Typological Determination of the Negligence Liability of Generative Artificial Intelligence Providers

Xinyue Song *

School of Humanities and Law, Northeastern University, Shenyang 110169, Liaoning

* Corresponding Author Email: 20240580@stu.neu.edu.cn

Abstract. The iterative evolution of generative artificial intelligence has given rise to novel criminal risks. For a long time, academic debate has remained locked in the binary controversy over whether strong artificial intelligence qualifies as a subject of criminal liability, while few studies have turned to the practical question of how liability should be attributed to providers. Attribution thinking remains trapped within the conventional paradigm of subject-status determination. Provider liability is often presumed to be liability for negligent crimes, with accountability pursued by stretching the causal chain, yet the preconditions for establishing negligence liability have not been adequately examined, the boundaries of attribution remain vague, and no systematic determination framework has been constructed. Moving beyond the controversy over subject status, this paper divides scenarios according to differences in the logic of harm production, and the paths of attribution diverge accordingly. In traditional negligence scenarios, harm mostly stems from human carelessness at the source stage, and autonomous algorithmic decision-making does not substantially intervene in the causal process; the traditional logic of negligent crimes can therefore be applied by relying on foreseeability, the duty of result avoidance, and adequate causation. In scenarios of independent harm infliction by artificial intelligence, by contrast, autonomous algorithmic output breaks the causal chain, the theory of supervisory negligence faces dogmatic difficulties, and the focus of attribution shifts to the post-hoc removal duty, so that legal interests are protected through the path of duty offenses. Between the protection of legal interests and technological innovation, the criminal law should maintain modesty and treat administrative regulation as the preferred path of governance.

Keywords: Generative artificial intelligence; providers; negligent crimes; typological determination.

1. The Origin of the Problem

The rapid development of generative artificial intelligence empowers society even as it brings profound criminal risks. President Xi Jinping has pointed out that today's global artificial intelligence technological innovation 'holds both great opportunities and governance challenges [1].' Facing the question of the era—when algorithms participate in decision-making, how can safety be ensured—how to let technology 'run fast' while ensuring that it 'runs steadily' has become a practical subject that legal research must answer.

In practice, typical cases such as the Uber autonomous-driving fatality [2] and the Tesla Autopilot fatal crash [3] both show that when a harmful result stems from the 'autonomous decision-making' of artificial intelligence, the existing attribution system faces structural difficulties; but if the harm can be traced to source defects in algorithm design, training data, or security protection, it still falls within the traditional negligence scenario typified by technical defects. As to whether strong artificial intelligence qualifies as a subject of criminal liability, theory is divided between the affirmative [4] and negative 5] views; yet in practice, directly attributing liability to artificial intelligence faces practical obstacles such as the application of penalties. This paper therefore does not dwell on the controversy over subject status but shifts its perspective to the criminal liability of natural persons such as providers.

Viewed more closely, the above controversy reflects a deep tension between the ideal and the actual. The affirmative view can hardly answer the practical difficulty that traditional penalties such as fines and custodial punishment cannot be applied to AI systems; the negative view, while avoiding that difficulty, may leave a regulatory vacuum. In the Uber case, the operator was ultimately not held criminally liable, and such responsibility absence neither satisfies the public's expectation of justice

nor effectively curbs the spread of risk. Breaking this deadlock requires re-examining the responsibility of human subjects such as providers, together with its normative nature.

However, when existing studies examine the criminal liability of providers, they tend to conclude directly that providers should bear liability for negligent crimes, or to trace attribution backwards by lengthening the causal chain, leaving several preliminary questions insufficiently answered. First, should providers bear negligence liability at all? Existing research concentrates on the constitutive conditions and determination paths of negligent crimes, while lacking adequate discussion of the prior question whether criminal liability should be imposed in the first place. Second, under what circumstances should providers bear negligence liability? Whether the logic of attribution differs substantially across different types of harm scenarios is a question that existing studies, usually analyzing only a single situation, have failed to answer with a systematic typological framework. Third, if negligence liability does not apply, might providers still bear other forms of criminal liability? Whether post-hoc duties of providers, such as the duty to remove content, may trigger the application of duty offenses remains controversial in scholarship.

