www.ijarp.com ISSN 2456-9992 Page: 01-17

International Journal Advanced Research Publications

Review Article

Volume: 02
Issue: 03

CRIMINAL LIABILITY OF AI INTERMEDIARIES: EVALUATING THE "SAFE HARBOUR" PROTECTION FOR SOCIAL MEDIA PLATFORMS IN THE AGE OF VIRAL DEEPFAKES

Aditi Sonal*[1] Dr. Mudra Singh[2]

[1]Research Scholar, Amity Law School, Lucknow.

[2]Assistant Professor, Amity Law School, Lucknow.

Article Received: 5 February 2026, Article Revised: 25 February 2026, Published on: 18 March 2026

***Corresponding Author: Aditi Sonal**

Research Scholar, Amity Law School, Lucknow.

DOI: https://doi-doi.org/101555/ijarp.5649

ABSTRACT

The proliferation of hyper-realistic synthetic media, commonly known as deepfakes, has fundamentally challenged the legal frameworks governing online content. The "safe harbour" provisions, which have historically shielded social media platforms from liability for user-generated content, are facing unprecedented strain. This paper examines the tension between these immunities and the urgent need to address the multifaceted harms caused by deepfakes, which range from electoral disinformation to gender-based violence. Through a comparative analysis of legislative responses in India, the European Union, the United Kingdom, and the United States, this paper argues that the passive intermediary model underpinning safe harbour is conceptually ill-suited to an ecosystem where platforms actively curate content and increasingly integrate generative AI tools. It concludes that a recalibrated framework is necessary, one that moves beyond notice-and-takedowntowards proactive duties of care, while carefully preserving constitutional free speech guarantees. A differentiated approach, distinguishing between a platform's traditional hosting functions and its role as an AI service provider, offers the most viable path forward.

KEYWORDS: Deepfakes, Safe Harbour, Intermediary Liability, Section 230, Generative AI, Criminal Responsibility, Information Technology Act, Online Safety Act.

1. INTRODUCTION

In early 2024, a sexually explicit deepfake video of the renowned Indian actor Rashmika Mandanna went viral across multiple social media platforms, sparking national outrage and prompting the Prime Minister to call for action against the misuse of artificial intelligence (AI). This incident was not an isolated one but a harbinger of a new digital reality. Just two years later, in January 2026, the platform X (formerly Twitter) faced the threat of a ban in the United Kingdom after its AI chatbot, Grok, was found to be generating non-consensual sexualised images of women and children in response to user prompts. These events underscore a fundamental crisis: the legal architecture designed to foster an open internet is now struggling to contain the harms of synthetic media.

For over two decades, the growth of the internet has been underpinned by "safe harbour" protections. Enshrined in Section 230 of the US Communications Decency Act of 1996 and Section 79 of India's Information Technology Act, 2000, these laws provide platforms with immunity from liability for content created and posted by their users. This "notice-and-takedown" regime was predicated on a model of the platform as a passive conduit, a neutral bulletin board for third-party speech. It was never designed for a world where AI can fabricate convincing evidence of events that never occurred, impersonate any individual with photographic realism, and weaponise this content at scale and at speed. A 2024 McAfee survey found that a staggering 75% of Indian respondents had encountered deepfakes, with 38% personally targeted by a deepfake-enabled scam.

The core challenge posed by deepfakes is that their harms are often incompatible with the reactive logic of safe harbour. The damage be it to reputation, psychological well-being, or democratic discourse occurs in the initial moments of viral spread, long before a takedown notice can be issued and processed. Furthermore, the lines of responsibility have blurred. Platforms are no longer just hosts; they are architects of engagement through algorithmic amplification and, increasingly, direct providers of AI generation tools, as the Grok incident demonstrates.

This paper seeks to answer a central question: Can the existing safe harbour framework be reconciled with the imperative to hold platforms accountable for deepfake-related harms, or does it require fundamental reconfiguration? This paper argues that the passive intermediary model is conceptually obsolete. It proposes a recalibrated framework for criminal liability that distinguishes between platforms' hosting functions and their novel responsibilities as AI service operators. Part II examines the theoretical foundations of safe harbour. Part III

analyses the distinctive harms of deepfakes. Part IV evaluates recent comparative legislative responses. Part V proposes a path forward for a differentiated liability framework.

