The obstacle of artificial intelligence embedded in the criminal law system and the limitation of producers' criminal responsibility

Wangxu Hao

National University of Singapore, 259776 (SG), Singapore

Abstract: *The core of artificial intelligence lies in its technical processes, which often lack transparency. This poses challenges to integrating it into legal frameworks, making it unsuitable to be regarded as an entity capable of bearing criminal responsibility. Given that AI lacks the capacity to assume criminal liability, discussions on the attribution of responsibility for crimes involving AI should focus on its developers or users. In most cases, developers, as the primary responsible parties, typically bear liability in cases of negligence. When determining such liability, the application scope of negligence-related offenses should be appropriately limited, guided by updated negligence theories. Developers may be held accountable if they fail to take measures to prevent adverse outcomes when potential risks were foreseeable, as this constitutes a breach of their duty of care, warranting prosecution for negligence.*

Keywords: *Artificial Intelligence; Criminal Responsibility; Duty of Care; Limited Attribution*

1. Introduction

In recent years, artificial intelligence (AI) technology has advanced rapidly, with generative AI systems like ChatGPT becoming deeply integrated into everyday life. This evolution has transformed tools from simple assistive instruments into more autonomous entities. Such a shift not only redefines the role of tools but also gives rise to new forms of crime. As methods of committing crimes evolve, the causal chains behind them, the mental state of perpetrators, and the definition of legal responsibility become increasingly complex[1].

Amid these challenges, a key question arises: Can AI be regarded as a subject capable of bearing criminal responsibility? In criminal cases involving AI, how should criminal liability be reasonably distributed? Furthermore, what evaluative standards must be established to accurately assess such crimes? Addressing these issues requires integrating the fundamental principles of computer science with criminal law theories, clarifying the unique nature of AI-related crimes, and constructing a legal regulatory framework that meets the demands of the AI era.

2. Challenges of Embedding Artificial Intelligence into the Criminal Law System

2.1 The Dilemma of Artificial Intelligence as a Subject of Criminal Liability

2.1.1 Affirmative Arguments

Proponents of this view tend to base their arguments on the trajectory of advancements in computer technology, proposing through theoretical assumptions and deductions that AI should be granted the status of a subject of criminal liability[2]. Their main arguments include:

Independence of Responsibility Capacity: When intelligent machines commit criminal acts beyond their programmed instructions, they demonstrate the ability to distinguish right from wrong and exercise self-control.

Anthropocentrism and Utilitarianism: From the perspective of safeguarding overall human interests and adopting a utilitarian viewpoint, incorporating AI into the scope of criminal law regulation helps better ensure public safety and well-being.

Beyond Traditional Product Liability: For autonomous AI systems, traditional theories of product liability, such as manufacturing defects or misuse, are difficult to apply. Recognizing them as subjects of liability is necessary to address gaps in the legal framework.

Similarity to Natural Persons and the Purposes of Punishment: Although AI is not a biological entity, it has the capacity to "experience" adverse consequences, enabling it to be deterred from reoffending. This aligns with legal practices that allow for the constructive recognition of non-biological entities as subjects of liability.

2.1.2 Negative Arguments

Opponents argue that neither purpose-specific AI nor multifunctional AI systems, as technical products based on programming code, should be granted the status of subjects of criminal liability[3]. Their main points include:

Lack of Understanding of Social Significance: AI cannot comprehend the societal implications of its actions. Its behaviors are merely advanced algorithmic responses to data, akin to biological instinct rather than choices based on free will.

Ineffectiveness of Penal Functions: Since AI lacks the capacity to experience pain or suffering, criminal punishment would fail to achieve either specific deterrence or general deterrence.

Stability of the Legal System: Recognizing AI as a subject of criminal liability could disrupt the existing legal framework by blurring the concepts of subjects and objects of law. To maintain the stability of the legal system and ensure it continues to serve human interests, AI should not be considered a legal subject.