In view of this, this paper takes the typological determination of providers' liability for negligent crimes as its object of study. Based on differences in the logic of harm production, it typifies the relevant situations into the traditional negligence scenario typified by technical defects and the independent harm infliction scenario typified by network platforms, examining the forms of providers' responsibility in each.

2. The Scenarios of Criminal Risk Involving Artificial Intelligence

The premise of typological determination lies in accurately identifying the normative features of different harm scenarios. Harm caused by generative artificial intelligence takes no single form: one type stems from source-stage carelessness of providers in research, development, and deployment, with a clear and traceable causal chain; the other stems from autonomous output of AI systems after they escape human intervention, so that the causal chain becomes complex through the intervention of autonomous decision-making.

2.1. Traditional Negligence Attribution Scenarios Exemplified by Technical Defects

AI-inflicted harm caused by technical defects refers to situations in which an AI system, by reason of the provider's technical carelessness in research, development, and deployment, infringes legal interests after being put into use. Although the harmful result is formally 'produced' by artificial intelligence, its true root lies in the technical defects introduced by the provider. As the creator of the source of risk, the provider has de facto control over the AI system.

Technical defects mainly manifest in three forms. First, algorithmic defects. The quality of algorithm design directly determines the decision logic and behavioral boundaries of artificial intelligence. In March 2018, an Uber self-driving vehicle struck and killed a pedestrian in Tempe, Arizona, becoming the first accident worldwide in which an autonomous vehicle caused a human death. In that case, the autonomous driving system failed to make a timely braking decision after detecting the pedestrian, exposing the provider's insufficient reliability review at the algorithm design stage. In the field of medical AI, if a provider ignores lesion characteristics of particular ethnic groups when annotating CT images, the system's misdiagnosis rate rises markedly. The core of algorithmic defects lies in the provider's lax review of core performance indicators such as algorithmic reliability and correctness at the design stage, allowing defective AI systems to enter the market. Second, training data bias. Take the American COMPAS recidivism risk assessment system: trained on historical crime data, it repeatedly output higher risk scores for Black defendants than for white defendants in actual use [6]. The particularity of training data bias is that it is usually not 'intentional' on the provider's part; rather, it constitutes negligence because the provider fails to ensure the purity, representativeness, and fairness of training data during collection and annotation. Third, security vulnerabilities. In September 2024, an algorithm engineer surnamed Wang in the AI short-drama

department of a technology company entered the notorious ‘delete the database and flee’ code, causing 89TB of core training data to be deleted and all R&D projects of the company’s AI gaming department to grind to a halt [7]. Although that case involved an insider, it fully exposed that if a provider fails to establish sound mechanisms for permission management, data backup, and security monitoring, the core assets of an AI system face the risk of destruction.

Taken together, this scenario has the following features. The root of the harm lies in the provider’s technical carelessness, manifested in poor algorithm design, lax data quality control, and an inadequate security protection system; the technical defects are under the provider’s direct control during the R&D stage, and the provider possesses the corresponding capacity to recognize and foresee the harmful consequences they may trigger; and there is no substantial intervention of an intermediate variable in the form of autonomous algorithmic decision-making, so the causal connection between the provider’s conduct and the harmful result is relatively clear. These features jointly show that this scenario is a typical one to which the traditional theory of negligent crimes can be properly applied.

2.2. Scenarios of Independent Harm Infliction by Artificial Intelligence Exemplified by Network Platforms

The dissemination of unlawful content generated by artificial intelligence on network platforms has become one of the risk scenarios peculiar to generative AI. Unlike the model of traditional internet platforms, where users actively upload unlawful information, generative AI can produce unlawful information such as fake news, obscene pornographic content, and infringing materials while operating autonomously or following user instructions and can achieve large-scale dissemination through network platforms. Viewed by generation path, there are ‘human-induced’ scenarios, in which users induce the AI to generate unlawful content through prompts, and ‘spontaneous’ scenarios, in which the AI produces unlawful content on its own, free from direct intervention; the core feature of the latter is the uncontrollability and randomness of the generation process, which may yield entirely different results even under identical input conditions.

