Advances in Social Behavior Research Vol.16 Issue 4 EWA Publishing
Available Online: 10 June 2025 DOI: 10.54254/2753-7102/2025.23777

Normative construction of platform criminal liability in the governance of deepfake technology

Xuanting Liu

School of Law, Guangzhou Xinhua University, Dongguan, China

liuxuanting@xhsysu.edu.cn

Abstract. Online platforms play a critical intermediary role in the generation and dissemination of deepfake content and should accordingly bear criminal liability commensurate with their status. However, China's current legal framework faces practical obstacles in defining platform criminal liability, including ambiguity in the delineation of duties, difficulties in establishing the "knowledge" standard, and procedural asymmetry in accountability mechanisms—factors which severely undermine the effectiveness of criminal law governance. To address these challenges, the normative construction of platform criminal liability must be grounded in the theoretical foundation of legal interests protected by criminal law, clearly define platforms' proactive duties of care, and introduce reasonable standards for the presumption of knowledge. At the same time, a differentiated and tiered compliance obligation system should be established, utilizing technological means to ensure transparency and verifiability of platform responsibility. Only in this way can criminal law regulation achieve an effective balance between safeguarding social security and fostering technological innovation.

Keywords: deepfake, platform criminal liability, technology governance, criminal law regulation

1. Introduction

Law has never existed in isolation; it is always rooted in specific social contexts and technological environments. The advent of the artificial intelligence era has made the relationship between technological governance and legal regulation increasingly complex, and the rise of "deepfake" technology vividly exemplifies this complexity. Since 2017, when a Reddit user first used this technology to synthesize and widely disseminate fake pornographic videos of celebrities, deepfake technology has rapidly drawn intense attention and regulatory responses worldwide. The United States has successively enacted the "Malicious Deepfake Prohibition Act" and the "Deepfake Accountability Act," explicitly defining the responsibilities of both content creators and platforms. The European Union, through the General Data Protection Regulation (GDPR) and the Code of Practice on Disinformation, has imposed constraints from the perspectives of data governance and algorithmic regulation.

By comparison, China issued several legal documents in 2023, including the "Regulations on the Administration of Deep Synthesis of Internet Information Services" and the "Interim Measures for the Administration of Generative Artificial Intelligence Services." However, the regulation of technologies such as "deep synthesis" and "generative artificial intelligence" in these documents demonstrates a pronounced tendency toward "technological neutrality." They fail to provide a clear and forceful negative assessment of the criminal harmfulness of deepfakes, resulting in an ambiguous theoretical foundation for criminal regulation. Therefore, how to define the boundaries of criminal regulation for platforms in the governance of deepfakes and to clarify their criminal liability has become an urgent issue that needs to be addressed.

At present, the malicious use of deepfake technology has already caused serious social harm. For example, an incident occurred at a university where AI technology was used to synthesize and disseminate fake pornographic videos, causing the female victims severe psychological trauma and damage to their dignity. However, legal remedies are often limited to administrative penalties, which fall far short of addressing the actual harm suffered by the victims. According to data from the Ministry of Public Security, since 2022, nearly a hundred cases of telecom fraud involving "AI face swapping" have occurred nationwide, resulting in economic losses exceeding 200 million yuan. Clearly, deepfake technology poses a real threat to individual rights, property security, and even societal trust. In the chain of deepfake information dissemination, online platforms play a critical role. Modern platforms not only control distribution channels but also possess content recommendation algorithms, allowing them to deeply intervene in the creation and dissemination of information. Therefore, platforms cannot simply claim to be neutral technological channels; they

must bear corresponding managerial responsibilities. The proactive or passive actions of platforms directly determine the scope and extent of harm caused by the dissemination of deepfake content.

Platforms are not outsiders to the risks posed by deepfakes; rather, they are significant participants in shaping and amplifying such risks. This article argues that it is necessary to re-examine the theoretical basis and regulatory pathways for criminal liability of platforms under criminal law. On one hand, clarifying the criminal liability of platforms is an inevitable response to public expectations for the protection of security, order, and individual rights. On the other hand, it is also an important means to promote platforms' proactive implementation of technological governance, thereby forming a collaborative governance structure involving technological, administrative, and criminal regulatory measures. On this basis, the abuse of deepfake technology can be effectively curbed, and the law can genuinely fulfill its governance function and value in the age of artificial intelligence.

