International Journal of Asian Social Science Research; Vol.3, No.1; 2026
E-ISSN: 3006-2179 P-ISSN：3006-754X
DOI: https://doi.org/10.70267/ijassr.260301.2031
Published by: Zeus Press

Criminal Regulatory Approaches to Deepfake-Related Offenses: Focusing on the Crime of Fraud

Songyang Sai [1*], **Zifan Wang**[2]

Law School (Intellectual Property Research Institute), Shandong Normal University, Jinan, 250358, China

**Corresponding author: Songyang Sai*

Abstract

As a representative form of AI-generated synthetic media, deepfake technology-characterized by its high degree of realism and increasing accessibility-has expanded the scope of legitimate technological applications while simultaneously providing novel tools and pathways for fraud-related crimes. Centered on the offense of fraud, this article systematically examines the role and functional mechanisms of deepfake technology within the criminal chain, identifying the associated criminal risks across key stages such as the illicit acquisition of personal information, content fabrication, and the execution of fraudulent schemes. The study finds that deepfake-enabled fraud exhibits distinctive features, including low technical barriers to entry, a modularized criminal chain, increasingly precise and multidimensional methods of deception, and the diffusion of harmful effects into the broader system of social trust. These characteristics pose significant challenges to existing criminal law frameworks, particularly with regard to the identification of criminal subjects, the classification of forms of accomplice liability, the allocation of platform responsibilities, and the legal characterization of relevant conduct. On this basis, the article proposes a systematic regulatory framework from three dimensions-legislative refinement, judicial response, and comprehensive governance. This includes promoting the interpretation and adaptation of the elements of fraud crimes, establishing an intelligent trial assistance system, implementing a dual-constraint mechanism for deepfake content labeling and informed consent, and the exploration of a collaborative “platform–government” regulatory model. These proposals aim to provide theoretical support and institutional reference points for the prevention and punishment of deepfake-enabled fraud.

Keywords

artificial intelligence, deepfakes, deepfake technology, AI-enabled fraud, crime of fraud

1. Introduction

Deepfakes refer to technologies and their generated outputs that employ artificial intelligence-particularly deep learning algorithms-to create or manipulate audio, visual, or textual content for specific purposes. Substantively, deepfakes constitute a form of AI-generated media, the technical core of which lies in self-learning deep neural network models capable of automatically producing highly realistic images, voices, videos, and other forms of information [1]. This technology integrates multiple artificial intelligence techniques, including video synthesis and computer vision.

The term “deepfake” originated in 2017 on the U.S.-based social news platform Reddit, where a user operating under the name “deepfakes” uploaded manipulated pornographic videos in which faces were replaced with those of public figures, drawing widespread public attention. Since then, deepfake technology has developed rapidly, with its application domains continually expanding. What initially emerged in the entertainment sector has gradually permeated diverse fields such as film and television production, medical imaging, and education, evolving into a powerful technological tool capable of reshaping information forms, transcending the boundaries of reality, and triggering far-reaching social transformations.

2. Research Foundations

2.1 Determining Illegality in the Context of Deepfake Technology

At present, deepfake technology demonstrates significant constructive value in a wide range of fields, including creative content production, film and television visual effects restoration, virtual influencers, and the digital reconstruction of cultural heritage. Applications of this kind generally fall within the scope of lawful and compliant technological innovation. In the audiovisual industry, for example, technological advancements driven by deepfake techniques have accelerated the development of new forms of productive capacity. Within China’s audiovisual sector, active exploration has taken place in areas such as text-to-video AI filmmaking and AI-generated digital humans, promoting human–machine collaborative creation and yielding a series of practical outcomes [2]. However, when the purpose of use deviates from its original creative or productive intent, technologies that initially empower industrial development may become instruments for infringing legally protected interests. In practice, certain malicious actors employ deepfake technology for purposes such as defamation, extortion, the incitement of social panic, or political manipulation, thereby infringing upon citizens’ lawful rights and interests, undermining corporate commercial reputations, and threatening public security. This functional transformation from technological innovation to technological abuse constitutes a critical juncture that legal regulation must precisely identify. It highlights the urgency of clearly delineating permissible boundaries of use in order to strike an appropriate balance between technological development and risk prevention.

From the perspective of industrial division of labor within the deepfake ecosystem, the chain of unlawful infringement may be analytically divided into three stages. The first stage concerns the illegal acquisition, sale, or unlawful provision of citizens’ personal information prior to the execution of deepfake activities. The second stage involves the unlawful use of deepfake technology and the improper production of deepfake content. The third stage consists of the subsequent utilization of such deepfake content to carry out a series of illegal or criminal acts, including fraud and defamation.

