentication safeguards are necessary to counter technologically mediated deception (Paglietti & Rabitti, 2022). Legal harmonization likewise requires aligning definitions, enforcement mechanisms, and jurisdictional rules to prevent regulatory gaps that malicious actors can exploit. Scholars studying harmful content regulation argue that regulatory coherence improves the predictability and efficacy of platform governance systems by reducing inconsistencies across national laws (Gorwa, 2020). Without coordinated standards, platforms and AI developers face unclear obligations that weaken both preventive and remedial measures.

Several proposals can contribute to a coherent regulatory model. A duty of care for AI developers and platforms would impose proactive obligations to mitigate foreseeable harms, reflecting arguments that intermediaries must assume greater responsibility for moderating harmful synthetic media within their ecosystems (Machado & Aguiar, 2023). Mandatory watermarking or authenticity indicators would support verification processes and reduce evidentiary disputes, aligning with platform governance efforts to label manipulated content (Gorwa, 2021). Cross-border cooperation is essential because deepfakes can be created in one jurisdiction and deployed in another with minimal traceability, a problem underscored in research examining jurisdictional fragmentation in digital regulation (Kira et al., 2022). AI governance sandboxes offer another avenue by allowing regulators, developers, and civil society to experiment with oversight models in controlled environments, enabling evidence-based policy design and reducing risks associated with premature or overly restrictive regulation. These proposals, collectively, highlight the need for multilayered interventions that reflect the technological complexity and global reach of deepfakes.

Despite growing scholarly attention, several research gaps remain. More empirical work is needed to assess the effectiveness of detection tools, especially given the rapid evolution of generative models. Cross-jurisdictional comparative studies are necessary to understand how legal definitions and enforcement mechanisms can be aligned across different systems. Further examination is required to determine how structural power imbalances between platforms and users influence deepfake harm distribution, an issue raised in critiques of digital inequality (Dyson, 2022). Additionally, scholars need to study how societal trust erodes in environments where synthetic media becomes pervasive, building on democratic misinformation research that documents the destabilizing effects of manipulated content (Guess et al., 2020). Addressing these gaps will be essential for developing regulatory systems that can withstand the challenges posed by increasingly sophisticated synthetic media technologies.

7. Conclusion

The rapid proliferation of deepfake technologies marks a profound turning point in the relationship between individuals, digital platforms, and the legal systems tasked with protecting fundamental rights and societal stability. As synthetic media becomes more sophisticated, accessible, and seamlessly integrated into everyday digital communication, it challenges long-standing assumptions about identity, authenticity, truth, and accountability. The analysis across this review demonstrates that deepfakes do not represent a singular or isolated threat; instead, they operate at the intersection of technological innovation, social vulnerability, political structures, and economic systems. Their capacity to manipulate appearances and fabricate expressive acts destabilizes essential foundations of trust that underpin both private interactions and democratic institutions.

The harms associated with deepfakes unfold across multiple dimensions, illustrating the diverse and increasingly severe consequences of synthetic manipulation. Reputational harms undermine personal dignity and damage careers, relationships,

and social standing. Privacy violations, particularly non-consensual sexual deepfakes, constitute profound invasions of autonomy and bodily integrity, disproportionately targeting women and vulnerable individuals. Political harms threaten democratic legitimacy by enabling misinformation, distorting electoral processes, and eroding public confidence in authentic communication. Economic harms exploit trust-based relationships, allowing fraud, impersonation, and market manipulation to occur with unprecedented ease. At a broader societal level, deepfakes contribute to an authenticity crisis in which citizens begin to doubt not only fabricated content but also real events, weakening shared epistemic foundations and amplifying social fragmentation.

The legal challenges surrounding deepfakes are equally multidimensional. Existing legal doctrines struggle to map onto the complex lifecycle of synthetic media, in which creation, distribution, amplification, and harm occur across decentralized and transnational digital networks. Traditional liability models, built around traceable authorship and identifiable intent, prove inadequate in environments where anonymous actors can generate and disseminate convincing fabrications within minutes. Intermediary liability frameworks further complicate these issues, as platforms simultaneously serve as facilitators of communication, amplifiers of harmful content, and potential regulators of their own ecosystems. As a result, legal systems face the difficult task of balancing innovation, speech rights, and safety while confronting actors whose operations often transcend jurisdictional boundaries.

Although national and regional regulatory initiatives have begun to address these challenges, the landscape remains fragmented. Some jurisdictions focus on criminalization, others on civil remedies, and still others on systemic platform governance. These approaches, while valuable, are insufficient when applied in isolation. Deepfakes require a holistic response that integrates civil law protections, criminal sanctions, regulatory oversight, technological safeguards, and cross-border cooperation. The emergence of broader digital governance frameworks such as the Digital Services Act, AI regulation initiatives, and platform duty-of-care models demonstrates growing recognition of the need for systemic alignment. Yet implementation remains uneven, and enforcement often lags behind technological development.

