New Research in Islamic Humanities Studies Online ISSN : 2980 - 8944
Lorestan University

Assessment of Criminal Liability for Artificial Intelligence in Emergency Situations or Legitimate Defense

Mohammad Ali Abbaszadeh Dibavar[1], **Ebrahim Rajabi Taj Amir** [2] ✉, and **Sajad Akhtari**[3]

1. Department of Criminal Law and Criminology, Sav. C., Islamic Azad University, Saveh, Iran. Email: m.abbaszadehdibavar@iau.ir
2. Corresponding Author, Department of Criminology, Amin Police University, Tehran, Iran. Email: e.rajabi.t@gmail.com
3. Department of Criminal Law and Criminology, Ka. C., Islamic Azad University, Karaj, Iran. Email: Sajadakhtari2025@iau.ac.ir

Article Info

Article type:
Research Article

Article history:
Received 31 March 2025
Received in revised form 13 July 2025
Accepted 27 August 2025
Available online 28 September 2025

Keywords:
Artificial Intelligence,
Criminal Liability,
Self-Defense,
State of Necessity,
Multi-Layered Liability,
Iranian Criminal Law

ABSTRACT

Objective: This study aims to elucidate a framework for criminal liability regarding the decisions of autonomous artificial intelligence systems in sensitive situations involving self-defense and the state of necessity. The core issue lies in the inadequacy of traditional criminal liability models (based on human will and intent) as well as the limitations of novel approaches, such as foreseeability-based or risk-based liability, in addressing the complexity and self-learning capabilities of these systems.

Method: The research adopts a descriptive-analytical approach and utilizes a comparative study of advanced legal systems to analyze this challenge.

Results: The findings indicate that an effective solution requires moving beyond the "human or machine" dichotomy and adopting a multi-layered liability model. This model distributes responsibility across three distinct layers: the design and production layer (involving the moral and warranty liability of the designer), the operation and supervision layer (involving the precautionary and supervisory liability of the user), and the regulation and compensation layer (involving the warranty liability of the regulatory body and a compensation fund). This framework, while managing technological complexity and emphasizing ultimate human accountability, enables a fair analysis of incidents and facilitates the adaptation of self-defense and necessity criteria to algorithmic decisions.

Conclusions: Within the context of Iranian criminal law, this research highlights the necessity of moving beyond the existing "legislative paralysis" and proposes two pathways: the dynamic interpretation of existing legal institutions (such as liability arising from indirect causation) and proactive legislation based on the multi-layered model to ensure criminal justice and legal security. Consequently, the future of criminal law depends on embracing such flexible and distributed models to safeguard human dignity and the rule of law in the age of autonomous technologies.

Cite this article: Abbaszadeh Dibavar, M. A.; Rajabi Taj Amir, E. & Akhtari, S. (2025). Assessment of Criminal Liability for Artificial Intelligence in Emergency Situations or Legitimate Defense. *New Research in Islamic Humanities Studies*, 4(7), 1-23. https://doi.org/10.22034/api.2025.2081092.1606

Publisher: Lorestan University.
DOI: https://doi.org/10.22034/api.2025.2081092.1606

Introduction

Within the realm of human knowledge, each science explores phenomena based on its unique paradigm. Law, as a normative and prescriptive discipline, does not merely seek to describe facts, but rather aims to regulate social relations through principles of "ought" and "ought not." This nature places law in a distinct position when confronting novel phenomena: on one hand, it requires insights from empirical and technical sciences to accurately understand complex realities, and on the other hand, it is obligated to evaluate and formulate norms for those same realities using standards of justice, equity, and ethics. The emergence of "Artificial Intelligence" (AI), particularly in the branch of machine learning, has brought this dialectical relationship to an unprecedented challenge. By creating systems capable of learning, decision-making, and action in complex environments, AI has impacted the foundational concepts of criminal law. The nature and consequences of criminal liability for acts originating from the will and intent of an algorithm have become a central question in contemporary legal literature. This complexity is compounded when we consider the application of these systems in the most sensitive arenas: emergency situations and conditions of self-defense. In such contexts, intelligent systems may be compelled to make decisions requiring the immediate assessment of criteria such as necessity, proportionality, and imminence criteria that in traditional criminal law have relied on conscious will and human intent. Furthermore, different machine learning methods significantly influence the degree of predictability and control over a system's behavior. In supervised learning, the output behavior is largely predictable, whereas in reinforcement learning or deep learning, the system may, through dynamic interaction with its environment, learn behaviors or discover patterns that are unexpected even for its designers. This inherent dynamism and the "black box" characteristic of some complex architectures make tracing and attributing causality difficult, casting ambiguity upon traditional foundations of liability.