As of June 2025, China’s internet user population had reached 1.123 billion, with an internet penetration rate of 79.7% [3]. In the domain of personal information protection, the widespread adoption of the internet and the acceleration of digitalization have rendered issues of personal data leakage and illegal data transactions increasingly prominent. Data acquisition, as the initial link in the chain of unlawful conduct, constitutes the first stage of illegality determination. The relevant illegal acts typically manifest in two interrelated forms. The first involves the unauthorized and unlawful collection of citizens’ personal information, which directly violates the principle of legality in data collection. The second concerns the unlawful purchase, receipt, or solicitation of such information through transactional channels, giving rise to extended criminal chains involving the illegal sale and unlawful provision of personal data.

Generally speaking, the training of deepfake algorithms typically requires large volumes of authentic data in order to enhance the accuracy of model performance and the precision of generated content [4]. In practice, however, such data are often obtained without the lawful authorization of the data subjects. A substantial amount of personal information relating to target individuals constitutes the foundational material for the production of deepfake content. Accordingly, infringement of citizens’ personal information frequently represents the point of origin for the misuse of deepfake technology. Crimes targeting the security of citizens' personal information may be regulated under the offense of infringing upon citizens' personal information. [5].

Online platforms and mobile applications have become major channels for the collection of citizens’ personal information. Through these channels, actors may not only passively receive personal data voluntarily uploaded by users, but also actively collect various categories of user data. It is worth noting that, in practice, improper collection of personal information commonly manifests in two principal forms. First, users are often compelled to passively accept platforms' data collection practices. Privacy authorization mechanisms employed by platforms and applications are typically embedded in standard-form contracts, under which users must accept predefined terms in order to access core functionalities. As a result, users' freedom of choice is substantially constrained. Moreover, such standard-form contracts are frequently lengthy and complex. Ordinary users rarely read them in full, and non-professionals generally lack the capacity to fully comprehend all contractual provisions, leading many users to upload personal information without an adequate understanding of the associated risks. Second, platforms frequently exceed reasonable and necessary limits in the collection of personal information. Online platforms and mobile applications often engage in excessive data collection, requesting access to user information that goes beyond what is required for the maintenance of their core functions. For example, a shopping application that requests access to a user's location data or contact list-despite the absence of a direct functional necessity-may be deemed to engage in excessive data collection. By aggregating large volumes of citizens' personal information, platforms and applications are able to extract significant commercial value from such data. At the same time, certain illicit actors seek to unlawfully obtain facial images, voiceprints, and other sensitive identifiers of target individuals-by means of illegal sale or provision, theft, or other unlawful methods-for use in the production of deepfake content, thereby infringing upon citizens' personal rights and interests [6].

With respect to the determination of illegality at the second stage-the production of deepfake content-a contextual and differentiated analysis is required. Where deepfake technology is used merely to replace a specific individual's facial features in audiovisual materials for personal use or for entertainment within a limited circle, issues of intellectual property infringement are set aside for present purposes. Such conduct does not infringe upon the lawful interests of the State, society, or other citizens, nor does it cause actual harm to legally protected interests or create a concrete risk thereof. Accordingly, it lacks the necessity for criminal regulation.Similarly, where facial substitution involving others is limited to ordinary forms of "parody" or "pranks" of a general nature, such conduct should not, in principle, be classified as criminal. By contrast, where deepfake technology is employed to produce or disseminate face-swapped pornographic images or videos, the conduct escalates into criminally relevant acts, potentially constituting offenses such as insult, defamation, or the dissemination of obscene materials. In judicial practice, it is therefore essential to draw a clear distinction between improper but non-criminal content production and genuinely illicit conduct, so as to prevent the excessive expansion of criminal law intervention and to safeguard the healthy development and reasonable application of artificial intelligence technologies within the criminal law framework.

Deepfake technology is primarily used in content creation, but creating such content without dissemination rarely infringes upon legal interests. Therefore, it is the use of deepfake technology to produce content that truly transforms potential risks into tangible harm [7].

2.2 Defining Deepfake-Related Fraud Offenses

Owing to its high degree of realism and low-cost replicability, deepfake technology provides fraud with unprecedented technical tools and implementation scenarios. It is precisely the close alignment between the technical characteristics of deepfakes and the internal logic of fraud that renders fraud the primary pathway through which the latent risks of deepfake technology are transformed into concrete harm. Structurally, fraud is characterized by an intent to unlawfully appropriate property and the use of deception premised on information asymmetry-features that allow it to fully exploit the controllable false information generated by deepfake technologies [8].

From a doctrinal perspective, deepfake-related fraud offenses may be defined as acts whereby an offender, with the intent of unlawful appropriation, employs deceptive means and incorporates deepfake technology as an instrumental component of the criminal scheme, thereby fraudulently obtaining public or private property of a relatively substantial amount. More specifically, such offenses encompass conduct in which the offender, for the purpose of unlawful appropriation and by using deepfake technology as an integral element of the criminal scheme, fabricates facts and engages in fraudulent acts that induce the victim to fall into a misconception. Acting on this misconception, the victim disposes of property, enabling the offender to obtain the property and causing the victim to suffer a corresponding property loss.