2. The Safe Harbour Doctrine: The Legal Immunity of Intermediaries

2.1 Theoretical Underpinnings and Policy Rationales

The safe harbour doctrine is built upon a foundational legal distinction: the difference between a publisher and a distributor (or intermediary). A publisher, such as a newspaper, exercises editorial control over content and is therefore strictly liable for what it publishes. An intermediary, by contrast, merely provides the infrastructure for others to communicate and should not be treated as the "speaker" of third-party information.

Section 230 of the US Communications Decency Act (CDA) is the most influential expression of this principle. It famously states, "No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider." This provision was a direct response to early internet defamation cases that threatened to stifle the nascent online ecosystem. Congress sought to create a "forum for a true diversity of political discourse, unique opportunities for cultural development, and myriad avenues for intellectual activity."

India adopted a parallel framework under Section 79 of the Information Technology Act, 2000. It grants intermediaries immunity from liability for third-party content, provided they "observe due diligence" and do not "initiate the transmission, select the receiver of transmission, and select or modify the information contained in the transmission." The core policy rationales are common across jurisdictions: to preserve free speech by preventing over-censorship, to acknowledge the practical impossibility of pre-screening all user content, and to foster innovation by protecting fledgling companies from ruinous liability. As senior advocate Sajan Poovayya notes, the safe harbour is "fundamental to the functioning of a free, open to all and borderless internet."

2.2 The Conditional Nature of Immunity

Safe harbour is not an absolute shield. Both Section 230 and Section 79 condition immunity on the platform's adherence to specific responsibilities, creating a "notice-and-takedown" regime. In India, Section 79 requires intermediaries to "expeditiously" remove or disable access to unlawful content upon receiving actual knowledge, typically from a government agency or court order. The Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021, further elaborate these "due diligence" requirements. They

oblige platforms to appoint grievance officers, publish clear terms of service, and remove content depicting non-consensual intimate images or pornography within 24 hours of receiving a complaint.

This framework reflects a legislative compromise: platforms are not required to proactively monitor, but they must cooperate in removing clearly illegal content once identified. The underlying assumption is that the primary harm lies in the continued *availability* of the content, and that removal upon notification is an adequate remedy.

2.3 The Erosion of the Publisher-Intermediary Distinction

The technological reality of the 21st century has severely eroded the conceptual foundation of the publisher-intermediary distinction. Social media platforms are far from passive. Their algorithmic recommendation systems actively curate, promote, and amplify certain content while suppressing others, a function that is undeniably editorial in nature. By designing systems that prioritise engagement, and given that engagement metrics often favour sensational or emotionally charged deepfakes, platforms become active participants in the content's success.

The integration of generative AI directly into platforms has shattered the distinction entirely. When a platform like X provides a tool like Grok that can generate illegal content at a user's behest, it is no longer acting as an intermediary for third-party content. It is the direct *creator* of the content, albeit one prompted by a user. This shift demands a corresponding shift in legal accountability. As Zhang Futian argues in the context of AI-facilitated fraud, the emergence of AI systems does not change the fact that machines lack consciousness, but it radically alters the nature of human agency and control in the value chain.

3. The Distinctive Challenge of Deepfakes

3.1 The Nature and Scale of the Harm

Deepfakes inflict harms that are qualitatively different from traditional forms of harmful content. Firstly, they exploit the epistemic trust we place in audiovisual media. Seeing is believing, and deepfakes weaponise this cognitive bias to create a "liar's dividend," where the mere possibility of fabrication can be used to deny authentic evidence.

Secondly, deepfakes enable scalable personalisation. The same technology can be used to create a fake video of a politician or a sexually explicit image of a private individual, using nothing more than publicly available photographs from social media. This has led to an

epidemic of gender-based violence online, with women being disproportionately targeted by non-consensual deepfake pornography.

Thirdly, deepfakes exploit the viral velocity of social media. The "notice-and-takedown" model is structurally inadequate for harms that materialise in the first few hours of a video's release. By the time a takedown notice is processed, the reputational damage is often complete, the financial market may have been manipulated, or the disinformation may have already shaped public opinion.