The existing research literature attests to the serious attention scholars have paid to various dimensions of this challenge. Domestic studies have focused, on one hand, on the strategic and defensive functions of AI (e.g., Mostafalassan and Dehestani, 1401 SH; Rostami, 1401 SH; Ghasemi et al., 1402 SH; Torki, 1403 SH; Hosseinzadeh et al., 1403 SH) and, on the other hand, on the primary foundations of its civil and criminal liability (e.g., Nooriso, 1404 SH; Zakerinia, 1402 SH). Internationally, discussions range from focusing on the legality of using autonomous systems in armed conflicts (e.g., Time Report, 2024; Exuis Magazine, 2024; Associated Press, 2023 in the context of warnings about automated weapons alongside analyses by the Lieber Institute, 2025) to examining the possibility of granting legal personality to AI (e.g., research by Al-Hamuri and Al-Kadi, 2024) and analyzing emerging crimes arising from it (e.g., studies by Panatoni, 2025; Abdel Aziz, 2025). A similar approach exists in comparative studies (USA, Germany, and the European Union). However, a clear gap is observed in connecting these two research streams: formulating a coherent framework for attributing criminal liability, while considering specific justificatory circumstances (such as self-defense and necessity) in algorithmic autonomous decisions. The main question is: how can we evaluate the actions of an intelligent system which

lacks human "will" and "intent" in light of the traditional rules of criminal liability, which are based on the three pillars of the material element, the mental element (intent), and attributability to a person? Can concepts like self-defense be virtually attributed to a system's behavior, with ultimate liability placed on the relevant humans (designer, operator, supervisor), or must we move towards entirely new models of multi-layered liability or risk-based liability?

Acknowledging this challenge, this article seeks, through a descriptive analytical approach and utilizing comparative studies, to examine the feasibility and method of attributing criminal liability to AI system decisions in situations of necessity and self defense. The central hypothesis is that traditional legal systems alone are incapable of adequately addressing this complexity, and that adopting a distributed, multi-level model of liability, in which ultimate accountability remains human-centric, represents a suitable solution for ensuring justice and legal security in the age of autonomous technologies..

Method

The research adopts a descriptive-analytical approach and utilizes a comparative study of advanced legal systems to analyze this challenge.

Results

The findings indicate that an effective solution requires moving beyond the "human or machine" dichotomy and adopting a multi-layered liability model. This model distributes responsibility across three distinct layers: the design and production layer (involving the moral and warranty liability of the designer), the operation and supervision layer (involving the precautionary and supervisory liability of the user), and the regulation and compensation layer (involving the warranty liability of the regulatory body and a compensation fund). This framework, while managing technological complexity and emphasizing ultimate human accountability, enables a fair analysis of incidents and facilitates the adaptation of self-defense and necessity criteria to algorithmic decisions.

Conclusions

This research was formed with the aim of answering the fundamental question: when faced with potentially harmful decisions by artificial intelligence systems in situations of self-defense and necessity, what is the appropriate legal and judicial framework for attributing criminal liability and ensuring justice? The findings revealed that traditional models of criminal liability, which are based on the triad of "material, mental, and personal elements" and center on the autonomous human agent, face theoretical and practical dead-ends when confronting the non-human, self-learning agency of AI. Alternative approaches are also individually incomplete: foreseeability-based liability grapples with the "black box" problem; strict (risk-based) liability may stifle innovation; and granting independent legal personality conflicts with the philosophical foundations of criminal law. Amidst this, the multi-layered liability model emerges as a comprehensive and practical solution. This model, by moving beyond the unproductive "human or machine" dichotomy, distributes responsibility along a value chain: from the moral and warranty liability of the designer and producer regarding the system's intrinsic safety and ethics-based training, to the supervisory and