Such offenses may also arise where the offender, with the intent of unlawful appropriation, is aware that the conduct constitutes deepfake-enabled fraud and bears a clear duty of disclosure toward the counterparty, yet deliberately conceals the truth, thereby maintaining the victim's mistaken belief. In reliance on this continued misconception, the victim disposes of property, the offender obtains the property, and the victim incurs financial loss.

3. Typologies of Contemporary Deepfake-Enabled Fraud

3.1 Face-Swapping Fraud

Face-swapping fraud refers to the use of deepfake technology to replace the fraudster's facial image with that of a target individual-such as a relative, superior, or public figure-thereby generating highly realistic video-call visuals in real time. This technology relies on generative adversarial networks to accurately map facial expressions, muscular textures, and lighting details, causing victims to develop a false sense of trust based on what appears to be "seeing with one's own eyes." Such fraud is often combined with social engineering techniques. Prior to the commission of the offense, perpetrators obtain information about the victim's social relationships through illicit channels and impersonate acquaintances to issue urgent requests for assistance or instructions for fund transfers, exploiting the strong psychological cues associated with visual trust to overcome victims' defenses.

3.2 Voice-Cloning Fraud

Voice-cloning fraud employs AI-generated technologies to collect, process, and manipulate samples of a victim's voice, producing audio segments that are virtually identical to the victim's original voice, often within a matter of minutes or even less. At present, this technology is relatively mature, features a low threshold for use, and demonstrates extremely high accuracy-reportedly reaching up to 99 percent-making it the most accurate among several generative modalities. Current artificial intelligence techniques are capable of replicating a person's intonation, speech rate, and accent with near-perfect fidelity. When combined with contextual simulation and signal interference, such forged voices become difficult for victims to distinguish from authentic ones. This form of fraud is particularly common in telephone scams, where perpetrators imitate the voices of familiar relatives or friends in order to induce victims into believing the deception.

3.3 Motion-Transfer Fraud

Motion-transfer fraud constitutes one of the principal forms of deepfake-enabled fraud. It relies on algorithmic techniques to capture and replace, in real time and with high precision, a specific individual's facial expressions, lip movements, eye gaze, and bodily actions. This type of fraud exhibits several defining characteristics. First, it features a high degree of technical concealment, as the generated video stream can be synchronized with audio and environmental contexts in real time, achieving an effect akin to "seamless grafting." Second, it demonstrates strong interactive deceptiveness, as dynamic details such as nodding, smiling, and specific gestures can be simulated during video interactions, creating the appearance of "authentic" communication. Third, it is highly target-specific, typically requiring the prior acquisition of static images or limited dynamic footage of the target individual as raw material in order to carry out precision fraud. In the course of implementation, perpetrators primarily employ this technology to accomplish identity impersonation. For example, in scams involving the impersonation of acquaintances or superiors, offenders may transfer their own real-time facial expressions and lip movements onto the facial images of the victim's relatives or supervisors and conduct video calls under this false identity. In such scenarios, victims are highly likely to accept the counterparty's identity as genuine and consequently lower their level of vigilance.

The consequences of motion-transfer fraud extend beyond substantial financial losses and often entail profound emotional betrayal and psychological harm, as the deception originates from individuals who appear to be trusted figures "seen with one's own eyes." At the societal level, this form of fraud significantly erodes the foundations of trust upon which social interactions depend-visual evidence no longer guarantees authenticity. Remote identity verification mechanisms thus face fundamental challenges, potentially giving rise to widespread identity security anxiety and markedly increasing the trust costs and risks associated with commercial transactions, remote work, and even familial communication. By algorithmically hijacking and real-time manipulating micro-expressions, eye movements, and bodily postures in video content-including facial expression transfer and pose transformation-perpetrators can precisely map their own actions onto fabricated targets. This technology is capable of animating static facial images, enabling fraudsters to create the illusion of "live, real-person interaction" during video calls, thereby greatly enhancing the immersive quality and perceived credibility of the deception and making it difficult for victims to assess authenticity based on behavioral logic.

3.4 Fabricated-Scenario Fraud

Fabricated-scenario fraud employs background transfer and environmental synthesis techniques to seamlessly embed individuals into artificially constructed spatiotemporal settings, such as virtual studios, conference rooms, or press briefing venues, thereby creating false contexts imbued with authority and credibility. Typical applications include forged celebrity endorsement videos, fabricated expert recommendation footage, and falsified scenes of official announcements. By relying on "scenario endorsement" effects, this form of fraud enhances the perceived credibility of information, reduces victims' critical thinking under contextual cues, and-when combined with fabricated identities and persuasive narratives-enables the multifaceted hijacking of brand reputation, consumer trust, and public cognition, with social harm spreading in a chain-like manner.

The core mechanism of this type of fraud lies in improving the efficiency of deception through the provision of highly realistic scenarios. By constructing virtual environments that appear authoritative, formal, or urgent, perpetrators activate audiences' trust expectations typically associated with such settings. When coupled with fabricated identities and inducive narratives, this process completes a chain from contextual misdirection to behavioral manipulation. The resulting social harm exhibits a cascading pattern: it not only directly infringes upon individual property interests, but also undermines public trust in media, expert systems, and official channels, and in the long term may weaken society's shared consensus and judgment concerning authenticity.