3.2 The Attribution Problem and the Limits of Criminal Law

Attributing criminal liability for deepfake harms is a complex doctrinal challenge. Criminal law requires a culpable human actor. In the AI supply chain, several potential actors exist: the individual user who prompts the AI, the platform that hosts the tool, the developer who trained the model, and sometimes even the victim who made their image publicly available.

Chinese legal scholarship offers a useful analytical framework by distinguishing between "self-developed" and "technology-integrated" AI service operators. Self-developed operators have complete control over the technical chain, including model training and data selection. They possess the greatest capacity to prevent harmful outputs and should therefore bear a stricter duty of care. Technology-integrated operators, who customise third-party models for specific applications, have more limited control confined to interface permissions, invocation frequency, and parameter settings. Their criminal responsibility should be calibrated to the scope of their actual control.

This distinction maps directly onto the social media context. A platform that merely hosts user-uploaded deepfakes has a different relationship to the harm than a platform that provides an integrated AI generation tool like Grok. The former may lack knowledge of the content; the latter has designed the very system that produces it and can monitor for patterns of misuse. To treat them identically under the safe harbour framework is a category error.

3.3 The Inadequacy of Reactive Frameworks

The notice-and-takedown model places the entire burden of enforcement on the victim. It requires them to locate the content across multiple platforms, document it, navigate often opaque reporting mechanisms, and await a decision. This burden is unsustainable for an individual targeted by dozens or even hundreds of deepfakes. Moreover, the reactive model fails to account for platforms' role in amplification. Even if a platform eventually removes the

content, its algorithm may have already served it to millions of users. The harm is not just in the content's existence, but in its promotion.

4. Comparative Legislative Responses

In response to these challenges, jurisdictions worldwide are moving beyond the traditional safe harbour framework.

4.1 India: Proactive Diligence and Labelling

India's approach has evolved from reactive takedowns to proactive obligations. The November 2025 amendments to the Intermediary Guidelines introduce the concept of "synthetically generated information," defined as content "artificially or algorithmically created, generated, modified or altered." The amendments mandate that such content be prominently labelled, with disclaimers covering a significant portion of the visual or audio medium. Crucially, significant social media intermediaries (SSMIs) "must remove such content using reasonable efforts," and their failure to do so "may cause them to lose their safe harbour protection."

This shift from reactive to proactive duties represents a fundamental reconceptualisation of the intermediary's role. Platforms are now expected to identify and label synthetic content independently, effectively collapsing the distinction between platform and publisher. While this creates a powerful transparency regime, it also raises concerns about over-compliance and the chilling of legitimate speech, such as AI-generated art or satire.

4.2 United Kingdom: Systemic Risk and Duties of Care

The UK's Online Safety Act 2023 takes a different tack, focusing on systemic risk management rather than individual pieces of content. It imposes a duty on platforms to assess the likelihood of their services being used to disseminate illegal content and to take proportionate measures to mitigate those risks. The Act empowers Ofcom, the communications regulator, to investigate and impose significant sanctions for non-compliance. The threat of a ban against X over the Grok incident demonstrates the coercive power of this regulatory model, which targets the platform's governance and processes rather than just its content.

4.3 United States: Carving Out Intellectual Property

The proposed NO FAKES Act of 2024 in the US takes a third approach. It creates a federal intellectual property right in an individual's voice and likeness, and explicitly exempts claims

under this right from Section 230 immunity. This allows individuals to sue both the creators and the hosting platforms for unauthorised digital replicas. Like the Indian model, it includes a notice-and-takedown safe harbour for platforms that remove content "expeditiously." Critics, however, warn that vague exceptions for "bona fide news" or parody could lead platforms to over-remove content to avoid litigation, chilling protected speech.

4.4 European Union: The Product Safety Model

The EU's AI Act adopts a product safety approach. It classifies AI systems by risk level and imposes strict transparency obligations on general-purpose AI models, including requirements for technical documentation and compliance with copyright law. For deepfakes, it mandates clear disclosure of AI-generated content. The Act's focus on the entire "value chain" of AI, from developer to deployer, provides a more holistic framework for addressing the risks posed by these technologies.