precautionary liability of the operator for correct use and intervention in crisis situations, and finally to the warranty and compensatory liability of regulatory bodies through establishing transparent frameworks and compensation funds. This model not only can manage technological complexity but also, by emphasizing ultimate human accountability, maintains its compatibility with the fundamental principles of criminal law. Moving from theoretical deadlock to a practical solution based on multi-layered liability in facing algorithmic self-defense and necessity can clarify the future horizon of using these intelligent systems from the perspective of liability across criminal, technical, and responsibility aspects. The findings show that the core of traditional criminal liability theory revolves around concepts that lack objective counterparts in the world of self-learning algorithms. "Free will" as the basis for choice and "criminal intent" as the moral condition for punishment undergo a kind of analytical collapse when faced with systems that make decisions based on probabilistic calculations and optimization of objective functions. The key question is not whether AI has will, but rather: how can a system whose behavior is the product of a chain of predefined "if-then" statements or the output of a complex statistical model be the subject of moral blame and criminal punishment? In a concrete scenario, an automated defense system that identifies and neutralizes a range of targets based on patterns learned from historical data lacks the "intent to defend" in the human sense. This action is a calculated reaction, not an ethical choice. Therefore, attempting to impose traditional frameworks onto this novel phenomenon leads only to hypothetical and unrealistic attributions or a simple denial of the problem. This void makes the search for alternative paradigms of liability unavoidable. Examining each of the novel approaches within the realistic context of self-defense and necessity situations reveals their inherent limitations:

- Foreseeability-based Liability: In entirely novel and unprecedented emergency situations (such as a combined cyber-physical attack with an unknown pattern), the fundamental criterion of "foreseeability for a reasonable expert" itself becomes ambiguous. The boundary between an unforeseeable risk and a foreseeable error in such conditions is unclear.· Strict Liability (Risk-based): While effective in the fair distribution of the burden of compensation, it lacks the necessary nuance to distinguish between a criminal act and a justified act. This approach cannot differentiate between an error arising from a technical defect and a justified action involving unavoidable collateral damage (e.g., a self-driving car's decision to swerve onto a sidewalk to save the passengers' lives).
- Independent Legal Personality: This approach is entirely incapable of solving the riddle of the "competent and beneficiary of defense." A system lacks the instinct for self-preservation, personal interest, or the ability for ethical reasoning about its own actions. Therefore, how can one claim it engaged in "self-defense" or defense of another? This incapacity weakens the foundation of this theory in emergency situations. The multi-layered liability model, by abandoning the fruitless search for a "single culprit" and focusing on a "distributed network of causality," enables a more precise and fairer analysis of incidents. This model examines an event across three interconnected horizons:

1. The Design and Development Horizon (First Layer - Moral and Warranty Liability): Was the algorithm trained with biased or incomplete data? Were fundamental ethical values such as the priority of civilian lives or the principle of proportionality properly embedded in its decision-making logic? Accountability at this horizon lies with the manufacturer and developer.

2. The Operation and Supervision Horizon (Second Layer - Supervisory and Precautionary Liability): Was the operator or end-user adequately trained? Did they have the meaningful opportunity and authority to intervene and override the command at the critical moment? Were there clear operational protocols for unexpected situations? Responsibility at this horizon rests with the human operator and supervisor.

3. The Regulation and Compensation Horizon (Third Layer - Warranty and Compensatory Liability): Had an independent, specialized regulatory body evaluated and approved this system prior to deployment? Was a swift and fair compensation mechanism for victims, without the need to prove fault, foreseen? This responsibility falls upon the legislator and regulatory institutions.

Regarding the impact of multi-layered liability, in a hypothetical incident where an automated medical diagnosis and treatment system, during an emergency caused by a prolonged power outage, is forced to allocate a limited drug resource to one of two critically ill patients, the multi-layered model raises the following analytical questions:

At the design layer: On what basis (survival probability, age, social utility) was the objective function and allocation algorithm of this system configured? Had these bases been reviewed by medical ethics committees? Accountability at this layer is conditional on the transparency of the algorithm's ethical foundations (such as the allocation criteria) and obtaining approval from independent ethics committees. Lack of transparency or design without foreseeing an emergency stop mechanism entails the fault and civil/criminal liability of the manufacturer.