2. The legal interest foundation and normative basis for platform criminal liability

From the perspective of criminal law's protection of legal interests, the responsibility of platforms to participate in the governance of deepfakes is both well-founded and legitimate. Firstly, online platforms control vast amounts of user data and content distribution channels, possessing strong management and organizational capabilities. Platforms are able to utilize artificial intelligence algorithms to analyze uploaded content, detect signs of illegal deepfake activity, and take measures such as deletion or blocking to control it. The technical and managerial capabilities of platforms are critical for preventing deepfake-related harm, as they are in a position to detect and block the dissemination of unlawful content. Criminal law assigns specific safety management obligations to entities with such control, in line with the legal logic that "the greater the power, the greater the responsibility"—that is, those who have the ability to prevent harm are obliged to do so. When a platform has stronger means than individuals to identify fake deepfake messages and protect the public from deception and harm, assigning them a criminal law duty to act is precisely to effectively safeguard legal interests such as personal information, reputation, privacy, and even public information security.

Secondly, there is a strong causal connection between the intermediary actions of platforms and the harms caused by deepfakes, as well as a genuine possibility for platform-enabled harm. Platforms, as amplifiers of content dissemination, are often the direct reason why the consequences of deepfake information are magnified due to their inaction. For example, in rumor-based deepfake incidents, if a platform's recommendation algorithm pushes sensational fake videos to more users, it undoubtedly exacerbates the damage to the victim's reputation and the degree of public deception. In the well-known "Qvod case," a video platform knowingly allowed users to upload large amounts of pornographic content and turned a blind eye, ultimately being found guilty as a joint offender in the dissemination of obscene materials due to profiting from illegal content. Thus, the inaction of platforms can easily become a part of the perpetration of harm, and the consequences can be no less severe than those of the direct actors of deepfake crimes. From the principle of criminal law restraint but not withdrawal, there is a necessity to include platforms as subjects of criminal law obligations; one cannot allow them to escape responsibility merely due to their intermediary position, nor can one overly emphasize technological neutrality while ignoring the actual role of platforms in the occurrence of harm. Some scholars argue that platforms often invoke "technological autonomy" as a defense [1], claiming they merely provide technical intermediation and have no subjective intent. However, the precondition for this defense is that the platform must prove it has fulfilled its reasonable duty of care. When a platform, equipped with advanced technology, fails to supervise clearly illegal deepfake content, its claim to neutrality is undermined, and there is both factual and legal justification for holding it criminally liable.

Finally, from the normative perspective, existing Chinese laws provide a certain basis for pursuing platform liability in the dissemination of deepfakes. According to Article 47 of the Cybersecurity Law, network operators are required to immediately stop transmission, take removal measures to prevent information dissemination, preserve records, and report to the relevant authorities when they discover information prohibited by laws and administrative regulations. This provision clearly establishes the statutory duty of platforms to handle illegal information. If a platform knowingly allows the dissemination of illegal deepfake content without intervention, this constitutes a failure to fulfill statutory management obligations. Article 286-1 of the Criminal Law further incorporates such conduct into the criminal law framework. When a network service provider fails to perform the information network security management obligations stipulated by laws or administrative regulations, and, after being ordered by regulatory authorities to make corrections, refuses to do so, resulting in the widespread dissemination of illegal information or other serious consequences, it constitutes a crime punishable by up to three years' imprisonment or a fine; for units committing this crime, both the unit and directly responsible individuals are subject to penalties. This offense, commonly known as the "crime of refusing to perform information network security management obligations," is aimed at punishing platforms that seriously neglect their supervision of illegal information. The harmful content generated by deepfake technology clearly falls within the category of "illegal information" prohibited by laws and regulations; once a platform allows such content to be widely disseminated and refuses to implement regulatory rectification orders, it may constitute this crime. In this sense, Article 286-1 provides an existing normative foundation under criminal law for platforms' involvement in deepfake governance. The application of this provision reflects the need to protect legal interests (preventing the societal harm caused by the widespread dissemination of illegal deepfake information) and also aligns with the subsidiary nature of criminal law (intervening only when administrative regulatory measures are insufficient to correct platform behavior and serious consequences result).