5. Towards a Recalibrated Framework for Liability

The preceding analysis demonstrates that the existing safe harbour framework, with its binary distinction between passive intermediary and active publisher, is no longer fit for purpose. A more nuanced, function-based approach is required.

5.1 A Functional Taxonomy of Platform Roles

An effective regulatory framework for digital platforms must begin by recognizing that not all platforms perform the same functions or exercise the same level of control over online content. In contemporary digital ecosystems, platforms can operate in several distinct roles, each involving different levels of technological capability, editorial influence, and responsibility. Treating all platforms as though they perform identical functions risks creating an overly simplistic regulatory system that either imposes excessive burdens on some actors or fails to hold others accountable for their influence. Consequently, the first step toward designing a balanced and effective legal framework is to distinguish between qualitatively different platform roles. A useful way to conceptualize these roles is by identifying three primary categories: the Host, the Curator, and the AI Service Provider. Each of these roles involves different forms of interaction with user-generated content and therefore requires a different set of legal duties and liabilities.

The first category is the Host. A host platform primarily provides the technical infrastructure that allows users to upload, store, and access digital content. In this role, the platform functions as an intermediary rather than as a creator or editor of the material that appears on

its service. Examples of hosting activities include cloud storage services, file-sharing platforms, or websites that allow users to post content without actively selecting or promoting specific material. Because host platforms typically do not exercise direct control over the substance of user-generated content, it would be unreasonable to expect them to prevent every instance of unlawful or harmful material from appearing online. As a result, many legal systems have adopted the notice-and-takedown model as the primary mechanism for regulating such intermediaries. Under this model, platforms are not automatically liable for content posted by users; however, once they receive a credible notification that specific material violates the law or platform policies, they must act promptly to remove or disable access to that content.

The notice-and-takedown framework strikes a pragmatic balance between protecting freedom of expression and addressing harmful or illegal material. By requiring action only after a platform becomes aware of problematic content, the system avoids imposing an unrealistic obligation to monitor every piece of information uploaded by millions of users. At the same time, it ensures that platforms cannot ignore complaints about unlawful content once they are informed about it. Nevertheless, improvements to the traditional notice-and-takedown model may be necessary in order to address contemporary challenges posed by large-scale digital platforms. For instance, procedures for submitting complaints should be accessible and transparent so that individuals who are harmed by unlawful content can easily request its removal. Platforms should also respond to such notices in a timely manner, minimizing the period during which harmful material remains accessible to the public. Furthermore, users whose content is removed should be provided with explanations and opportunities to challenge the decision, thereby safeguarding procedural fairness and preventing unjustified censorship.

The second category of platform function is the Curator. Unlike a host platform that simply stores and displays user-generated content, a curator platform actively shapes the visibility and reach of that content through algorithmic recommendation systems. Modern social media services often rely on sophisticated algorithms that analyze user behavior, preferences, and engagement patterns in order to recommend posts, videos, or articles that are likely to attract attention. These recommendation systems determine which pieces of content appear prominently in news feeds, trending sections, or suggested viewing lists. As a result, curator platforms exert a significant influence over the flow of information within the digital environment. Their algorithms do not merely display content; they selectively amplify certain materials while reducing the visibility of others.

This curatorial role introduces a different dimension of responsibility because algorithmic amplification can significantly magnify the social impact of harmful or misleading content. For example, synthetic media such as manipulated videos or AI-generated images may remain relatively obscure when viewed by only a small group of users. However, if a platform's recommendation algorithm promotes such material to millions of users based on its capacity to generate engagement, the potential harm increases dramatically. In such cases, the platform cannot be viewed merely as a passive intermediary. Instead, its design choices and algorithmic priorities contribute directly to the dissemination and visibility of the content. Therefore, a legal framework that distinguishes between hosting and curating functions can more accurately reflect the platform's role in shaping the digital information ecosystem.

For platforms acting as curators, liability should not focus solely on individual pieces of content posted by users. Attempting to assign responsibility for every harmful post would be both impractical and ineffective given the enormous volume of online material. Instead, regulatory attention should focus on the systemic design and operation of algorithmic amplification systems. If evidence demonstrates that a platform's algorithms consistently promote sensational, misleading, or harmful synthetic content because such material generates high engagement, regulators may reasonably question whether the system has been designed with adequate safeguards. In such circumstances, liability could arise not from the mere existence of harmful content but from the structural characteristics of the recommendation system that encourage its spread.