At the operation layer: Was the responsible nurse or doctor present on-site aware of the incident, and did they have clear instructions and legal authority to override the system's decision in exceptional cases? Liability at this level depends on adequate staff training, the existence of transparent protocols for overriding system decisions, and the real possibility of human intervention. If hospital management was negligent in training or developing protocols, it is the primary responsible party. Negligence by aware personnel also creates direct liability.

At the regulatory layer: Was the hospital obligated to have a backup power source with sufficient capacity? Had the health regulatory institution developed mandatory standards for autonomous medical systems in crisis conditions? The responsibility of regulatory institutions lies in developing mandatory standards (e.g., requiring backup power), granting conditional licenses after evaluation, and continuous post-deployment oversight. Negligence in these duties can establish the liability of the regulatory body.

Ultimately, it must be said that the AI challenge in criminal law is a structural and paradigmatic challenge, not a simple exceptional case. The multi-layered liability model, by accepting this reality, focuses on empowering and holding the entire value chain accountable rather than finding a single culprit. For Iran's legal system, adopting this model is not about choosing a foreign template, but an endogenous strategic necessity to enable it, while preserving its own dogmatic principles, to provide

a just, practical, and preventative response to one of the most complex legal issues of the present era. This analysis provides a solid theoretical foundation for the objective, layered policy proposals presented in the conclusion section. However, applying this framework within Iran's legal system, especially in the sensitive domains of self-defense and necessity, requires urgent legislative and jurisprudential rethinking. Articles 156 and 159 of the Islamic Penal Code and other related regulations were drafted based on the presumption of a "human agent," and there is a clear gap regarding autonomous algorithmic decisions. This gap can lead to injustice, legal insecurity, and evasion of responsibility.

The transition from a human-based society to a human-machine society is not possible by relying solely on yesterday's rules. This research showed that the future of criminal law depends on accepting flexible, distributed, and interdisciplinary models like multi-layered liability. Achieving this is not a choice, but an urgent necessity to protect human dignity, uphold the rule of law, and ensure national security in the age of autonomous technologies. The actions of the legislator and policymaker today will determine how history answers the question of "justice in the age of artificial intelligence."

Author Contributions

All authors contributed equally to the conceptualization of the article and writing of the original and subsequent drafts.

Data Availability Statement

Data available on request from the authors.

Acknowledgements

The authors would like to thank the anonymous reviewers for their insightful comments and constructive feedback, which significantly improved the quality of this manuscript. We also extend our gratitude to our colleagues for their valuable discussions and technical support throughout this research.

Ethical Considerations

The authors strictly adhered to the highest standards of research integrity. The authors avoided data fabrication, falsification, plagiarism, and any other form of scientific misconduct.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of Interest

The authors declare no conflict of interest.

پژوهش های نوین در مطالعات علوم انسانی اسلامی
شاپا الکترونیکی: ۸۹۴۴-۲۹۸۰
دانشگاه لرستان

شناسایی مسئولیت کیفری هوش مصنوعی بر اساس موقعیت‌های اضطراری یا دفاع مشروع

محمدعلی عباس‌زاده دیباور[۱]، ابراهیم رجبی تاج امیر[۲] ✉، سجاد اختری[۳]

۱. گروه حقوق جزا و جرم‌شناسی، واحد ساوه، دانشگاه آزاد اسلامی، ساوه، ایران. رایانامه: m.abbaszadehdibavar@iau.ir
۲. نویسنده مسئول، گروه جرم‌شناسی، دانشگاه علوم انتظامی امین، تهران، ایران. رایانامه: e.rajabi.t@gmail.com
۳. گروه حقوق کیفری و جرم‌شناسی، واحد کرج، دانشگاه آزاد اسلامی، کرج، ایران. رایانامه: Sajadakhtari2025@iau.ac.ir