From the ‘spontaneous’ generation of unlawful content in network platform scenarios, one can abstract a more general scenario of independent harm infliction by artificial intelligence: the provider has not actively intervened in the content generation mechanism, yet the AI system produces harmful results on its own, detached from direct human intervention. Whether the provider bears a prior duty to prevent harmful results then becomes the preliminary question for liability determination. However, most crimes of omission under traditional criminal law depend on statutory duties or contractual duties; whether AI model providers have a positive duty to prevent harmful outputs has no clear legislative support. In ‘spontaneous’ scenarios, the provider’s role is closer to a neutral position, lacking substantial prior control over the system’s outputs, and the generation of unlawful content can hardly be fully predicted or blocked beforehand.

It is worth noting an important difference between network platform scenarios and the general scenario of independent harm infliction: providers of AI services in network platforms also bear a duty to take timely disposal measures such as deletion and blocking when improper information spreads on a large scale, as expressly stipulated in Article 14 of the Interim Measures for the Management of Generative Artificial Intelligence Services [8]. Based on this removal duty, providers may face the application of the crime of refusing to fulfill information network security management obligations. It should be emphasized that this removal duty arises from the particularity of the network platform scenario; the general scenario of independent harm infliction does not involve such post-hoc content management duties.

In sum, the core feature of this scenario is that the harmful result does not stem from the provider’s technical carelessness but is produced by the AI system on its own, detached from direct human intervention. This feature forms a sharp contrast with the technical-defect scenario described above: the latter centers on the provider’s source-stage negligence with a clear causal chain, whereas this scenario is characterized by the AI system’s autonomous decision-making, which complicates the causal chain. The two together constitute the two poles of the problem of AI-inflicted harm.

3. The Basic Concepts for Determining the Criminal Liability of AI Providers

The distinction between the two scenarios shows that the logic of attributing harm caused by generative artificial intelligence varies by scenario. In technical-defect scenarios the causal chain is clear, and the traditional theory of negligent crimes remains applicable; in independent harm infliction scenarios the AI system autonomously produces harmful results, and the causal chain becomes complex through the intervention of autonomous decision-making. Accordingly, the determination of providers' criminal liability cannot stop at technical judgments within concrete scenarios; it must answer, at the normative level, the questions of the criminal law's functional positioning and the boundaries of its intervention.

3.1. Advocating the Concept of Preventive Criminal Law in the Context of the Risk Society

In independent harm infliction scenarios, if attribution waits until the harm has materialized, the retrospective tracing of causation will face significant obstacles because of the intervention of autonomous decision-making, and the effect of legal-interest protection will be greatly compromised. This creates a practical demand for the concept of preventive criminal law, whose core appeal is to curb risks before harmful results occur by adding preventive norms.

The risk society theory proposed by the German sociologist Ulrich Beck constitutes a central interpretation of the essential features of post-modern society [9]. The theory reveals that while modern society creates enormous wealth, it also systematically produces unprecedented forms of risk that spread across regions, are irreversible, and resist precise quantification. When a clear causal correspondence between the makers of risk and its bearers cannot be established, the post-hoc punitive model premises on whoever acts, answers' faces fundamental limits. Against this background, the functional positioning of the criminal law begins to shift progressively from post-hoc punishment to prior prevention.

The risk structure peculiar to generative artificial intelligence provides a typical setting for applying the concept of preventive criminal law. Its risks are hierarchically transmissible, spanning technical-level risks, content-level risks, and social-level risks; they are also universally diffusive, as the low entry barrier leads to a socialized spread of risk and raises the urgency of prevention [10]. By moving the point of criminal-law intervention forward to the stage of risk formation, preventive criminal law regulates situations in which the risk has acquired a concrete direction before the harm occurs, thereby making up for the functional deficiency of the ex post accountability model in independent harm infliction scenarios.

However, the introduction of the concept of preventive criminal law must be handled with caution regarding the normative boundaries it may overstep. The indirect governance model that the criminal law adopts toward digital technology risks shows that, given the policy consideration of encouraging technological innovation, premature reactions to technological risks are undesirable [11]. The continuing contest between preventing technological risks and preventing technological atrophy constitutes the core tension in applying the concept of preventive criminal law in the AI field. On the whole, this concept is a forceful theoretical response to the novel risks of artificial intelligence, but its application must abide by the normative boundaries of taking concrete danger of legal-interest infringement as the precondition for intervention and of treating indirect governance as the basic path, and it must rely on the balance and constraint of the modesty principle of the criminal law.