A coherent legal framework for addressing deepfakes must begin with recognition of the layered nature of the harm and the distributed nature of responsibility. It requires policymakers to adopt forward-looking regulatory designs that anticipate rather than merely react to technological evolution. Duty-of-care obligations for platforms and AI developers can help shift the burden toward proactive risk mitigation rather than reactive removal. Technical safeguards such as watermarking, authenticity indicators, and content provenance systems offer essential tools for restoring trust in audiovisual evidence. International cooperation mechanisms are critical for addressing cross-border dissemination and ensuring that malicious actors cannot exploit inconsistent or weak regulatory environments. Experimental regulatory spaces, such as AI governance sandboxes, create opportunities for testing oversight mechanisms that can adapt as the technology evolves.

Equally important is the need to account for the human and societal dimensions of deepfake harm. Legal responses must be sensitive to the disproportionate impact on marginalized groups, the psychological trauma associated with identity manipulation, and the democratic risks posed by the erosion of shared reality. Regulatory models should therefore be developed through inclusive, participatory processes that involve civil society, technologists, legal scholars, affected communities, and industry stakeholders. Only through broad engagement can policymakers craft mechanisms that are both effective and socially legitimate.

Finally, addressing deepfakes requires sustained research, empirical analysis, and theoretical development. Legal scholarship must continue to explore how synthetic media reshapes evidentiary norms, challenges assumptions about authorship and consent, and alters the distribution of power in digital environments. Technological researchers must advance detection and authentication tools, while social scientists examine the psychological and societal impacts of widespread media manipulation. Coordinated interdisciplinary inquiry will be essential for creating flexible, resilient regulatory frameworks.

In sum, deepfakes represent a complex, evolving phenomenon that tests the boundaries of current law and governance. Building effective safeguards will depend on integrating diverse legal mechanisms, aligning global standards, strengthening technological infrastructure, and deepening our understanding of the social conditions that enable synthetic manipulation to flourish. By adopting a comprehensive, forward-looking approach, legal systems can better protect individuals, preserve democratic integrity, and sustain public trust in an era where the boundaries between the real and the artificial continue to blur.

Ethical Considerations

All procedures performed in this study were under the ethical standards.

Acknowledgments

Authors thank all participants who participate in this study.

Conflict of Interest

The authors report no conflict of interest.

Funding/Financial Support

According to the authors, this article has no financial support.

References

Ali, A. H. S., Saidin, O. K., Roisah, K., & Ediwarman, E. (2021). Liability of Internet Intermediaries in Copyright Infringement: Comparison Between the United States and India. https://doi.org/10.4108/eai.29-6-2021.2312595

Dyson, M. R. (2022). Combatting AI's Protectionism &Amp; Totalitarian-Coded Hypnosis: The Case for AI Reparations &Amp; Antitrust Remedies in the Ecology of Collective Self-Determination. *Smu Law Review*, *75*(3), 625. https://doi.org/10.25172/smulr.75.3.7

Gorwa, R. (2020). The State of Global Harmful Content Regulation: Empirical Observations. *Aoir Selected Papers of Internet Research*. https://doi.org/10.5210/spir.v2020i0.11221

Gorwa, R. (2021). Elections, Institutions, and the Regulatory Politics of Platform Governance: The Case of the German NetzDG. https://doi.org/10.31235/osf.io/2exrw

Guess, A. M., Barberá, P., Siegel, A., Woolley, S., Fowler, E. F., Nielsen, R. K., Wittenberg, C., Fukuyama, F., Keller, D., Hwang, T., Gorwa, R., & Persily, N. (2020). Social Media and Democracy. https://doi.org/10.1017/9781108890960

Hörnle, J. (2021). The Jurisdictional Challenge Answered—Enforcement Through Gatekeepers on the Internet. 33-80. https://doi.org/10.1093/oso/9780198806929.003.0003

Kira, B., Tavengerwei, R., & Mumbo, V. (2022). Points À Examiner À l'Approche Des Négociations De Phase II De La ZLECAf: Enjeux De La Politique Commerciale Numérique Dans Quatre Pays d'Afrique Subsaharienne. https://doi.org/10.35489/bsg-dp-wp_2022/01

Machado, C., & Aguiar, T. H. (2023). Emerging Regulations on Content Moderation and Misinformation Policies of Online Media Platforms: Accommodating the Duty of Care Into Intermediary Liability Models. *Business and Human Rights Journal*, *8*(2), 244-251. https://doi.org/10.1017/bhj.2023.25

Paglietti, M. C., & Rabitti, M. (2022). A Matter of Time. Digital-Financial Consumers' Vulnerability in the Retail Payments Market. *European Business Law Review*, *33*(Issue 4), 581-606. https://doi.org/10.54648/eulr2022027

Ray, A. (2021). Disinformation, Deepfakes and Democracies: The Need for Legislative Reform. *University of New South Wales Law Journal*, *44*(3). https://doi.org/10.53637/dels2700

Reich, R., Sahami, M., & Weinstein, J. M. (2022). System Error: Where Big Tech Went Wrong and How We Can Reboot. *Perspectives on Science and Christian Faith*, *74*(1), 62-64. https://doi.org/10.56315/pscf3-22reich

Watney, M. (2022). Regulation of Social Media Intermediary Liability for Illegal and Harmful Content. *European Conference on Social Media*, *9*(1), 194-201. https://doi.org/10.34190/ecsm.9.1.104