اطلاعات مقاله

نوع مقاله:
مقاله پژوهشی،

تاریخچه مقاله:
تاریخ دریافت: ۱۴۰۴/۰۱/۱۱
تاریخ بازنگری: ۱۴۰۴/۰۴/۲۲
تاریخ پذیرش: ۱۴۰۴/۰۶/۰۵
تاریخ انتشار: ۱۴۰۴/۰۷/۰۶

کلیدواژه‌ها:
هوش مصنوعی،
مسئولیت کیفری،
دفاع مشروع،
حالت اضطرار،
مسئولیت چندلایه،
حقوق کیفری ایران

چکیده

هدف: پژوهش حاضر با هدف تبیین چارچوب مسئولیت کیفری در قبال تصمیمات سامانه‌های هوش مصنوعی خودمختار در موقعیت‌های حساس دفاع مشروع و حالت اضطرار انجام شده است. مسئله اصلی، ناکارآمدی الگوهای سنتی مسئولیت کیفری (مبتنی بر اراده و نیت انسانی) و نیز محدودیت رویکردهای نوینی چون مسئولیت مبتنی بر پیش‌بینی یا ریسک در مواجهه با پیچیدگی و خودآموزی این سامانه‌هاست.

روش پژوهش: این پژوهش با رویکرد توصیفی–تحلیلی و با بهره‌گیری از مطالعه تطبیقی در نظام‌های حقوقی پیشرو، به واکاوی این چالش پرداخته است.

یافته‌ها: یافته‌ها نشان می‌دهد که راه‌حل کارآمد، خروج از دوگانه «انسان یا ماشین» و اتخاذ الگوی مسئولیت چندلایه است. این مدل، مسئولیت را به‌صورت توزیع‌شده در سه لایه پی می‌گیرد: لایه طراحی و تولید (مسئولیت اخلاقی و تضمینی طراح)، لایه بهره‌برداری و نظارت (مسئولیت احتیاطی و نظارتی کاربر) و لایه تنظیم‌گری و جبران (مسئولیت تضمینی نهاد ناظر و صندوق جبران خسارت). این چارچوب ضمن مدیریت پیچیدگی فناورانه و تأکید بر پاسخگویی نهایی انسان، امکان تحلیل منصفانه حوادث و تطبیق معیارهای دفاع مشروع و اضطرار با تصمیمات الگوریتمی را فراهم می‌سازد.

نتیجه‌گیری: در بستر حقوق کیفری ایران، این پژوهش ضرورت گذار از «بهت تقنینی» موجود را گوشزد کرده و دو مسیر تفسیر پویا از نهادهای فعلی (مانند مسئولیت ناشی از تسبیب) و تقنین پیش‌نگر بر مبنای الگوی چندلایه را برای تضمین عدالت کیفری و امنیت حقوقی پیشنهاد می‌دهد. در نتیجه، آینده حقوق کیفری در گرو پذیرش چنین الگوهای انعطاف‌پذیر و توزیع‌شده‌ای است تا بتواند در عصر فناوری‌های خودمختار، از کرامت انسانی و حاکمیت قانون حراست نماید.

استناد: عباس‌زاده دیباور، محمدعلی؛ رجبی تاج امیر، ابراهیم و اختری، سجاد. (۱۴۰۴). شناسایی مسئولیت کیفری هوش مصنوعی بر اساس موقعیت‌های اضطراری یا دفاع مشروع. *پژوهش های نوین در مطالعات علوم انسانی اسلامی*، (۷) ۴، ۲۳-۱. https://doi.org/10.22034/api.2025.2081092.1606

ناشر: دانشگاه لرستان.