This systemic approach shifts the focus of regulation from isolated moderation decisions to the broader architecture of digital platforms. Regulators may examine factors such as the incentives embedded within recommendation algorithms, the extent to which platforms conduct risk assessments, and the safeguards implemented to prevent the amplification of harmful content. For instance, platforms could be required to implement mechanisms that reduce the visibility of content identified as misleading or manipulated, especially when it relates to sensitive topics such as elections, public health, or personal reputation. If a platform knowingly maintains algorithmic systems that prioritize engagement at the expense of safety and accuracy, and if those systems repeatedly amplify harmful synthetic media, such behavior could justify regulatory sanctions or corrective measures.

The third category of platform function is the AI Service Provider. This role involves platforms that offer integrated generative artificial intelligence tools enabling users to create new forms of digital content. Unlike hosts or curators, AI service providers do not merely store or distribute existing material; they supply the technological systems that generate text,

images, audio, or video based on user prompts. These generative tools represent a significant technological advancement, but they also raise unique regulatory challenges. Because the platform itself provides the mechanism through which new content is created, it possesses a greater degree of control over the design, capabilities, and limitations of the technology.

In this context, the responsibilities of AI service providers extend beyond traditional content moderation. They must consider the potential risks associated with the outputs generated by their systems and implement safeguards to prevent foreseeable misuse. For example, generative AI systems could be designed to refuse requests that attempt to produce illegal material, impersonate real individuals, or fabricate deceptive media intended to mislead the public. Platforms might also implement watermarking technologies that allow AI-generated content to be identified more easily, thereby improving transparency and accountability. By integrating such safeguards into the design and operation of their AI tools, service providers can reduce the likelihood that their technologies will be used for harmful purposes.

Recognizing these three distinct roles host, curator, and AI service provider allows policymakers to create a regulatory framework that assigns responsibilities in proportion to the platform's capabilities and influence. A host platform that simply stores user-generated content may reasonably operate under a notice-and-takedown regime. A curator platform that amplifies content through algorithmic recommendation systems may face obligations related to the design and impact of those systems. An AI service provider that enables the creation of new digital material may bear additional responsibilities related to the safe design and deployment of generative technologies. By tailoring legal duties to these roles, regulators can avoid imposing uniform obligations that fail to account for the diversity of modern digital platforms.

For a platform acting as an **AI Service Provider**, the most stringent duties apply. As Zheng and Zeng argue in the Chinese context, the level of control should determine the scope of responsibility. A provider like X with its Grok chatbot has a duty to conduct pre-deployment risk assessments, implement robust technical safeguards (e.g., filtering prompts and outputs), and monitor for patterns of misuse. Failure to do so, especially after becoming aware of harms, could give rise to criminal liability for recklessness or negligence.

5.2 The Case for a Differentiated Duty of Care

Criminal liability is the most severe form of legal responsibility imposed by the state, as it carries serious consequences such as fines, sanctions, reputational damage, and in certain circumstances even imprisonment of responsible individuals. Because of these grave

consequences, criminal law traditionally reserves liability for the most serious and blameworthy forms of conduct. In the rapidly evolving digital environment, particularly with the emergence of artificial intelligence (AI) technologies and platforms that enable the creation and dissemination of content such as deepfakes, determining when criminal liability should arise has become increasingly complex. Platforms that host or develop AI systems often serve as intermediaries rather than direct creators of harmful content. Therefore, imposing blanket criminal liability for every harmful outcome associated with AI-generated content could be both unjust and counterproductive. A more balanced and principled approach is the adoption of a differentiated "duty of care" framework, which calibrates the level of responsibility according to the platform's role, technological capacity, and degree of control over the AI systems it deploys.

The concept of a duty of care originates from legal principles that require individuals or organizations to act with reasonable caution to avoid causing harm to others. When applied to digital platforms and AI developers, this duty of care recognizes that companies operating powerful technologies have a responsibility to anticipate and mitigate foreseeable harms arising from their tools. However, the level of responsibility should correspond to the extent of influence and control the platform exercises over the technology. Platforms that merely host user-generated content without significant control over the design or functioning of AI tools may have a different level of responsibility compared to companies that design, deploy, and maintain advanced generative AI systems capable of producing synthetic media. By differentiating responsibilities based on roles and capacities, the duty of care framework prevents both over-criminalization and regulatory gaps, ensuring that liability is imposed only when a platform has failed to act responsibly despite having the capacity to prevent harm.