4. Characteristics of Deepfake-Enabled Fraud

Deepfake-enabled fraud exhibits three core characteristics: a low technical threshold combined with highly realistic forgery effects; a modularized criminal chain that leads to difficulties in tracing and attributing liability; and harmful consequences that extend from individual property losses to the erosion of the broader social trust system. Specifically, the accessibility and deceptiveness of the technology lower barriers to criminal participation and challenge traditional cognitive boundaries; the division of labor and cooperative structure within the criminal chain render perpetrators increasingly concealed and tracing accountability more difficult; and the precise and multidimensional methods of offending not only cause direct financial losses but also undermine society's shared consensus on authenticity, giving rise to systemic crises of trust.

4.1 Low Technical Threshold and Highly Realistic Forgery Effects

The open-source nature and rapid diffusion of deepfake technology have significantly reduced the threshold for its use, enabling ordinary fraudsters to generate forged content that is difficult to distinguish with the naked eye even without advanced technical expertise. The striking realism of the generated outputs directly destabilizes the long-standing cognitive foundation of "seeing is believing," thereby subjecting the credibility of traditional forms of evidentiary media-such as audio and video-to fundamental doubt. When technological accessibility converges with extreme realism, it effectively constructs an "infrastructure" for the proliferation of deepfake fraud, endowing such crimes with characteristics of low cost, high efficiency, and strong concealment. As demonstrated by joint research conducted by German and Italian scholars, untrained individuals are almost incapable of reliably distinguishing the authenticity of face-swapped videos. This further underscores the systemic impact of deepfake technology on public cognitive defenses and the structure of social trust.

4.2 The Composite Nature of Criminal Objects Involving Multiple Legally Protected Interests

Deepfake-enabled fraud departs from the traditional model of fraud that targets only property ownership as a single criminal object and instead exhibits a composite pattern involving the cumulative infringement of multiple legally protected interests. By fabricating the identities and images of specific subjects-such as public officials, relatives, or celebrities-it simultaneously infringes upon citizens' personal information rights, portrait rights, and reputation rights, among other personality interests. Moreover, through the large-scale dissemination of fabricated crisis-related information, such conduct directly threatens public order in cyberspace and may even endanger national security. This expansion of the criminal object-from "individual property rights" to a combination of personality rights, property rights, and social administrative order-causes a single criminal act to implicate multiple categories of interests protected under criminal law. As a result, issues of statutory concurrence and offense classification become increasingly complex, significantly raising the difficulty of judicial determination.

4.3 Modularized Criminal Chains and Difficulties in Tracing and Attribution

In deepfake-enabled fraud, the criminal industry chain is tightly interconnected, rendering tracing and accountability particularly difficult. Fraudsters typically operate through a division-of-labor model, implementing criminal activities across multiple stages, including technical support, equipment management, algorithmic prediction, audience targeting, and the execution of fraudulent acts. The resulting industrialized fraud chain poses substantial challenges to fraud governance. In many cases, even when criminal elements at a particular stage are identified during investigation, it remains difficult to trace responsibility back to the actual organizers or planners. Furthermore, the quality of biometric data-illegally obtained or provided by personal information traffickers-directly determines the realism of forgeries and the success rate of fraud. Such acts of "data poisoning" should be regarded as co-perpetration in deepfake-enabled fraud rather than as independent offenses. In addition, network service providers may incur corresponding liability for omission where they fail to fulfill their information network security management obligations.

4.4 Deep Integration of Precision Inducement and Multidimensional Offending

Criminal methods have undergone a shift from indiscriminate "wide-net casting" to precise, individualized cognitive attacks. By using AI tools to scrape and analyze victims' social media data, perpetrators can accurately identify emotional vulnerabilities and economic conditions, enabling the customized design of fraudulent schemes and significantly increasing success rates. At the same time, this approach breaks away from traditional linear fraud models and constructs a multidimensional "point–line–network" structure: initiating deception through a single forged element, remotely directing operations through fabricated identities, and forming a networked diffusion pattern based on multi-level division of labor and cross-platform dissemination. This non-contact, distributed, and intelligent multidimensional architecture causes the social harm of such crimes to expand geometrically and substantially increases the difficulty of governance.

4.5 Extension of Harm from Individual Property Loss to the Collapse of the Social Trust System

Deepfake-enabled fraud not only results in direct financial losses, but also undermines habitual modes of thinking based on "seeing is believing" and "hearing is believing," thereby breaching individual psychological defenses and triggering a broader crisis of social trust. Once deepfake content is disseminated on a large scale in the absence of reliable identification technologies, the public may develop systemic skepticism toward all forms of audiovisual information, accelerating the transition into a so-called "post-truth era" [9]. This foundational erosion of consensus regarding authenticity far exceeds the destructive scope of traditional fraud and poses fundamental challenges to core social mechanisms, including judicial fact-finding, news dissemination, and interpersonal interaction.