It is worth noting that the boundaries of Article 286-1's application should be rationally expanded and understood in light of the deepfake context. On one hand, the provision requires regulatory authorities to first issue a corrective order, and only if the platform refuses to comply does criminal intervention occur. This means that only when a platform is aware of its management negligence (made clear through regulatory notifications or similar means) yet remains passively inactive and causes serious harm, will criminal law intervene. The intent behind this design is to provide platforms with an opportunity to correct their conduct and avoid premature criminal intervention. However, in the case of rapidly spreading, novel risks like deepfakes, waiting for regulatory notification before acting is often too late, as illegal content may have already spread widely. Therefore, how to interpret the requirement of "refusal to make corrections after being ordered" and whether law enforcement response times should be shortened in urgent deepfake situations is a matter that needs further discussion regarding the boundaries of application. On the other hand, Article 286-1 lists consequences such as "widespread dissemination of illegal information," "leakage of user information," "destruction of evidence in criminal cases," and provides "other serious circumstances" as a catch-all category. For example, if a large number of fake, harmful synthesized videos are disseminated or biometric information is leaked through face-swapping technology, and if the harm caused by deepfake conduct does not clearly fall within the specified categories, "other serious circumstances" may still be invoked. However, in judicial practice, judges tend to take a cautious approach to what constitutes "serious" [2]. Therefore, in applying the existing provision, it is necessary to comprehensively consider factors such as the scale and duration of dissemination, the number of victims, and the degree of loss in order to ensure that criminal law covers platform dereliction in deepfake governance without overextending its scope. While platforms have a solid legal interest foundation and legal basis for being criminal law obligors in deepfake governance, there remain certain difficulties and limitations in the practical application of current norms. It is necessary to further sort out these real-world dilemmas and seek improvements.

3. Practical dilemmas and jurisprudential contradictions in regulating deepfake platforms

3.1. "Technological transfer" of duty sources and the abstraction of regulation

The rapid evolution of deepfake technology has led to a trend of "technological transfer" in the source of regulatory obligations, meaning that legal norms are continually incorporating new technological variables and requiring platforms to assume corresponding duties of technical governance. However, as technological updates far outpace the law's capacity to respond, many obligations are set out in principle-based or general terms, resulting in regulations that are abstract and ambiguous, and leaving platforms at a loss when it comes to implementation. For example, under the deep synthesis management regulations promulgated in 2023, service providers are required to add "user-unobtrusive marks" or "conspicuous labels" to certain synthesized information—a typical technical obligation, such as watermarking or labeling AI-generated content. In contrast, Article 286-1 of the Criminal Law only generally refers to "information network security management obligations stipulated by laws and administrative regulations," without specifying concrete technical measures. There are huge differences in technical conditions between platforms of different industries and sizes, and the realization of the same obligation may vary greatly in practice. Judicial authorities must first confirm whether relevant technical obligations exist and whether the platform has breached them, but this often involves specialized technical assessments. For instance, if a platform fails to add a watermark to a particular deepfake video, was it due to technical limitations in recognition, or was it negligence in fulfilling the labeling obligation? Similarly, regulations require algorithmic recommendation service providers to label unmarked synthesized information once detected, but what exactly "detection" means depends on the algorithm's capability and threshold settings, and the law does not specify what sensitivity level is required. As a result, legal obligations remain at an abstract level; so long as platforms argue "technical limitations" or "best efforts," regulators often struggle to determine whether obligations were genuinely breached. Many obligations in the deepfake arena—such as compliant data training, content watermarking, or opening source-identification interfaces—are all cutting-edge technical measures. Once incorporated into regulations but lacking operational detail, criminal regulation is left with ambiguous standards [3].