One important dimension of this duty of care framework is the obligation related to design duties. Design duties require platforms and developers to create AI systems that are safe by default. This means that safety and ethical considerations must be integrated into the architecture of the technology from the earliest stages of development. Rather than treating safety as an afterthought, developers must anticipate foreseeable misuse scenarios and incorporate technical safeguards that reduce the likelihood of harmful applications. For instance, AI systems capable of generating images or videos could include built-in filters that prevent the creation of explicit or illegal content. Similarly, watermarking technologies and traceability mechanisms can help identify AI-generated material and discourage malicious uses such as the production of deceptive deepfakes. By implementing such safeguards during the design phase, developers can significantly reduce the probability that their technologies

will be used for unlawful purposes. Failure to incorporate reasonable safety measures, especially when risks are well known, may indicate negligence or recklessness and could justify stronger legal consequences.

Another crucial component of the duty of care is the obligation of monitoring duties. Monitoring duties require platforms to actively observe how their technologies are being used and to identify patterns of misuse. In the digital environment, harmful activities often occur at scale and can spread rapidly across networks. Waiting passively for individual users to report harmful content is often insufficient, as it allows damage to occur before any intervention takes place. Instead, platforms are expected to develop and deploy detection systems capable of identifying suspicious behavior, abnormal usage patterns, or content that violates legal and ethical standards. For example, if an AI tool begins to show patterns indicating that users are repeatedly generating harmful or illegal materials, the platform should be able to detect these trends through automated monitoring tools or algorithmic analysis. Effective monitoring does not necessarily mean constant surveillance of every user interaction, but it does require reasonable steps to ensure that the platform remains aware of significant risks associated with its technology. This proactive approach aligns with the broader principle that companies benefiting from powerful technological systems must also bear responsibility for preventing foreseeable misuse.

The third element of the duty of care framework involves response duties. Even with strong design and monitoring mechanisms, it is impossible to eliminate all risks associated with complex AI technologies. Therefore, platforms must also demonstrate the capacity to respond quickly and effectively when harmful activities are detected. Response duties include taking immediate action to remove illegal or harmful content, restricting or suspending accounts responsible for misuse, and modifying or disabling features that facilitate harmful outcomes. In serious cases, platforms may also need to cooperate with law enforcement authorities or relevant regulatory bodies. The effectiveness of a platform's response is often judged by the speed and adequacy of its actions once it becomes aware of the harm. A platform that promptly addresses misuse and implements corrective measures demonstrates responsible behavior, whereas a platform that ignores warnings or delays intervention may be seen as failing to fulfill its duty of care.

Importantly, the duty of care framework does not automatically impose criminal liability on platforms for every harmful piece of content that appears on their services. The mere presence of a deepfake or other harmful AI-generated material on a platform does not by itself establish criminal wrongdoing on the part of the platform operator. In an open digital

environment where millions of users interact simultaneously, it is unrealistic to expect platforms to prevent every instance of misuse. Criminal liability arises only when there is clear evidence of a culpable failure to fulfill the duties described above. In other words, liability depends not simply on the existence of harm, but on whether the platform knowingly or recklessly ignored its responsibilities despite having the capacity to prevent or mitigate the harm.

A practical illustration of this principle can be seen in situations where a platform becomes aware that its AI tool is being widely used to generate illegal content such as child sexual abuse material. If credible evidence shows that the platform received repeated warnings, internal reports, or external notifications about this misuse yet failed to implement safeguards, monitoring systems, or effective responses, such conduct may demonstrate a culpable disregard for legal obligations. In such cases, criminal liability may be justified because the platform's failure goes beyond mere oversight and reflects a conscious neglect of its duty of care. Conversely, if a platform actively attempts to prevent misuse through safety features, monitoring mechanisms, and prompt responses, it would generally not be appropriate to impose criminal sanctions even if isolated incidents of misuse occur.