5. Issues in Deepfake-Enabled Fraud Crimes

5.1 The Issue of Criminally Liable Subjects

Whether generative artificial intelligence can constitute a subject of criminal liability has become a highly debated issue in contemporary criminal law scholarship, both in China and abroad. In the context of deepfake technology, the question of whether criminal responsibility may be attributed to generative AI likewise warrants examination. Under the current criminal law framework, criminal liability is limited to natural persons and legal entities, and generative artificial intelligence cannot qualify as a criminal subject. However, according to the doctrine of indirect perpetration, where an offender exploits a person lacking criminal capacity or criminal intent to carry out a crime in order to achieve their own criminal purpose, the principal nature of the offense lies in the user's control over the causal process of the crime through means such as coercion or deception, treating the exploited party as a tool of the offense. In the specific context of fraud committed through the use of deepfake technology, the so-called "constitutive object of the crime" generally refers to the material elements indispensable to the commission of the offense and forms part of the objective aspect of the crime. As the foregoing analysis indicates, deepfake-type fraud cannot be constituted without the use of deepfake technology. Accordingly, this article argues that even if deepfake technology is characterized not as a criminal tool but as a constitutive object of the offense, the individual who manipulates and controls it from behind may still be held liable as an indirect principal offender.

5.2 Issues in the Determination of Criminal Conduct

In deepfake-enabled fraud crimes, traditional forms of joint perpetration at the execution stage have decreased, while cooperative conduct at the preparatory stage has increased significantly. This structural shift primarily stems from the extensive scope of interests infringed by such crimes. Although deepfake-enabled fraud is characterized by a high degree of technical realism, its actual conversion rate is constrained by multiple factors, as not all individuals who come into contact with fraudulent content are deceived. To obtain a certain amount of criminal proceeds, offenders must therefore expand the pool of potential victims and cover as many targets as possible, thereby ensuring the success of fraud on a probabilistic basis. Moreover, the realization of such extensive infringement relies heavily on technological and large-scale preparatory conduct. This mainly includes the mass acquisition of data and materials, such as the collection of facial images, voiceprints, and social information of specific groups. These activities are often carried out through database intrusions or the scraping of publicly available information. In addition, the batch generation and dissemination of forged content require the use of automated tools to simultaneously produce personalized fraudulent materials targeting different individuals and to distribute them on a large scale through social networks, messaging groups, and similar channels. Given the substantial workload and diverse technical requirements involved, such tasks are difficult for a single individual to complete independently, thereby giving rise to a division of labor and cooperation at the preparatory stage centering on the acquisition of "raw materials," technical tools, and promotional channels.

As a result, deepfake-enabled fraud has come to exhibit a new pattern characterized by "cooperation at the preparatory stage and individualization at the execution stage." Traditional "joint execution" in accomplice liability has diminished, while upstream activities-such as material collection, model training, script drafting, and dissemination-are more likely to be organized and coordinated in the form of criminal groups.

For example, where A provides B with deepfake technology and instructs B on how to use it to commit fraud, including the transmission of criminal techniques, A may possess only abstract awareness regarding how B selects victims, the specific methods employed, and the extent of the resulting property losses. If B independently masters the relevant deepfake technology and carries out large-scale property crimes, causing substantial harm to numerous victims, the question arises as to whether A should bear criminal liability for all offenses as an instigator or an aider. Furthermore, if B upgrades the deepfake technology or updates the data during its use-transforming a model originally intended for fraud into one used for extortion, or significantly enhancing the conversational scripts to increase deceptiveness and harmfulness-this raises the issue of whether such conduct constitutes an excess in execution. For instance, where A connects a chat-based generative AI interface to the WeChat platform, preconfigures it to simulate customer service for loan-related business, and requires users to transfer RMB 5,000 to a designated account, but the AI system "exceeds" the assigned task during actual operation and fraudulently obtains as much as RMB 50,000, this article argues that A possesses an abstract and general intent regarding the amount of fraud (indirect intent). If A in fact receives the entire RMB 50,000 in illicit proceeds, indirect intent may be established for the full amount. However, if B, without A's knowledge, improves the deepfake technology and obtains the excess proceeds, the portion exceeding RMB 5,000 should be regarded, with respect to A, as an excess in execution, and A should be held criminally liable as a joint offender only within the amount of RMB 5,000.

5.3 Issues Concerning Platform Regulatory Responsibility

Progress in collaborative governance between platforms and government authorities remains slow. Significant disparities exist among platforms in terms of the completeness and granularity of their internal regulatory frameworks. Following the implementation of the *Measures for the Labeling of Artificial Intelligence–Generated and Synthetic Content*, some platforms have still failed to update their user agreements in a timely manner, reflecting a lag in responding to national governance requirements.