Furthermore, the weakening effect of the "technological neutrality" doctrine on regulatory efficacy cannot be ignored. Chinese legislation often uses the term "deep synthesis" when referring to deepfakes, emphasizing the neutrality of the technology itself. While "technology itself is not guilty" is a sound principle, if one overemphasizes neutrality while neglecting the social harm of technological applications, regulation risks lagging behind actual risks [4]. Current regulations emphasize that platforms should strengthen algorithmic and data governance, but rarely directly evaluate the infringement of deepfake content on personality rights or property rights. Objectively, this can easily give platforms the illusion that as long as they formally comply with technical obligations (such as filing, labeling, etc.), they can claim "technological neutrality" even if their platforms are used to commit deepfake crimes. In contrast, criminal law concerns itself with actual harmful outcomes and subjective fault; the generalization and abstraction of technical obligations may obscure the connection between platform conduct and social harm, raising the threshold for accountability. Thus, the technological transfer of duty sources requires a balance in criminal regulation: acknowledging the importance of technical obligations while avoiding excessive abstraction that would hinder legal enforceability.

3.2. The predicament of presuming knowledge under the "notice-and-takedown" model

For a long time, content regulation on online platforms has followed a "notice-and-takedown" model: platforms have no pre-screening obligation for user-posted information, but must promptly delete illegal content after being notified by a right holder or regulatory authority—otherwise, they may face liability. This model is common in civil and administrative contexts, reflecting the "safe harbor" principle for platforms. However, in the criminal sphere, the "notice-and-takedown" approach poses the challenge of determining whether a platform had knowledge ("knew" or "should have known") of illegality. Under criminal law theory, only when a platform "knows" of illegal conduct and nevertheless allows it to continue does subjective culpability arise. Yet if platforms uniformly claim "ignorance," law enforcement often struggles to prove actual knowledge unless there is evidence of official notification or user reports. Article 286-1 of the Criminal Law, by requiring "refusal to correct after being ordered by regulatory authorities," provides an objective standard for establishing knowledge. The problem, however, is that the spread of many harmful deepfakes is concealed and sudden, and victims often only realize the harm after it has already occurred. Under the notice-and-takedown model, platforms can claim ignorance before receiving notice and thus evade liability; even if they delete content upon notification, there is no criminal risk. This system allows passive and negligent conduct by platforms to go unpunished: as long as no report or regulatory attention is drawn during the dissemination of illegal content, the platform's acquiescence escapes legal sanction, which no doubt encourages a sense of impunity among some platforms. For deepfake victims, who often lack the capacity to promptly detect infringement, their rights are left unprotected before notification. When seeking accountability afterward, it is hard to prove the platform's subjective indifference, as the content will have already been removed. Because platforms tend to be passive—waiting for others to find problems rather than proactively monitoring—this predicament in determining "knowledge" may lead to wrongful acquittal of negligent platforms, weakening the criminal law's guidance toward proactive supervision.

Attempts to resolve this at the criminal law level encounter a jurisprudential dilemma. If the standard for "knowledge" is lowered, adopting presumed knowledge or "should have known," these risks conflicting with the principle of culpability and could lead to over-criminalization. But if one insists on strict proof of subjective intent, many derelictions that deserve punishment will escape criminal regulation. For example, some scholars argue that for clearly illegal deepfake content, such as explicit synthetic pornography or overt fraudulent inducements, platforms should be presumed to have known, even without direct evidence that staff actually viewed the content [5]. However, there is currently no clear "red flag rule" or legal presumption of knowledge in Chinese criminal law, and the adoption of such a rule would require legislative or judicial clarification. In sum, under the current notice-and-takedown model, proving knowledge is a dilemma: without presumption, wrongful acquittal; with presumption, risk of wrongful conviction. This contradiction must be addressed by clarifying the boundaries of the platform's duty of care and improving evidentiary rules.