This differentiated approach to liability offers several advantages. First, it encourages responsible innovation by providing clear expectations for technology developers without discouraging legitimate technological progress. Second, it promotes accountability by ensuring that companies cannot escape responsibility when they knowingly allow harmful uses of their technologies to continue. Third, it protects fundamental legal principles by reserving criminal punishment for situations involving genuine culpability rather than unavoidable technological limitations. By focusing on the platform's conduct and its adherence to the duty of care, the legal system can strike a careful balance between safeguarding society from harm and preserving the benefits of technological advancement.

5.3 Preserving Free Speech and Ensuring Proportionality

Any effort to recalibrate legal liability for digital platforms must carefully consider the unintended consequences that may arise from stricter regulatory obligations. One of the most significant risks in this context is the phenomenon often described as "collateral censorship." This term refers to a situation in which online platforms, in an attempt to avoid legal penalties or reputational damage, remove or restrict large amounts of user-generated content even when such content may be lawful or socially valuable. When platforms face the possibility of heavy fines, criminal liability, or regulatory sanctions for hosting prohibited material, they

may adopt an overly cautious approach to moderation. As a result, rather than conducting a careful assessment of each piece of content, platforms might simply remove or block anything that appears even remotely risky. While such behavior may reduce legal exposure for the platform, it can simultaneously undermine the fundamental right to freedom of expression and limit the diversity of voices in the digital public sphere.

The risk of collateral censorship becomes particularly significant in the context of artificial intelligence–generated content, including deepfakes, synthetic media, and algorithmically generated text or images. These technologies can be used for harmful purposes such as misinformation, harassment, or exploitation, but they also have legitimate applications in fields such as education, entertainment, art, journalism, and political commentary. If legal frameworks impose strict liability on platforms for hosting potentially harmful AI-generated material, companies may respond by adopting sweeping restrictions on such content. In practice, this could lead to the removal of legitimate works of parody, satire, artistic experimentation, or political critique that rely on synthetic media techniques. Such outcomes would create a chilling effect on creativity and democratic discourse, discouraging individuals from exercising their right to communicate ideas, challenge authority, or engage in cultural expression through digital technologies.

The concept of a chilling effect is closely connected to concerns about collateral censorship. When individuals believe that their speech may be arbitrarily removed, restricted, or punished, they may choose to remain silent rather than risk negative consequences. This self-censorship can gradually erode the openness of the online environment and weaken the democratic function of digital platforms as spaces for debate, dissent, and public participation. Freedom of expression has long been recognized as a foundational value in democratic societies because it allows individuals to exchange ideas, criticize institutions, and participate in cultural and political life. Therefore, any regulatory system designed to address the harms associated with online content must also ensure that legitimate speech is not unnecessarily suppressed. Striking this balance is one of the most challenging aspects of modern digital governance.

To address this challenge, any new liability framework must incorporate strong procedural safeguards that protect the rights of users while still enabling effective regulation of harmful content. One essential safeguard is the establishment of clear and narrowly defined categories of prohibited content. Vague or overly broad definitions can create confusion and encourage platforms to interpret legal requirements in an excessively restrictive manner. For example, if a law simply prohibits “harmful” or “offensive” content without specifying the exact criteria,

platforms may feel compelled to remove a wide range of material to avoid legal risk. In contrast, precise definitions help ensure that enforcement actions target only clearly unlawful or dangerous material, such as content that incites violence, facilitates exploitation, or violates privacy rights. By limiting the scope of prohibited content to well-defined categories, regulators can reduce the likelihood that legitimate expression will be mistakenly removed.

Another important safeguard involves the recognition of meaningful exceptions for forms of expression that play a vital role in cultural and political discourse. Parody, satire, and artistic expression have historically served as powerful tools for social critique and creative exploration. Satirical works often imitate or exaggerate real individuals, events, or institutions in order to highlight contradictions or expose wrongdoing. Similarly, artistic projects frequently experiment with new technologies, including AI-generated imagery or video manipulation, to produce innovative forms of storytelling or commentary. If regulatory frameworks fail to recognize these legitimate uses of synthetic media, they risk stifling creative freedom and cultural innovation. Therefore, laws governing digital platforms should explicitly acknowledge that certain uses of AI-generated content, even when they resemble deceptive or manipulated media, may be permissible when they clearly serve purposes such as parody, satire, research, or artistic experimentation.