Platforms further exploit their structural advantages to evade liability for harms arising from the misuse of technology. Through standard-form contractual clauses, platforms explicitly exclude their responsibility for risks associated with relevant content. User agreements across platforms commonly assign all losses caused by deepfake content exclusively to content creators and disseminators. For example, Article 14 (“Disclaimer”) of the Douyin User Agreement expressly states that the platform provides no guarantees with respect to risks encountered during user activities and strictly limits its own compensation liability. Such “liability exemption” clauses weaken the institutional protection of user rights and are inconsistent with the social responsibilities that platforms, as governors of content ecosystems, ought to bear.

Platforms have also failed to adequately fulfill their duty of risk disclosure. Regulatory standards for the processing of sensitive data, such as users’ facial information, remain insufficient, resulting in inadequate protection of individual rights. At present, video recording and live streaming have become standard features on short-video platforms, and the activation of embedded AI effects typically triggers the platform’s access to users’ facial data and its upload to servers. Research indicates that only Bilibili (B站) explicitly and comprehensively discloses in its privacy policy that it accesses users’ sensitive personal information. Other platforms lack such disclosures in relation to users’ engagement with AI effects. This absence of adequate notification makes it difficult for users to clearly perceive the boundaries of their data rights and further impedes their ability to guard against potential harms arising from data misuse.

5.4 Issues Concerning the Standards for Identifying Deepfake-Enabled Fraud

At present, the number of adjudicated cases available for research on deepfake-enabled fraud remains limited, while media reports on such fraud are frequent, indicating a high incidence of related crimes. With respect to whether conduct involving deepfakes constitutes fraud, the principal difficulty in current legal determination lies in the following question: If A commits acts such as infringing upon citizens' personal information or improperly creating deepfake works prior to using synthetic works to commit fraud, should the multiple offenses be treated as a single, more serious crime under the principle of selecting the most severe punishment, or should they be punished cumulatively? The academic community has yet to reach a consensus.

6. Criminal Law Regulatory Pathways for Deepfake-Enabled Fraud

6.1 Building a Comprehensive Legal and Regulatory Framework to Achieve Ex Ante Legal Safeguards

In response to the technological characteristics of deepfake-enabled fraud, it is necessary to promote a moderate expansion in the interpretative application of the existing statutory provisions on fraud [5] Through judicial interpretations, the use of “highly deceptive AI-generated audio-visual materials, images, and similar synthetic content” may be expressly incorporated into the forms of “fabricating facts or concealing the truth,” thereby achieving substantive congruence with the constituent elements of fraud under Article 266 of the Criminal Law. At the same time, face-swapping conduct undertaken for different purposes-such as extortion, reputational harm, or illicit financial gain-should be distinguished and addressed under the corresponding offenses, including defamation, insult, and extortion, so as to ensure precise and differentiated criminal punishment. Furthermore, in light of the “non-human” characteristics of AI-based deepfake conduct, it is necessary to explore the introduction of corporate criminal liability or a framework of accomplice liability for entities in platform-based participation (such as platforms providing automated content synthesis services). Such reforms would help fill the existing accountability gaps arising from claims of platform neutrality in criminal law. Meanwhile, for abusive uses of deepfake technology that do not reach the threshold of criminalization, appropriate coordination should be ensured through civil law, administrative law, and other relevant branches of law to provide a coherent and layered regulatory response.

6.2 Constructing Intelligent Adjudication Systems and Exploring Governance Pathways for Fraud Offenses

The introduction of artificial intelligence–assisted mechanisms into the field of judicial adjudication holds significant value. Against the backdrop of rapid technological evolution, the integration of generative artificial intelligence is of critical importance for advancing the intelligent transformation of judicial proceedings. At present, judicial systems around the world are actively exploring AI system architectures compatible with their respective institutional frameworks. The application of generative AI technologies is expected to bring about profound changes in the standardization of adjudicative criteria, thereby effectively promoting judicial fairness, enhancing the efficiency of criminal punishment, reducing the likelihood of wrongful convictions, and exerting a preventive deterrent effect on the formation of criminal intent in fraud-related offenses. In China’s judicial practice, preliminary attempts have already emerged in which local courts employ algorithmic technologies to assist adjudication, some focusing on the intelligent handling of specific categories of cases, while others utilize element-based analytical models to promote the standardization of criminal case processing. Looking ahead, with further technological maturation and the improvement of supporting institutional arrangements, the application of generative AI within judicial processes is expected to expand gradually.