3.3. Burden of proof and procedural asymmetry in platform regulation

Even if it is established that a platform objectively failed to fulfill its obligations and subjectively acted with willful disregard, the criminal prosecution process still faces challenges of proof and procedural asymmetry. First, key evidence is often controlled by the platform, while law enforcement and victims are far less capable of obtaining it. Whether the platform received user complaints about deepfake content, whether it reviewed particular content, how its internal risk controls operate—these are all documented within the platform's internal databases and communication records, and are not publicly accessible. Unless an official criminal investigation is initiated and compulsory measures are taken, it is difficult for outsiders to access direct evidence. After incidents occur, platforms often quickly delete problematic content and update logs, or even "whitewash" themselves by claiming they never detected anomalies. Reversal of the burden of proof is of limited use in criminal proceedings, and in most cases, the prosecution still bears the burden of proof for the defendant's guilt [6]. As a result, this information asymmetry greatly increases the difficulty of proving platform dereliction. Because AI-related criminal behavior is especially covert and complex, gathering evidence is even more difficult; for "technical omission crimes" like platform dereliction in deepfakes, investigators must not only possess ordinary legal evidence-collection skills, but also utilize technical tools to audit the platform's data logs, which is costly and challenging [7].

Second, procedural asymmetry is reflected in the mismatch between the cost of rights protection and response mechanisms. Deepfake victims are often individually weak, while platforms possess legal and technical teams to handle complaints and investigations. On one hand, victims must initiate police reports to prompt criminal investigation, while public security organs decide resource allocation based on case impact and available clues. If deepfake infringement is regarded as an isolated or "not serious" incident, the initiation of criminal proceedings becomes difficult. Conversely, even when a case is initiated, platforms can use their resources to vigorously contest proceedings, or influence public opinion to steer cases in their favor. In many past cases, corporate executives have argued that their companies had compliance systems in place or that they themselves were not directly responsible, seeking to mitigate liability. Under such circumstances, the contest between individual victims and large platforms is extremely lopsided. Third, from a procedural design perspective, charges like Article 286-1 require administrative authorities to first order corrective action, meaning criminal proceedings largely depend on prior and coordinated administrative enforcement. If administrative oversight is inadequate or delayed, criminal accountability is unattainable. Moreover, some platform violations may be more appropriately dealt with through administrative fines or content takedowns, with criminal prosecution considered a

last resort. With this mindset, investigative agencies may be reluctant to initiate criminal proceedings, preferring administrative penalties instead, thereby weakening the deterrent effect of criminal law.

Thus, the regulation of platforms suffers from severe procedural asymmetry: platforms hold an advantage in evidence and resources, while victims and public authorities are disadvantaged in detection, evidence collection, and litigation. This asymmetry undermines the ability of criminal law to safeguard platform obligations. It is necessary to improve legal mechanisms to balance the procedural positions of all parties and enhance the enforceability of criminal duties for platforms.

4. Paradigm shift in platform criminal duties and pathways for restructuring liability

4.1. Reconstruction of the duty of knowledge: introducing presumptions and the standard of reasonable care

To resolve the dilemma of platform knowledge under the "notice-and-takedown" model, it is necessary at the level of criminal law to reconstruct platforms' duty of knowledge by establishing a composite mechanism that combines objective presumptions with subjective fault assessment. On the one hand, the legislature or judicial interpretation may introduce a "should have known" standard, such that for manifestly illegal deepfake content, the platform's inaction can be deemed as constructive knowledge. For example, in cases involving the public dissemination of AI-generated pornographic videos or major fake information about public figures—where illegality is obvious according to social common sense—the platform is presumed to have discovered and addressed the issue. If it fails to act, such inaction should be regarded as "knowing" and give rise to criminal liability. Of course, to avoid the over-expansion of presumed knowledge, this presumption should be restricted to a minority of cases involving particularly severe circumstances where illegality is readily apparent, and the platform should be afforded the opportunity to present rebuttal evidence (such as proof that technical reasons prevented detection).

On the other hand, a mechanism for reviewing reasonable care obligations should be introduced. Platforms should be required to demonstrate, according to industry standards, that they took sufficient preventive measures in advance. Once a deepfake infringement incident occurs, if a platform claims ignorance, it must prove that it has fulfilled a substantial portion of these obligations. In other words, the platform must provide evidence that it exercised reasonable care, thereby excluding its own fault. If the platform cannot present compelling evidence and only makes vague claims of "never having received notice," it should be presumed to have acted with willful disregard. The combination of a duty of care and a presumption of knowledge will incentivize platforms to proactively fulfill their monitoring responsibilities in daily operations—since, once an incident arises, passively waiting for notification will no longer serve as a shield from liability; only by taking preventive measures and demonstrating diligence in advance can platforms be exempted from criminal responsibility. Through this reconstruction of obligations, platforms' tendency to "avoid knowing or being unwilling to know" in order to evade liability will be replaced by a compliance culture of active monitoring and preventive risk management.