DOI: https://doi.org/10.22034/api.2025.2081092.1606

مقدمه

در گستره معرفت بشری، هر علمی با اتکا به پارادایم منحصربه‌فرد خود به کاوش در پدیده‌ها می‌پردازد. علم حقوق، به‌عنوان دانشی هنجاری و تجویزی، نه در پی توصیف صرف واقعیات، بلکه در جستجوی تنظیم روابط اجتماعی از رهگذر اصول «باید» و «نباید» است. این ماهیت، حقوق را در مواجهه با پدیده‌های نوین در موقعیتی ویژه قرار می‌دهد؛ از سویی برای فهم دقیق واقعیت‌های پیچیده، نیازمند بهره‌گیری از یافته‌های علوم تجربی و فنی است و از سوی دیگر، موظف است با معیارهای عدالت، انصاف و اخلاق، به ارزیابی و هنجارگذاری برای همان واقعیات بپردازد. ظهور فناوری «هوش مصنوعی[1]» به‌ویژه در شاخه یادگیری ماشین، این رابطه دیالکتیکی را به چالشی بی‌سابقه کشانده است. هوش مصنوعی با خلق سامانه‌هایی که قادر به یادگیری، تصمیم‌گیری و اقدام در محیط های پیچیده هستند، مفاهیم بنیادین حقوق کیفری را تحت تأثیر قرار داده است. ماهیت و پیامدهای مسئولیت کیفری در قبال اعمالی که ریشه در اراده و نیت یک الگوریتم دارد، به پرسشی محوری در ادبیات حقوقی معاصر تبدیل شده است.

این پیچیدگی زمانی مضاعف می‌شود که کاربرد این سامانه‌ها را در حساس‌ترین عرصه‌ها، یعنی موقعیت‌های اضطراری و شرایط دفاع مشروع، مد نظر قرار دهیم. در چنین بستری، سامانه‌های هوشمند ممکن است ناگزیر به اتخاذ تصمیماتی شوند که مستلزم سنجش فوری معیارهایی مانند ضرورت، تناسب و فوریت است؛ معیارهایی که در حقوق کیفری سنتی، بر اراده آگاهانه و قصد انسانی متکی بوده‌اند. افزون بر این، روش‌های متفاوت یادگیری ماشین، درجه پیش‌بینی‌پذیری و کنترل رفتار سامانه را به شدت تحت تأثیر قرار می‌دهند. در یادگیری تحت نظارت، رفتار خروجی تا حد زیادی قابل پیش‌بینی است، حال آنکه در یادگیری تقویتی یا یادگیری عمیق، سامانه ممکن است در تعامل پویا با محیط، رفتارهایی را فرا بگیرد یا الگوهایی را کشف کند که حتی برای طراحان آن نیز غیرمنتظره باشد. این پویایی ذاتی و ویژگی جعبه سیاه در برخی معماری‌های پیچیده، ردیابی و انتساب علیت را دشوار ساخته و مبانی سنتی مسئولیت را با ابهام روبه‌رو می‌کند.

ادبیات پژوهشی موجود، گواه توجه جدی محققان به ابعاد گوناگون این چالش است. پژوهش‌های داخلی، از یک سو بر کارکردهای راهبردی و دفاعی هوش مصنوعی (مانند مصطفی السان و دهستانی،۱۴۰۱؛ رستمی، ۱۴۰۱؛ قاسمی و همکاران، ۱۴۰۲؛ ترکی، ۱۴۰۳و حسین زاده و همکاران، ۱۴۰۳) و از سوی دیگر بر مبانی اولیه مسئولیت مدنی و کیفری آن(مانند نوری‌سوا، ۱۴۰۴؛ ذاکری‌نیا، ۱۴۰۲) تمرکز داشته‌اند. در عرصه بین‌المللی نیز مباحث از تمرکز بر مشروعیت استفاده از سیستم‌های خودمختار در درگیری‌های مسلحانه (مانند گزارش تایم[2]، ۲۰۲۴؛ مجله اکسویس[3]،۲۰۲۴؛ آسوشیتدپرس[4]،۲۰۲۳) در زمینه هشداردهی سلاح‌های خودکار به همراه تحلیل‌های مؤسسه لیبر[5]، ۲۰۲۵) تا بررسی امکان اعطای شخصیت حقوقی به هوش مصنوعی (همچون پژوهش حموری[6] و الکدی[7]، ۲۰۲۴) و تحلیل جرایم نوپدید ناشی از آن (همچون مطالعات پاناتونی[8]، ۲۰۲۵؛ عبدالعزیز[9]، ۲۰۲۵) گسترده است؛ رویکرد مشابهی نیز در مطالعات تطبیقی (آمریکا، آلمان واتحادیه اروپا[10]) وجود دارد. با این وجود، خلاء آشکاری در پیوند دادن این دو جریان پژوهشی مشاهده می‌شود: تدوین یک چارچوب منسجم برای انتساب مسئولیت کیفری، با درنظرگیری شرایط خاص موجهه جرم(مانند دفاع مشروع و اضطرار) در تصمیمات خودمختار الگوریتمی. پرسش اصلی این است