To further optimize the functional deployment of artificial intelligence within the judicial system, systematic development should be advanced along the following dimensions. First, efforts should be made to promote the systematic integration and high-quality accumulation of judicial data resources, with the aim of constructing a comprehensive database of judicial activities. The development of generative AI relies fundamentally on learning and training based on large-scale, structured data. Only by systematically aggregating various types of information on illegal and criminal conduct, judicial decisions, and outcomes of legal application into a unified platform can the accuracy and adaptability of intelligent analysis in fraud cases be continuously enhanced. Second, coordination mechanisms among judicial authorities should be strengthened by establishing cross-departmental information-sharing and operational linkage systems. In responding to emerging forms of fraud, while adjudicative authority resides with the courts, investigative and prosecutorial stages involve public security organs and procuratorial authorities, respectively. It is therefore necessary to build an AI-enabled collaborative platform that connects all case-handling bodies, optimizes the allocation of judicial resources, integrates multi-source information generated throughout the entire case-handling process, and breaks down data silos between institutions. Third, the level of awareness and practical competence of judicial personnel in relation to artificial intelligence technologies should be enhanced. In the face of rapid technological advancement and the continuous evolution of criminal forms, judicial practitioners must maintain an ongoing update of their knowledge structures. This is both an intrinsic requirement for professional development and a practical necessity for addressing new types of criminal challenges. Without a basic understanding of and operational capacity in generative AI technologies, case-handling personnel may be unable to fully leverage the auxiliary functions of technological tools, thereby affecting judicial quality and efficiency. Accordingly, the judicial workforce should proactively adapt to digital transformation and strengthen its mastery of emerging technologies and specialized knowledge, so as to enhance its capacity to perform judicial functions in an environment of technological integration.

6.3 Establishing Consent as the Governing Principle with Explicit Exception Clauses

Where deep synthesis technologies involve natural persons’ biometric identifiers, such as facial images or voiceprints, an authorization mechanism centered on *informed consent* must be firmly established. That is, the collection and use of biometric data shall be permitted only upon obtaining the explicit consent of the data subject, and such consent must be given voluntarily, free from any form of coercion or disguised compulsion. At the same time, even where consent has been obtained, certain categories of conduct should remain prohibited. Research and development activities involving deep synthesis technologies must comply with ethical standards; in particular, the development of high-risk products or services should be subject to prior ethical review procedures. Entities engaged in the production and dissemination of deep synthesis content are strictly prohibited from employing such technologies for unlawful purposes, including political manipulation, dissemination of obscene or pornographic content, and identity impersonation. Apart from the aforementioned absolute prohibitions, producers and disseminators of deep synthesis content must refrain from misleading the public in their use of such content, a requirement that corresponds to the obligation of conspicuous labeling discussed above. For example, where an artificial intelligence entity interacts with the public for commercial or political purposes, it must proactively disclose its AI-generated nature. In addition, online platforms-particularly social media platforms-should bear heightened supervisory responsibilities, including the establishment of mechanisms for identifying false information, strengthening content moderation, verifying whether user-generated content has been duly labeled as deep synthesis content, and fulfilling their duties to identify and block unlabeled content that may endanger national security or the public interest.

6.4 Promoting a Joint "Platform–Government" Regulatory Model

The abuse of digital deepfake technologies infringes upon multiple legal interests and therefore requires coordinated regulation involving multiple governmental departments and industries at the national level. In order to effectively address the regulatory complexity and highly variable application scenarios of deepfake-related misinformation, a "hub-and-spoke" regulatory model may be adopted, under which the respective supervisory responsibilities of governmental authorities are clearly defined and regulatory boundaries are rationally allocated. Under this model, the Ministry of State Security functions as the central hub, with key participating bodies including the Central Propaganda Department, the Cyberspace Administration of China, the Ministry of Public Security, and the National Radio and Television Administration. Regulatory authorities should clearly delineate the legal "red lines" governing the use of deepfake technologies by online users. At the same time, China may consider incorporating specific provisions on the regulation of deepfake misinformation into the *Data Security Law*, so as to enhance regulatory specificity and inter-agency coordination.

China may further clarify the periodic reporting obligations of digital platforms, thereby correcting existing asymmetries of information and authority between platform self-regulation and governmental supervision, while preserving necessary regulatory flexibility. Products and services based on deep synthesis technologies are characterized by rapid dissemination and wide social impact. Once misinformation arises, both the scope and intensity of its effects can be substantial. From the platform perspective, effective remedial mechanisms must therefore be in place. Where social misunderstanding occurs, platforms should provide convenient and efficient means for clarification and debunking. Such debunking responsibilities should primarily rest with content publishers or the relevant platforms; additionally, consideration may be given to enabling social organizations or government-established online platforms to undertake clarification and explanatory functions, so as to minimize the adverse social consequences caused by misinformation.

Platforms and governmental authorities may also establish a tiered and categorized regulatory framework. Given the rapid development of deep synthesis technologies and the complexity of their application scenarios, it is inappropriate to adopt a uniform, undifferentiated regulatory approach. Instead, deep synthesis technologies, products, and services should be subject to graded and classified management. Within the industry, trade associations and enterprises may participate deeply in governance efforts, assisting regulatory authorities in jointly formulating classification standards for technologies and products, thereby enabling differentiated regulation.