4.2. Tiered compliance obligations: building a differentiated platform compliance system

Platforms vary greatly, and their obligations in deepfake governance should not be subject to a one-size-fits-all approach, but should instead be differentiated according to the nature and risk profile of the platform. For general social or video-sharing platforms—which are major channels for deepfake dissemination—it is essential to require robust content review mechanisms, with a focus on areas prone to deepfakes (e.g., political rumors, sexual exploitation), thus achieving a combination of ex-ante prevention and ex-post response. For technology-oriented platforms (such as face-swapping applications), the emphasis should be on technical safeguards against misuse, such as embedding digital watermarks or restricting high-risk functions. Furthermore, large, leading platforms with a vast user base and wide influence, given the greater harm caused by any negligence, should be held to higher compliance standards, including allocating more reviewers, deploying advanced recognition technologies, and establishing dedicated risk-warning teams. For small and medium-sized platforms, requirements may be appropriately reduced to avoid stifling innovation due to excessive compliance costs, but the basic duty to address illegal content cannot be omitted. Additionally, regulatory authorities can periodically disclose platform risk ratings based on their performance in deepfake governance, with high-risk platforms subject to stricter routine inspections and supervision. Such tiered obligations facilitate optimal allocation of resources and matching of responsibility, ensuring that key platforms genuinely fulfill their duties while avoiding the inefficiencies and inequities of blanket regulation.

At the level of criminal liability, tiered compliance obligations also serve as the basis for assessing whether a platform has exercised due care and for determining the severity of culpability. On the one hand, platforms that strictly comply with the relevant tiered obligations should be evaluated leniently in criminal proceedings. For example, if a large platform has implemented advanced AI review systems and provides round-the-clock monitoring, yet a few exceptionally covert deepfakes still slip through, it may be deemed to have made its best efforts and should not be held easily liable—at the very least, this should be a mitigating factor. On the other hand, platforms that are required to meet higher-level obligations but fail to act accordingly should face stricter accountability. For instance, if a large platform is well aware of the severe risks posed by deepfakes but fails to allocate sufficient review personnel or make the necessary technical investments, such an attitude of willful disregard cannot be excused by the claim of "not having received adequate notification" once serious harm occurs. It is worth noting that "compliance" should be the

baseline for delineating the boundaries of platform criminal liability: as long as a platform can demonstrate that it has an effective and robust compliance system in operation, criminal law should intervene with restraint. Conversely, a laissez-faire approach lacking compliance should fall squarely within the regulatory scope of criminal law. Recent compliance reform pilots by the Supreme People's Procuratorate also suggest that for first-time offenders who actively rectify, non-prosecution or lenient punishment may be considered. This mechanism could be extended to deepfake governance, encouraging platforms to develop robust self-regulation systems and giving opportunities for rectification when compliance failures occur, while resolutely punishing those who refuse to comply or recklessly allow risks. Through the design of tiered and differentiated compliance obligations, platforms can be motivated to take a more proactive stance against deepfake crimes and, at the same time, the system will allow for precise and proportionate enforcement.