2.2 Technical and Legal Barriers to Embedding Artificial Intelligence into the Criminal Law System

2.2.1 Technical Aspect: The Complexity of the Nature of AI Behavior

To make AI systems perform more like humans, researchers have translated human thinking processes into symbolic language, encoding brain functions and psychological activities to construct AI generation systems based on symbolic processing[4]. Despite demonstrating remarkable self-learning capabilities and human-like traits of "autonomous choice," allowing for adaptive responses to environmental changes, AI fundamentally remains an embodiment of computationalism and functionalism.

2.2.2 Legal Aspect: The Insufficiency of Criminal Law Theory in Adapting to AI

According to the principles of criminal law interpretation, proving that AI can serve as a subject of criminal liability requires a clear analysis of its self-recognition, independent decision-making, autonomous actions, and learning and creative processes. This is because the effective resolution of any issue relies on a rational foundation of argumentation. However, the lack of transparency in the complex algorithms, models, and operational mechanisms of AI systems makes it challenging to predict whether their decisions originate from the so-called "algorithmic black box." The essence of criminal responsibility lies in accountability, and accountability requires the ability to adequately explain behavior. Given the unexplainable nature of AI behavior, this poses a significant obstacle to holding AI accountable.

2.2.3 Comprehensive Analysis: The Difficulty of Allocating Responsibility Among Multiple Actors

Addressing the issue of AI responsibility allocation within the criminal law framework requires a dual consideration of technological and legal constraints. Theoretically, it is essential to define the fundamental attributes of AI behavior and its basis for assuming responsibility. Practically, a reasonable mechanism for responsibility-sharing must be established, ensuring that developers, users, regulators, and AI systems collectively form a cohesive responsibility framework. This approach aims to address gaps in criminal law interpretation and liability attribution.

3. Attribution of Criminal Liability in AI-Related Offenses

3.1 The Basis and Limitation of Producers' Criminal Liability

3.1.1 Basis of Producers' Criminal Liability

From the perspective of criminal law, the actions of developers exhibit both causal links and normative requirements. During the design, development, testing, and market deployment phases, developers may, through negligence or intentional acts, cause AI products to bring adverse effects to society. Therefore, developers should bear corresponding criminal liability based on the following aspects:

Duty of Foresight: Developers should anticipate the potential adverse outcomes of their AI products and implement effective preventive measures to mitigate risks.

Duty of Compliance: Developers must adhere to existing technical standards and legal regulations, ensuring that their AI products meet safety and ethical requirements.

Negligence: When developers fail to fulfill reasonable duty of care, leading to harmful incidents caused by AI products, they should be held legally responsible for their negligent behavior. In such cases, the key to determining whether developers are guilty of negligence lies in whether a breach of duty of care can be established. If it cannot be proven that developers violated necessary duties of care, even though this may create a liability gap, the principles of "presumption of innocence" and "principle of restraint in criminal law" dictate that such gaps should be considered socially acceptable risks, ultimately borne collectively by society.

3.1.2 The Necessity of Limiting Producers' Criminal Liability

Although producers play a critical role in AI systems, the self-learning capabilities and complex operational patterns of AI make it difficult for producers to fully control the outcomes of AI behavior. Expanding their criminal liability excessively would conflict with the principle of fairness[5]. The necessity of limiting producers' liability is based on the following reasons:

Autonomy Issue: During operation, AI may generate unforeseen behaviors due to its self-learning and adaptive characteristics, making it difficult to attribute such actions to the producer.

Technical Unpredictability: The complexity of AI algorithms and the "black box" nature of its processes prevent producers from fully predicting the consequences of AI behavior.

Encouraging Social Innovation: Overburdening producers with liability could stifle innovation and development in AI technologies.