1. Artificial Intelligence:AI
2. The TIME Journal
3. Axios
4. Associated Press
5. Lieber Institute
6. Hammouri
7. El-Kady
8. Panattoni
9. Abdelaziz

۱۰. در سطح بین‌المللی نیز این بحث در جریان است. پارلمان اروپا در سال ۲۰۲۱ طی قطعنامه‌ای ضرورت تصویب مقررات کیفری خاص در زمینه هوش مصنوعی را مورد تأکید قرار داد و شورای اروپا در سال ۲۰۲۲ دستورالعمل‌هایی در ارتباط با رعایت حقوق بشر در بهره‌گیری از این فناوری صادر کرد (European Parliament, 2021; Council of Europe, 2022).

که چگونه می توان در پرتو قواعد سنتی مسئولیت کیفری که بر سه رکن عنصر مادی، عنصر روانی(نیت) و قابلیت انتساب به شخص استوار است، به ارزیابی عملی یک سامانه هوشمند پرداخت که فاقد «اراده» و «قصد» انسانی است؟ آیا می‌توان مفاهیمی مانند دفاع مشروع را به صورت مجازی به رفتار سامانه نسبت داد و مسئولیت نهایی را بر داشت انسان‌های مرتبط (طراح، بهره‌بردار، ناظر) گذاشت، یا باید به سوی الگوهای کاملاً جدیدی از مسئولیت چندلایه یا مسئولیت مبتنی بر ریسک حرکت کرد؟

این مقاله با پذیرش این چالش، در پی آن است تا با رویکردی توصیفی–تحلیلی و با بهره‌گیری از مطالعات تطبیقی، امکان‌سنجی و چگونگی انتساب مسئولیت کیفری به تصمیمات سیستم‌های هوش مصنوعی در موقعیت‌های اضطرار و دفاع مشروع را بررسی کند. فرضیه محوری این است که نظام‌های حقوقی سنتی به تنهایی قادر به پاسخگویی به این پیچیدگی نیستند و اتخاذ یک الگوی توزیع شده و چندسطحی از مسئولیت، که در آن پاسخگویی نهایی انسان محور باقی می‌ماند، راهکار مناسبی برای تضمین عدالت و امنیت حقوقی در عصر فناوری‌های خودمختار به شمار می‌آید.

اهمیت و ضرورت پژوهش

سرعت تحولات هوش مصنوعی، پدیده‌ای صرفاً فنی یا اقتصادی نیست؛ این فناوری بنیان‌های مفهومی علوم انسانی و به‌طور خاص حقوق کیفری را با بازتعریف روبه‌رو ساخته است. پارادایم کلاسیک حقوق جزا، همواره «انسان» را به‌عنوان فاعل مختار، آگاه و صاحب اراده، محور مسئولیت کیفری می‌دانسته است. با این حال، پیدایش سامانه‌های هوشمند خودمختار که توانایی اتخاذ تصمیمات مؤثر و گاهی برگشت‌ناپذیر را بدون مداخله آنی انسان دارا هستند، این اصل بنیادین را به چالش کشیده است. این چالش تنها به بعد نظری محدود نمی‌ماند و در صحنه عملی، هنگام وقوع حوادثی با پیامدهای کیفری، خود را نمایان می‌سازد.

از منظر نظری، این پژوهش به چند پرسش بنیادین پاسخ می‌گوید: آیا می‌توان مفاهیمی چون «عنصر مادی جرم» و «عنصر روانی» (قصد و خطای کیفری) را به عملکرد یک الگوریتم که فاقد شعور و نیات انسانی است، تسری داد؟ چگونه می‌توان شرایط موجهه جرم، مانند دفاع مشروع و حالت اضطرار، را که مبتنی بر ارزیابی انسانی از خطر و تناسب اقدام است، به تصمیم‌گیری‌های خودکار یک سامانه نسبت داد؟ پاسخ به این پرسش‌ها مستلزم بازخوانی و احتمالاً بازسازی بخشی از نظریه مسئولیت کیفری در مواجهه با عاملیت غیرانسانی است.