4.3. Verifiable mechanisms integrating technology and criminal law: strengthening evidence and transparency of duty performance

To address the challenges of supervising platform performance and the risk of evidence destruction, it is essential to establish verifiable mechanisms combining technology and criminal law, thereby enhancing transparency and traceability of platform duty performance. First, improve log retention and audit systems. The law may require platforms to record key actions related to content review and disposal and retain these logs for a designated period. For example, each deleted or released suspicious deepfake content should have associated human review opinions or algorithmic decision rationale logged. In case of disputes, such logs become critical evidence to determine whether the platform has exercised due care. If a platform fails to retain logs or refuses to provide them as required, it should be subject to adverse inference—regardless of subjective intent, it should be deemed to have failed its review obligations. Second, introduce third-party algorithm evaluation and oversight. For platforms using algorithms to detect deepfakes, the effect of these algorithms can be difficult to assess; an "algorithm filing" and "algorithm impact assessment" system may be established, allowing independent agencies to periodically evaluate content review algorithms. Regulatory agencies may require large platforms to submit performance reports for their deepfake detection models, including recognition accuracy and false positive/negative rates, and conduct random tests. When necessary, platforms could even be required to provide partial access to algorithm code or models for regulatory review, analogous to the rigorous pre-market testing required for pharmaceuticals to verify safety and efficacy [8]. While there may be commercial confidentiality concerns, a reasonable degree of technical transparency should be regarded as part of fulfilling a social responsibility, especially in matters of public safety.

Furthermore, with the current rise of technologies such as blockchain, content traceability and watermark verification mechanisms can be explored. Deepfake generation services could be required to embed tamper-proof identification codes in every synthesized content item, with platforms automatically detecting these codes during dissemination; any unmarked content should be flagged or restricted. This technical approach provides objective verifiability of whether platforms have met labeling obligations, and also facilitates enforcement: if a platform disseminates unlabeled deepfakes, it is direct evidence of flaws in its review system. At present, Chinese public security agencies, in cooperation with research institutes, are intensifying technical efforts against AI face-swapping crimes, conducting security assessments of facial recognition and liveness detection on key platforms such as instant messaging, live streaming, and social media, to enhance the detection and evidence-gathering capabilities for deepfakes [9]. In this process, platforms are obliged to cooperate with technical audits, regularly undergo safety assessments, and rectify and report any problems discovered within prescribed timeframes. In summary, institutionalized log records, algorithm audits, and content traceability can visualize and substantiate platform duty performance, clarifying facts for subsequent accountability and exerting continuous compliance pressure in daily operations.

4.4. Differentiated approaches for corporate and executive liability: dual-layer accountability model

Platform dereliction often involves both the overall governance of the company and decisions by individual executives, making it necessary to distinguish between corporate liability and personal executive liability, and to implement a dual-layer accountability model. Article 286-1, paragraph 2 of the Criminal Law already stipulates that when a unit commits this crime, the unit shall be fined and the directly responsible supervisory and other directly responsible personnel shall be punished in accordance with paragraph 1. This provides a legal basis for dual-track accountability. In practice, further clarification is needed as to when focus should be placed on corporate versus personal liability to achieve both deterrence and prevention.

On the one hand, emphasis should be placed on corporate criminal liability to enhance the legal deterrent effect on platform companies. When platform negligence stems from profit-driven motives of corporate decision-makers or insufficient investment in compliance, only penalties imposed on the company can truly drive change. For example, in the "Qvod case," the company's business strategy tolerated piracy and pornography for profit, reflecting a company-wide disregard for legal obligations. For such profit-motivated and systematic dereliction, criminal prosecution should focus on the company as a unit, using high fines and business restrictions to target corporate pain points and compel reform. In tandem with China's ongoing compliance reforms, criminal penalties can be linked with compliance remediation: companies facing criminal liability should be required to implement corrective and compliance plans, with the possibility of reduced penalties upon satisfactory evaluation. This approach not only punishes illicit gains but also incentivizes lawful operations in the future.