3.2. Proposing the 'Prudent and Tolerant' Concept in the Context of Technological Innovation

The expansionist tendency inherent in the concept of preventive criminal law needs to be balanced by the modesty principle of the criminal law. The distinction between the two scenarios shows that indiscriminately elevating providers' supervisory duties to criminal liability may confuse normative logic and unduly suppress technological innovation. The 'prudent and tolerant' concept emphasizes that, in a field of highly uncertain technology such as artificial intelligence, criminal law should act

in some respects and refrain in others, observing the normative bottom line of not hindering technological innovation [12].

First, the ultima ratio nature of the criminal law. As the most severe means in the state sanction system, criminal law intervention should be premised on the insufficiency of non-criminal means such as civil and administrative measures to achieve the function of legal-interest protection. The EU Artificial Intelligence Act relies mainly on administrative fines, and the Interim Measures for the Management of Generative Artificial Intelligence Services also mainly adopt administrative-law regulatory paths; both reflect the subsidiary position of criminal law.

Second, the theory of 'permitted risk' and the narrowing of the attribution scope. Many activities in modern society carry certain risks but are tolerated by law because of the notable social benefits they bring. The theory of objective attribution constructed by Professor Claus Roxin first excludes, at the stage of risk creation, conduct falling within the scope of permitted danger from attribution [13]. As a strategic technology leading a new round of scientific and industrial revolution, artificial intelligence necessarily involves inherent risks such as the algorithmic black box and uncertain output. When harm stems from inherent technological attributes, adequate causation between the provider's conduct and the harmful result fails because the chain breaks at the point of 'autonomous algorithmic decision-making.' Attempting to expand the determination of causation in order to pursue providers' liability for negligent crimes would not only contradict the basic logic of that theory but also produce a 'chilling effect' on technological innovation. It must be stressed that the application of this theory is not unconditional: a developer who violates basic ethical standards or criminal law norms in developing an intelligent system falls within the category of impermissible risk.

Third, review under the proportionality principle. The proportionality principle comprises three sub-principles: suitability, necessity, and balancing. Suitability requires that the means be apt to achieve the end; in independent harm infliction scenarios, harm stems from the AI's autonomous decision-making rather than the provider's negligent conduct, so regulating it through liability for negligent crimes lacks substantive connection. Necessity requires choosing the means least invasive to fundamental rights; some scholars have introduced the 'regulatory pyramid theory,' advocating a tiered structure composed of enterprise self-regulation, administrative regulation, and criminal regulation, and using the proportionality principle to move some conduct step by step from the criminal-law tier back to the administrative tier or even the tier of platform self-governance [14]. Balancing requires that the harm caused by the means be proportionate to the benefits gained; an excessively heavy expectation of criminal liability will prompt developers to adopt overly conservative technical strategies and to avoid exploratory R&D directions that might bring significant social benefits. Examined in this light, in technical-defect harm the causal connection between source-stage negligence and the harmful result is clear, and applying traditional negligence liability remains suitable and necessary; in independent harm infliction scenarios, the causal chain is broken by the intervention of autonomous decision-making, so the intervention of negligence liability is neither suitable nor the least invasive means, and the focus of criminal attribution should therefore shift to the post-hoc stage.

In sum, ultima ratio demarcates the external boundary of criminal-law intervention; the theory of permitted risk excludes the attributability of harm caused by inherent technological risks; and the proportionality principle provides a refined measuring standard. Together they form a progressive narrowing system running from the macro to the micro level.

4. The Typological Determination Paths for Negligent Crimes by AI Providers

Based on the above concepts, the two scenarios differ fundamentally in their logic of harm production, and their determination paths should differ accordingly. The technical-defect scenario centers on the provider's source-stage negligence with a clear causal chain, leaving ample room for the traditional theory of negligent crimes; the network-platform independent harm infliction scenario involves the intervention of autonomous decision-making by AI systems, complicating the causal

chain, and requires separate examination of the prior supervisory duty and the post-hoc removal duty in order to determine whether, and within what scope, the provider bears criminal liability.