3.1.3 Framework for Limitation

To reasonably limit the criminal liability of producers, the following framework should be established, Shown in Table 1:

Table 1: Framework for Limitation

Framework Element	Details
Definition of Behavioral Scope	Clearly define producers" direct causal liability for Al behavior, excluding outcomes caused by uncontrollable factors.
Risk Prevention Obligations	Require producers to fulfill reasonable risk forecasting and control measures, exempting liability if technical and legal standards are met.
Multi-party Responsibility Allocation	Extend liability to Al users, managers, and other relevant parties, allocating responsibility based on their degree of involvement.
Legal Exemption Provisions	Apply liability exemptions if producers can demonstrate they have fulfilled necessary duties of care and implemented safety measures.

3.2 Attribution of Users' Criminal Liability

3.2.1 Determining Negligence of Direct Users' Conduct

When artificial intelligence fails to act according to the illegal intentions of its producer or user, this scenario can be regarded as a case of "cognitive error regarding a tool." It is akin to a situation where an individual mistakenly believes they are firing a loaded gun but fails to cause the intended harm. In such cases, although the criminal objective is not fully achieved, the producer or user already fulfills the subjective and objective elements of the offense[6]. The misunderstanding of the actual effects of the AI as a means of committing the crime warrants treating the incident as an attempted crime, holding the producer or user criminally responsible accordingly.

3.2.2 Attribution of Users' Responsibility in Complex Scenarios

Based on the specifics of the situation, the producer or user may be charged with attempted intentional injury against Party A. Regarding Party B’s injury, since no negligent form of the offense of obstructing official duties exists, the act would only constitute the crime of causing grievous bodily harm to Party B through negligence. Under the doctrine of imagined concurrence of offenses, the more severe charge would be applied.

Alternatively, under the principle of legal consistency, the producer or user may be charged with attempted intentional injury against Party A and, despite the harm to Party B being caused negligently, it could be treated as intentional. This would result in the producer or user being charged with completed intentional injury against Party B. Similarly, under the doctrine of imagined concurrence, the more severe charge would be applied.

3.3 Adjusting Liability Attribution under Cognitive Errors

3.3.1 Cognitive Bias Caused by Technical Limitations

If the harmful consequences occur because current human technological and cognitive levels are unable to identify defects in artificial intelligence products, in such cases, even though the harmful result happens within the user’s domain, the user should not necessarily be held criminally responsible for negligence. The determination should be made based on the interaction between the user and the AI product. When the AI product independently completes a task, the user should not be presumed to have acted negligently and thus should not face punishment. However, if the user, while collaborating with the AI product to complete a task, triggers a serious harmful outcome to society, the user should bear criminal responsibility within the scope of the legal duties of care imposed on them.

3.3.2 Conditions and Scope of Liability Exemption

In certain cases, the criminal law principles can be applied to grant exemption clauses for the responsible parties involved in artificial intelligence-related actions:

Conditions:

The user can prove that they have exercised reasonable care as required by operational standards.

The harmful outcome was entirely caused by the unpredictable characteristics of the AI product, rather than the user's actions.

The user did not significantly participate in the harmful actions in a manner of gross negligence or intent during the task collaboration.

Scope of Application:

The user reasonably trusted the AI product to function normally but encountered uncontrollable consequences.

The user, given their current level of understanding, failed to detect potential technical defects.

The harmful result is closely related to the AI’s independent decision-making, with no direct causal link to the user's actions.

4. Framework for Limiting Producers’ Criminal Liability

4.1 Grounds for Limiting Producers’ Negligent Criminal Liability

4.1.1 Application of the “Permissible Risk” Theory

Risks that are neither foreseeable nor preventable are considered “permitted risks.” While such risks may create gaps in responsibility attribution, imposing these responsibilities on producers, users, or victims is clearly unfair and unreasonable[7]. Considering that the benefits brought by artificial intelligence technology are shared by society as a whole, the damages resulting from the materialization of these risks should also be borne collectively by society.