On the other hand, holding individual executives personally liable is crucial to prevent the dilution of responsibility and to reinforce the deterrent effect on individuals. In cases of platform negligence, directly responsible managers such as heads of content moderation, chief security officers, and decision-makers such as CEOs often bear inescapable responsibility. Punishing only the company while neglecting personal accountability risks shifting responsibility, reducing the pressure on management and weakening incentives for compliance. Therefore, for managers with clear subjective fault and a high degree of decision-making relevance, criminal accountability should be strictly pursued. For example, if a platform executive is aware of major flaws in the company's deepfake review system but refuses to invest in improvements, or, after receiving reports of illegal content, instructs staff not to act, this willful indifference should trigger direct personal liability. Similarly, if evidence shows that executives discussed the traffic gains from deepfake content and gave instructions to allow it for profit, those involved should be regarded as joint offenders and held criminally responsible. By holding executives accountable, the principle that "criminal compliance starts at the top" is reinforced, encouraging management to prioritize internal governance and not to sacrifice legal boundaries for short-term gain. Of course, in practice, distinctions must be made regarding different roles and degrees of fault. For front-line reviewers unable to prevent problems due to objective difficulties, liability may be mitigated. For leaders with control over resources and decision-making who act passively, the law must hold them strictly accountable. Only when both companies and individuals bear the real costs of illegal dereliction will platforms develop a top-down compliance pressure network and fully implement measures to prevent deepfake-related offenses.

5. Conclusion

This article centers on the issue of platform criminal liability, responding to the real-world challenges that deepfake technology poses for legal governance, and seeks to demonstrate that platforms should bear legal responsibilities commensurate with their role. The particularity and unique responsibility of platforms do not stem from the technology itself, but from their deeply embedded intermediary position within the information dissemination chain. First, platforms control the channels of content distribution and the mechanisms of algorithmic recommendation, thereby deeply participating in both the creation and amplification of risks. Second, platforms must shoulder proactive compliance responsibilities, rather than passively awaiting administrative directives. Third, criminal law, as a measure of last resort, should intervene only when administrative and civil measures are insufficient to curb serious platform dereliction. Fourth, the criminal liability of platforms must be predicated on clear compliance standards and duties of care, so as to ensure the legitimacy and proportionality of criminal law intervention. Fifth, the effectiveness of criminal regulation depends on its synergy with administrative supervision and technological governance, to achieve comprehensive management of deepfake-related risks. In short, the construction of platform criminal liability is not intended to restrain technological development, but rather to ensure that the application of technology aligns with the needs of social security and order. Only by clarifying the legal obligations and boundaries of platform responsibility and encouraging platforms to proactively fulfill their duties can effective deepfake governance be achieved—so that technological innovation truly serves the welfare of society and the public interest.

References

[1] Fan, Y., & Yu, Y. (2022). Criminal Law Regulation of "Deepfake" Technology and Its Products in Online Communication. *Crime Studies*, (1), 51-60.

[2] Liu, Y. (2025). The Iterative Upgrade from Cybercrime to Digital Crime and the Criminal Law Response. *Comparative Law Research*, (1), 1-15.

[3] Feng, M., & Jiang, T. (2023). Criminal Law Regulation of the Abuse of "Deepfake" Technology. *Hubei Social Sciences*, (4), 127-135. https://doi.org/10.13660/j.cnki.42-1112/c.016084

[4] Chen, R. (2024). Criminal Law Regulation of Deepfake-Related Sexual Information. *Jurisprudence*, (3), 76-90.

[5] Zhou, J. (2024). Criminal Attribution of Online Platform Omissions: Rationale, Pathways, and Limits. *Journal of Shanghai Jiao Tong University (Philosophy and Social Science Edition), 32*(7), 98-117. https://doi.org/10.13806/j.cnki.issn1008-7095.2024.07.008

[6] Wang, L. (2025). Presumption of Fault of Network Service Providers in Information Network Crimes. *Journal of Henan University of Economics and Law, 40*(1), 103-116.

[7] Liu, W. (2024). Application Risks and Legal Regulation of AI Face-Swapping Technology. *Journal of University of Electronic Science and Technology of China (Social Sciences Edition), 26*(2), 60-69. https://doi.org/10.14071/j.1008-8105(2023)-4006

[8] Fang, H. (2024). The "Duality" of Digital Platform Governance and Mechanisms for Criminal Law Intervention. *Journal of East China University of Political Science and Law, 27*(5), 68-77.

[9] Gao, J., Sun, J., Cai, Y., Wang, C., Yang, Y., & Wang, K. (2025). Research on Artificial Intelligence Crime and China's Countermeasures. *Bulletin of the Chinese Academy of Sciences, 40*(3), 408-418. https://doi.org/10.16418/j.issn.1000-3045.20241025004