4.1. The Determination Path for Traditional Negligent Crimes Exemplified by Technical Defects

The core feature of technical-defect harm is that its root can be traced to the provider's source-stage carelessness in research, development, and deployment; the causal chain is clear, and there is no substantial intervention of autonomous algorithmic decision-making. This feature determines its inherent compatibility with the traditional theory of negligent crimes, but the constitutive elements of negligence still need to be concretized in light of the characteristics of AI technology.

First, the determination of the possibility of foreseeing the result. The traditional theory of negligence takes the possibility of foreseeing the result as the core of negligence. Latent bias in training data, loopholes in algorithmic logic, and deficiencies in permission management systems are not wholly unforeseeable unknown risks; they are risks that can be identified and avoided through standardized R&D procedures and adequate quality inspection. As the creator of the risk source, the provider possesses the corresponding capacity to recognize and foresee the above technical defects and the legal-interest infringement they may cause. It should be emphasized that the foresight required here is not foresight of every concrete detail of the harmful result, but foresight of the danger of legal-interest infringement inherent in conduct that violates the duty of care.

Second, breach of the duty of result avoidance. The new theory of negligence shifts the core of negligence from the possibility of foreseeing the result to breach of the duty of result avoidance. Criminal attribution for generative-AI-related crimes should not simply adhere to the traditional theory of negligence; the provider's duty of result avoidance should be assessed within the framework of the new theory, while the possibility of foreseeing the result under the traditional theory serves as a prior limiting condition [15]. In technical-defect scenarios, the provider bears duties of result avoidance such as ensuring algorithmic reliability, safeguarding the quality of training data, and improving the security protection system during the R&D and deployment stages. Failure to take reasonable measures of result avoidance constitutes breach of the duty of care. But not every breach of a technical norm or industry standard by a provider automatically constitutes negligence under criminal law; it must further be judged whether the breach substantially increased the risk of legal-interest infringement.

Third, the determination of causation. In technical-defect scenarios, the causal chain between the provider's negligent conduct and the harmful result is clear; there is no substantial intervention of an intermediate variable such as the autonomous decision-making of the algorithmic black box, and the doctrine of adequate causation can properly resolve attribution [16]. Judging by the experience of ordinary members of society, technical defects introduced through a provider's carelessness will ordinarily suffice to cause abnormal decisions of artificial intelligence and resulting harm; the two stand in a relationship of adequacy. The obstructive effect of the algorithmic black box reaches only scenarios in which the autonomous decision-making of artificial intelligence cuts the causal chain; for technical-defect harm caused by the provider's source-stage carelessness, the attribution logic of the traditional theory of negligent crimes still has ample room for application.

In sum, the determination path for this scenario should be as follows: the provider's state of foreseeability and controllability over technical defects serves as the basis of attribution; breach of the duty of result avoidance serves as the core of negligence determination; and adequate causation serves as the standard of causal judgment. This path takes the possibility of foreseeing the result under the traditional theory of negligence as its premise while responding to the new theory's emphasis on the duty of result avoidance, constituting the basic paradigm for applying the traditional theory of negligent crimes in AI scenarios.

4.2. The Determination Path for Independent Harm Infliction by Artificial Intelligence Exemplified by Network Platforms

The 'spontaneous' generation of unlawful content on network platforms exhibits the core feature of independent harm infliction by artificial intelligence: the AI system produces harmful results on its own, detached from direct human intervention, and the provider has not actively intervened in the content generation mechanism. This feature determines that the determination of negligent crimes must examine two levels, the prior and the post-hoc, separately.

4.2.1. The Collapse of the Supervisory Duty and the Exclusion of Negligent Crimes

In scenarios of independent harm infliction by artificial intelligence, elevating the provider's supervisory responsibility to criminal liability for negligent crimes encounters structural obstacles to the premises of the theory of supervisory negligence. The basic structure of that theory is that the supervisor bears criminal liability for failing to effectively supervise the supervised person's negligent conduct, and its core premise is that the supervised person's conduct can be evaluated as negligent conduct under criminal law. In independent harm infliction scenarios, however, the harm stems from autonomous algorithmic decision-making rather than the provider's carelessness; adequate causation is hard to establish, foreseeability becomes overly abstract because of the algorithmic black box, and the AI system's output cannot be evaluated as negligent conduct in the sense of criminal law [17]. Some scholars have noted that the involvement of multiple actors in AI technology makes unclear causation the core obstacle to attribution, and that a theoretical path for narrowing or excluding negligence attribution should be constructed on the basis of the doctrine of adequate causation. When the causal chain is broken by the intervention of autonomous decision-making and foreseeability remains at an abstract level, holding the provider liable for a negligent crime on the ground of failing to supervise a non-human subject lacks sufficient dogmatic justification.