4.1.2 Exemption from Liability for Damages Caused by Justifiable Actions

In criminal law, a justifiable act refers to actions that objectively result in harmful consequences and superficially meet the elements of a crime but lack social harmfulness and thus do not constitute criminal wrongdoing. Besides self-defense and necessity, justifiable acts include actions performed under legal mandates and those meeting industry standards as professional conduct. In cases where AI product producers might face liability exemptions, the most relevant scenarios involve compliance with legal mandates and adherence to industry standards. Such acts, due to their justification and legality, should not be grounds for producers to bear criminal responsibility.

4.2 Breach of Duty of Care as the Core Criterion for Negligent Criminal Liability

4.2.1 Theoretical Basis: From Traditional Negligence Theory to Modern Negligence Theory

Traditional negligence theory focuses on the actor's responsibility to foresee possible outcomes, often neglecting the obligation to prevent harm, which can lead to setting overly low standards for violations of duty of care[8]. The new negligence theory, with its emphasis on vague foresight concepts like unease or premonition, attempts to address this gap but often proves too abstract for practical application. By contrast, new negligence theory takes baseline norms as its starting point, stressing that as long as actions remain within the boundaries of these norms, they should not be considered unlawful or meet the elements of a crime. Thus, when assessing whether a producer has violated their duty of care and may have committed a negligent crime, the new negligence theory offers a more reasonable framework.

4.2.2 Practical Application: Standards for Determining Breach of Duty of Care

A producer’s responsibility to foresee potential risks forms the basis of their obligation to prevent those risks. If a producer cannot foresee a specific risk, they cannot be held responsible for preventing it. However, even when a producer is capable of foreseeing potential risks, this does not automatically impose a legal obligation to prevent those risks in the context of criminal law. Under the principle of trust, producers are justified in believing that once an AI product enters another party's management domain, the latter will adhere to relevant regulations and take appropriate actions.

5. Conclusion

Despite the human-like characteristics of AI, it fundamentally differs from humans in essential ways. These differences preclude AI from becoming a subject of criminal liability. The issue of criminal liability arising from AI-related crimes should instead fall on its producers or users. Before AI-related crimes become pervasive, exploring the criteria for judging the unlawfulness and culpability of producers and users can help preemptively mitigate technological risks and balance the responsibilities of producers and users with the risks and appropriateness of criminal liability in the context of AI products.

References

[1] Zhang Jingang. Criminal Risks in the Era of Artificial Intelligence: A Focus on Intelligent Robots [J]. Journal of Chongqing University of Technology (Social Sciences), 2019(8):87-94.

[2] Shi Fang. The Denial of Artificial Intelligence as a Criminal Subject [J]. Science of Law (Journal of Northwest University of Political Science and Law), 2018(6):67-75.

[3] Zhang Jinsong. Man as the Measure of Machines: On Artificial Intelligence and Human Subjectivity [J]. Studies in Dialectics of Nature, 2017(1):49-54.

[4] Liu Yanhong. A Study on the Explainability of Artificial Intelligence and Its Legal Responsibility [J]. Law and Social Development, 2022(1):78-91.

[5] Yang Fengyi. Between Adherence and Reform: Criminal Risks of Artificial Intelligence and the Path of Criminal Law Responses [J]. Journal of Jilin University (Social Sciences), 2021, 61(05):79-90+236-237.

[6] Li Wenjing, Yang Hongbin. An Analysis of Algorithmic Legal Regulation and Digital Human Rights Protection in the Era of Artificial Intelligence [J]. Journal of Sichuan Police College, 2022(3):52-59.

[7] Fang Huiying. The Doctrinal Analysis of Attribution and Determination of Criminal Responsibility for Artificial Intelligence Crimes [J]. Shandong Social Sciences, 2022(4):142-148.

[8] Xu Yongwei. The Position of Criminal Governance in the Era of Artificial Intelligence: Reflection and Construction Based on Subjects, Risks, and Responsibilities [J]. Shandong Social Sciences, 2022(10):169-175.