In addition to clear definitions and expressive exceptions, accessible appeal mechanisms are another critical component of a fair regulatory system. Content moderation decisions are often made quickly and at large scale, sometimes through automated systems or algorithmic filters. While these technologies can help identify harmful material efficiently, they are not infallible and may produce errors. Legitimate content may be mistakenly flagged or removed due to misinterpretation by automated systems or human moderators. Without effective mechanisms for review, such mistakes can lead to unjust restrictions on users' ability to communicate and share ideas. An accessible appeal process allows users to challenge moderation decisions and request reconsideration of their content. Ideally, such processes should be transparent, timely, and conducted by trained reviewers who can assess the context and intent behind the content.

Furthermore, liability must be proportionate. The goal is not to bankrupt platforms but to align their incentives with public safety. A graduated approach, starting with regulatory sanctions and escalating to civil and finally criminal liability for the most egregious and knowing failures, is the most appropriate path.

6. CONCLUSION

The age of viral deepfakes has exposed the foundational weaknesses in the legal architecture that governs the internet. The safe harbour doctrine, a product of the 1990s, is conceptually incompatible with the active, algorithmic, and increasingly generative role of modern social media platforms. To treat a host of user-generated text and a provider of a powerful AI image generator under the same legal framework is no longer tenable.

The solution, however, is not to abandon safe harbour entirely. Its core principles protecting free speech and fostering innovation remain vital. The task is to build upon this foundation, creating a more sophisticated and functional legal framework for the AI age. This requires moving beyond the reactive logic of notice-and-takedown towards a proactive framework of differentiated duties of care. By distinguishing between a platform's role as a host, a curator, and an AI service provider, the law can more precisely target the source of the harm. Platforms that design, deploy, and profit from powerful AI tools must bear the corresponding responsibility to ensure they are not used as weapons of mass deception and abuse. The future of a trustworthy and open internet depends on getting this balance right.

REFERENCES

1. Zheng Zexing and Zeng Qinwen, 'On the Criminal Responsibility of Generative Artificial Intelligence Service Operators' (2025) Aisixiang https://www.aisixiang.com/data/169829.html
2. Probir Roy Chowdhury, 'AI deepfakes go mainstream, exposing limits of safeguards and regulations' (JSA, 6 January 2026) https://www.jsalaw.com/jsa_news/al-deepfakes-go-mainstream-exposing-limits-of-safeguards-and-regulations/
3. Gowling WLG, 'Too late now to say sorry?: Rethinking personality rights in the era of dupes and deepfakes' (11 December 2025) https://gowlingwlg.com/en/insights-resources/articles/2025/rethinking-personality-rights-in-the-era-of-dupes-and-deepfakes
4. Sajan Poovayya, 'Deepfakes here's the relevance of safe harbour protection' (CNBC TV18, 13 March 2024) https://www.cnbctv18.com/information-technology/deepfakes-digital-india-act-safe-harbour-protection-information-technology-act-sajan-poovayya-19255261.htm
5. Futian Zhang, 'The Dilemma and Solutions in Applying Criminal Law to Generative AI Fraud Crimes: The Case of Sora' (2025) 4(5) Studies in Law and Justice 78

6. Lewis Nedas, 'Grok in the Dock: What X's AI scandal reveals about legal risks ahead' (17 February 2026) https://lewisnedas.co.uk/grok-in-the-dock-what-xs-ai-scandal-reveals-about-legal-risks-ahead/
7. Shwetashree Majumder, 'Deepfake regulation: It needs to be smarter, not stricter' Mint Hyderabad (22 December 2025)
8. Dhruv Anand and Dhananjay Khanna, 'Fighting deepfakes needs nimble but realistic laws' (Law.asia, 28 November 2025) https://law.asia/india-deepfake-regulations/
9. Jeffrey Westling and Barbara Batycka, 'The Pros and Cons of the NO FAKES Act' (American Action Forum, 13 November 2024) https://www.americanactionforum.org/insight/the-pros-and-cons-of-the-no-fakes-act/