Second, the conflict between attributed responsibility and self-responsibility. Attributed responsibility assigns one party's responsibility to another even though the latter did not directly participate in the occurrence of the conduct. Applying it to scenarios of independent harm infliction by artificial intelligence would fundamentally conflict with the principle of self-responsibility in criminal law. The basic requirement of criminal attribution is that the actor bears responsibility only for his own conduct; attributing harm caused by artificial intelligence to the provider essentially requires the provider to bear criminal liability for the autonomous conduct of a non-human subject, which can hardly find adequate justification under the principle of responsibility. If artificial intelligence itself cannot become a subject of criminal liability, then the path of attributing the consequences of its conduct to the provider likewise lacks a logical intermediary in dogmatic terms [18].

Finally, the 'responsibility gap' and the realistic path of legal-interest protection. The scenario of independent harm infliction by artificial intelligence raises the so-called 'responsibility gap' problem in criminal law theory: when an AI system causes harm, it may be difficult to identify a subject who can be held accountable. The traditional framework of criminal law is built on human intention and voluntary conduct and responds poorly to harm caused by autonomous or semi-autonomous AI systems; shifting responsibility to developers or operators faces the double obstacle of causation and blameworthiness. On the one hand, the harm stems from autonomous algorithmic decision-making rather than the developer's specific negligent conduct, so the 'adequacy' of adequate causation cannot be established. On the other hand, the blameworthiness of a negligent crime rests on the possibility of foreseeing the result and the possibility of avoiding it, while AI output is inherently unpredictable; requiring the developer to bear negligence liability for every concrete harm amounts to requiring him to bear criminal responsibility for inherent technological attributes falling within the scope of 'permitted risk [19].'

However, the existence of a responsibility gap does not mean that legal-interest protection must fail. Foreign legislative practice also provides a reference: the EU Artificial Intelligence Act relies mainly on administrative fines, and conduct violating the 'prohibited' domains is subject to

administrative fines of up to 35 million euros or 7% of total global annual sales, whichever is higher [20]. This legislative choice shows that, for harm caused by AI systems, the legislator's primary response is to set behavioral standards and compliance obligations through administrative regulation rather than to impose criminal sanctions directly. This governance approach of 'administration first, criminal law as the exception' offers a useful reference for allocating responsibility in independent harm infliction scenarios.

4.2.2. The Activation of the Removal Duty and the Application of Duty Offenses

In contrast to the absence of a duty of prior prevention, providers bear a peculiar removal duty in network platform scenarios. Article 14 of the Interim Measures for the Management of Generative Artificial Intelligence Services expressly stipulates that when a provider discovers unlawful content, it shall promptly take disposal measures such as stopping generation, stopping transmission, and elimination. This duty is directed at the post-hoc stage in which unlawful content has already been generated and begun to spread.

The core structure of the crime of refusing to fulfill information network security management obligations is that a network service provider, having failed to perform the information network security management obligations prescribed by laws or administrative regulations, refuses to take corrective measures after being ordered by the regulatory department to do so, thereby causing unlawful information to spread in large quantities. This crime is a duty offense: the obligation of information network security management is rooted in the network service provider's advantageous position as a 'digital gatekeeper.' Zheng Zexing and Zeng Qinwen point out clearly that operators of generative AI services are the operating subjects that connect foundation models with end users and exercise substantial control over the service process; their criminal liability unfolds precisely within the framework of the crime of refusing to fulfill information network security management obligations, and the assumption of obligations should be distinguished from the technology development responsibility of foundation model providers.