A dual content-labeling system should be constructed as an integral component of platform regulation. Existing norms require deep synthesis service providers to apply specific labels to the information content they generate. Such labels may be divided into two primary forms. The first is traceability labeling, which aims to ensure that synthesized content distributed or disseminated online possesses identifiable sources and traceable dissemination paths, thereby facilitating content tracking by service providers and enabling the identification of the original synthesis entity in cases of dispute or source verification. The second is conspicuous disclosure labeling, which requires that the public be clearly and perceptibly informed that the content has been generated through deep synthesis technologies, thereby alerting audiences to potential issues of authenticity and accuracy and preventing associated legal and ethical risks. Specifically, for image-based content, explicit disclosure may follow practices analogous to copyright attribution; for audio content, clear verbal statements should be provided at the beginning and end of the audio track (for example: "This audio has been generated using deep synthesis technology and is not an original recording").

7. Conclusion

The iteration and application of deepfake technologies have transcended their role as mere tools, becoming a significant variable in reshaping social trust structures and legal governance paradigms. Focusing on fraud as the regulatory lens, this paper systematically elucidates the technical enabling mechanisms and criminal risk spectrum of deepfake technologies within the criminal chain, revealing their characteristic features: "low threshold-high realism-precision-modularity-multidimensionality-broad societal diffusion," as well as the extension of consequences from individual property harm to erosion of the social trust system. Faced with multiple challenges-including ambiguous identification of criminal subjects, complex forms of joint offenses, imbalanced allocation of platform responsibilities, and difficulties in characterizing conduct-reliance on criminal law alone is insufficient, necessitating a shift toward a coordinated "legislation–judiciary–platform–society" governance framework.

At the legislative level, it is essential to promote interpretative adjustments to the elements of fraud, clarifying the correspondence between technological misuse and the infringement of legal interests. At the judicial level, intelligent adjudication support systems should be established to enhance the efficiency of case recognition and the consistency of rulings. In terms of platform governance, dual constraints of content labeling and informed consent should be reinforced, establishing the principle of "consent as default, prohibition as exception." Within regulatory mechanisms, a "platform–government" hub-and-spoke coordination model should be explored to achieve a dynamic balance between technological development, risk prevention, and rights protection. These pathways provide not only an interpretative framework and institutional reference for the criminal law response to deepfake-enabled fraud but also offer cross-domain perspectives for the theoretical construction and practical innovation of new crime governance in the AI era.

Looking forward, as deep synthesis technologies evolve toward immersive, real-time, and personalized applications, the risks of misuse are expected to become more covert and diffuse. Criminal regulation must remain open and forward-looking, continuously monitoring technological evolution and changes in criminal modalities, in order to seek a dynamic equilibrium between safeguarding innovation and preventing social risks. Only through multidisciplinary dialogue, inter-agency collaboration, and globalized governance can a digital social rule system be established in which technology serves the public good, legal principles are clear, and rights and responsibilities are well defined, thereby fortifying the rule-of-law defenses in the post-truth era.

References

[1] Bao, Y. X., Lu, T. L. and Du, Y. H. Overview of deepfake video detection technology. Computer Science. 2020, 47(09), pp. 283-292. https://doi.org/10.11896/jsjkx.200400130.

[2] Zou, Y., He, Y. X. and Zhou, A. C. Generative AI collaborative creation: An innovative path for developing new productivity in the audiovisual industry. China Television. 2024(6), pp. 91-98. https://doi.org/10.3969/j.issn.1002-4751.2024.06.019.

[3] website, C. I. N. I. C. C. o. The 56th Statistical Report on Internet Development in China. Media Forum. 2025(15), p. 121. https://doi.org/10.3969/j.issn.2096-5079.2025.15.038.

[4] Li, H. On the criminal responsibility about the abuse of personal biometric information: Taking artificial intelligence "Deepfake" as an example Tribune of Political Science and Law. 2020, 38(04), pp. 144-154.

[5] Li, M. The criminal regulation approach to deep fake technology abuse. Law-Based Society. 2021(6), pp. 40-47,73. https://doi.org/10.19685/j.cnki.cn11-2922/n.2021.06.005.

[6] Ye, X. B. Personal information protection in the context of generative artificial intelligence: Paradigm shift and rule improvement. The Jurist. 2025(04), pp. 61-73+192. https://doi.org/10.16094/j.cnki.1005-0221.2025.04.004.

[7] Zheng, G. J. Regulatory logic and governance pathways: Criminal law responses to the abuse of deepfake technology. Science of Law(Journal of Northwest University of Political Science and Law). 2025, 43(3), pp. 56-68. https://doi.org/10.16290/j.cnki.1674-5205.2025.03.009.

[8] Yao, Z. W. and Li, Z. L. Legal regulations governing AI fraud. Nomocracy Forum. 2023(04), pp. 301-308.

[9] Zhang, T. Legal risks and regulation of deepfakes in the post-truth era. E-Government. 2020(4), pp. 91-101. https://doi.org/10.16582/j.cnki.dzzw.2020.04.009.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output ofthe'science project.

Copyrights