Viewed from the path of attribution, as a duty offense, this crime does not depend on complete proof of causation between the conduct and the harmful result; it focuses on whether the provider performed the statutory information network security management obligations. In independent harm infliction scenarios, insisting on the path of prior supervisory negligence to pursue the provider's criminal liability always faces the obstacle that adequate causation is hard to establish; the structure of the duty offense bypasses this problem, because the provider knows or should know that the unlawful content exists and is spreading on the platform, and the focus of the question shifts from 'could it have been foreseen' to 'whether action was taken.' Whenever the provider, after being ordered by the regulatory department to take corrective measures, still refuses to perform the removal duty and thereby causes unlawful information to spread in large quantities, crime is constituted.

As to the subjective element, the requirement of this crime for the provider's subjective aspect is also particular. In independent harm infliction scenarios, the provider lacks concrete foreseeability of the specific path by which the AI system autonomously generates unlawful content; his subjective state is closer to a general, abstract foresight of danger. Through the objective element of 'refusing to take corrective measures after being ordered by the regulatory department, crime focuses subjective blameworthiness on the refusal to comply with the regulatory order rather than on concrete foresight of the harmful result. Once the provider receives the corrective notice from the regulatory department, he is already in a state of actual knowledge of the objective existence of the unlawful content, and his refusal to correct consequently possesses the basis of blameworthiness under criminal law.

To be sure, the application rate of this crime in judicial practice has long been low—in 2023 it accounted for only 0.39% of cybercrime cases [21]—and its boundaries of application and preconditions remain to be clarified. But in terms of normative logic, the crime of refusing to fulfill information network security management obligations provides a feasible path, distinct from traditional negligent crimes, for determining providers' criminal liability in network platform scenarios: where the causal chain is broken by the intervention of autonomous decision-making, legal interests are protected through the structure of a duty offense, which both responds to the particularity

of the network platform scenario and avoids the dogmatic dilemma of forcing criminalization by expanding the determination of causation.

5. Conclusion

The attribution of harm caused by generative artificial intelligence cannot be properly resolved within a single theoretical framework. Existing studies either directly affirm the provider's liability for negligent crimes or trace attribution backwards by lengthening the causal chain, but whether the logic of attribution differs substantially across harm scenarios has always lacked systematic explanation. This paper answers the question through typological analysis: according to differences in the logic of harm production, it distinguishes the relevant situations into two scenarios and examines the forms of the provider's responsibility in each.

As for the traditional negligence attribution scenario exemplified by technical defects, the root of the harm can be traced to the provider's source-stage carelessness in research, development, and deployment, and autonomous algorithmic decision-making does not substantially intervene. The provider has the capacity to foresee and avoid such technical defects and the legal-interest infringement they may cause; the traditional theory of negligent crimes remains applicable, and its determination should take the state of foreseeability and controllability as the basis of attribution, breach of the duty of result avoidance as the core of negligence, and adequate causation as the standard of causal judgment.

As for the scenario of independent harm infliction by artificial intelligence exemplified by network platforms, the AI system produces harmful results on its own, the causal process becomes complex with the intervention of autonomous decision-making, and the prior supervisory duty faces multiple dogmatic obstacles: the core premise of the theory of supervisory negligence cannot be satisfied, attributed responsibility conflicts with the principle of self-responsibility, and the responsibility gap deprives causal attribution of a dogmatic intermediary. Since the duty of prior prevention is hard to establish, the focus of attribution shifts to the post-hoc stage. The provider still bears a peculiar removal duty; the crime of refusing to fulfill information network security management obligations, as a duty offense, does not depend on a complete proof of causation but focuses on whether the statutory duty was performed, thereby providing a feasible path distinct from traditional negligent crimes for independent harm infliction scenarios.

All in all, criminal-law intervention should follow the principle of typological distinction and differentiated treatment: giving play to the regulatory effect of the traditional theory of negligent crimes where causation is clear, observing the modest character of the criminal law where the causal chain is broken, taking administrative regulation as the dominant governance means, and coordinating development and security on the track of the rule of law.

References

[1] Xi Jinping: 'Working Together to Build a Fair and Reasonable Global Artificial Intelligence Governance System'—Keynote Speech at the Opening Ceremony of the 2026 World Conference on Artificial Intelligence and the High-Level Conference on Global AI Governance, published in the Gazette of the State Council, July 17, 2026, No. 21 of 2026.

[2] Sha Rina. Determination of the World's First Fatal Autonomous-Driving Accident: Uber Not Criminally Liable. People's Daily Online. March 7, 2019. Information on http: // auto. people.com.cn/n1/2019/0307/c1005-30962032.html (last visited August 26, 2026).

[3] Tesla Held Partly Liable in the Fatal Autopilot Crash Case. Tencent News. September 16, 2025. Information on: https://news.qq.com/rain/a/20250916A058MF00 (last visited August 26, 2026).

[4] Liu Xianquan. A Re-Examination of Theoretical Research on the Evolution of the Subject of Criminal Liability in the Artificial Intelligence Era. Law Science. No. 9 (2025), p. 82-94.

[5] Ren Yanjun, Nie Ran. The Negation of the Subject Qualification of Strong Artificial Intelligence for Criminal Liability. Tribune of Social Sciences. No. 4 (2024), p. 132-141.

[6] ProPublica. Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks. ProPublica. 23 May 2016. Information on: https: //www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[7] Zhang Xuehong. Final Judgment in Beijing's First Criminal Case of Damaging an Artificial Intelligence Model: Engineer Convicted of the Crime of Damaging Computer Information Systems and Sentenced. Guangming Online. August 19, 2026. Information on: https: // www.toutiao.com/article/7675557807453651506/?wid=1787113442186 (last visited August 27, 2026).

[8] Interim Measures for the Management of Generative Artificial Intelligence Services, Article 14: 'Where a provider discovers unlawful content, it shall promptly take disposal measures such as stopping generation, stopping transmission, and elimination, take measures such as model optimization and retraining to rectify the situation, and report to the relevant competent department.'

[9] Ulrich Beck. Risk Society. Trans. He Bowen. Yilin Press, 2004, p. 66.

[10] Zheng Zexing, Zeng Qinwen. On the Criminal Liability of Operators of Generative Artificial Intelligence Services. Contemporary Law. No. 5 (2025), p. 52-62.

[11] Chu Chencheng. Indirect Governance: The Basic Model of Criminal Law's Response to Digital Technology Risks. Legal Forum. No. 1 (2025), p. 65-77.

[12] Fang Huiying. Criminal Risks and Prevention Strategies of Generative Artificial Intelligence—Taking ChatGPT as an Example. Journal of Nanchang University (Humanities and Social Sciences Edition). No. 4 (2023), p. 52.

[13] Claus Roxin. German Criminal Law (Vol. 1). Trans. Wang Shizhou. Law Press, 2005, p. 392.

[14] Colin Scott. Regulation, Governance and Law: Research on Frontier Issues. Trans. An Yongkang. Tsinghua University Press, 2018, p. 290-291.

[15] Zhang Weiyao. Result Attribution in Negligent Crimes of Generative AI Producers. Journal of East China University of Political Science and Law. No. 6 (2025), p. 75-85.

[16] Wang Chunyu. The Determination and Regulation of Criminal Liability for Crime Methods Output by Large Models. Jiangxi Social Sciences. No. 6 (2026), p. 197-205.

[17] Fabio Basile, Beatrice Fragasso. Intelligenza artificiale e diritto penale: quattro possibili percorsi di indagine. Diritto Penale e Uomo. 2019, p. 1-33. Information on https: // www.penalecontemporaneo.it/upload/3089-basile2019.pdf.

[18] Gabriel Hallevy. The Basic Models of Criminal Liability of AI Systems and Outer Circles. SSRN Electronic Journal. (11 June 2019), p. 1-16. Information on: https://ssrn.com/abstract=3402527.

[19] Wang Shan, Wang Jing. Attribution of Responsibility for the Abuse of Artificial Intelligence Technology and the Criminal Law Response. DeHeng Law Offices website. March 6, 2026. Information on: https://www.deheng.com/content/26410.html (last visited September 25, 2026).

[20] EU Artificial Intelligence Act, Article 99(3): 'Non-compliance with the prohibited AI practices referred to in Article 5 shall be subject to administrative fines of up to 35,000,000 euros; where the offender is an enterprise, the fine may be up to 7% of its total worldwide annual turnover in the preceding financial year, whichever is higher.'

[21] Editorial Board of China Procuratorial Yearbook (ed.). China Procuratorial Yearbook (2024 Volume). China Procuratorial Press, 2024, Part VI 'Statistical Data.'