از منظر عملی و کاربردی، ضرورت این پژوهش پررنگ‌تر می‌شود. سیستم‌های خودران، ربات‌های پزشکی، سامانه‌های تسلیحاتی خودکار و نرم‌افزارهای تصمیم‌ساز قضایی، همگی از جمله مصادیقی هستند که هماکنون در حال ورود به عرصه اجتماع هستند. فقدان چارچوب حقوقی روشن در مواجهه با خطاها یا تبعات زیان‌بار عملکرد این سیستم‌ها می‌تواند به بروز هرج و مرز قضایی، نقض حقوق بزه‌دیدگان و ایجاد مانعی در مسیر توسعه ایمن فناوری بینجامد. بنابراین، این پژوهش با هدف ارائه الگویی منسجم برای انتساب مسئولیت، می‌تواند نقشه راهی برای نهادهای زیر فراهم آورد:

برای قانون‌گذاران و سیاست‌گذاران: ارائه مبانی علمی جهت تدوین قوانین و مقررات پیش‌بین، منصفانه و کارآمد در حوزه فناوری‌های خودمختار، به نحوی که هم ازاین فناوری حمایت کند و هم از حقوق شهروندان حراست نماید.

برای مراجع قضایی و دادگاه‌ها: ارائه معیارها و چارچوب‌های تحلیل روشن جهت اتحاد رویه قضایی و صدور آراء مستدل و قابل پیش‌بینی در پرونده‌های پیچیده مرتبط با هوش مصنوعی، که به تضمین عدالت کیفری و امنیت حقوقی می‌انجامد.

برای توسعه‌دهندگان، تولیدکنندگان و بهره‌برداران فناوری: روشن‌سازی حدود و ثغور مسئولیت حقوقی، انگیزه‌بخشی برای طراحی مسئولانه‌تر، افزایش استانداردهای ایمنی و سرمایه‌گذاری در مکانیزم‌های نظارت و کنترل از همان مرحله طراحی.

در مجموع، این پژوهش نه تنها گامی ضروری در غنای دانش حقوقی تطبیقی و میان‌رشته‌ای محسوب می‌شود، بلکه پاسخی عملی به یک نیاز فوری اجتماعی است؛ بنابراین با توجه به این جنبه‌ها چگونگی حکمرانی، کنترل و پاسخگویی در عصر فناوری‌هایی که به شکلی روزافزون از ظرفیت توسعه شتابانی برخوردارند؛ مؤید اهمیت وضرورت مطالعه است.

پیشینه و چارچوب نظری پژوهش

تحلیل و دسته‌بندی پیشینه پژوهشی

بررسی نظام‌مند آثار منتشرشده در حوزه هوش مصنوعی و حقوق، نشان‌دهنده تمرکز ادبیات موجود بر دو جریان اصلی و در عین حال کم‌تر پیوندخورده است. شناسایی این دو جریان و خلاء موجود در میان آن‌ها، نقطه عزیمت ضروری برای تبیین چارچوب نظری این پژوهش محسوب می‌شود.

جریان اول: پژوهش‌های راهبردی–دفاعی و آینده‌نگر

بخش قابل توجهی از تحقیقات داخلی، با نگاهی کارکردی و مدیریتی به تحلیل نقش هوش مصنوعی به عنوان یک قدرت افزاینده ملی و یک ابزار راهبردی پرداخته‌اند. برای نمونه، رستمی (۱۴۰۱) و قاسمی و همکاران (۱۴۰۲) بر شناسایی و بومی‌سازی ظرفیت‌های کاربردی هوش مصنوعی در سازمان‌های نظامی و فرآیندهای آینده‌نگاری دفاعی تأکید داشته‌اند. در امتداد همین نگاه، ترکی (۱۴۰۳) هوش مصنوعی را به مثابه یک «قوت ژئوپلیتیک» در رقابت میان ابرقدرت‌ها تحلیل کرده و حسین‌زاده و همکاران (۱۴۰۳) لزوم به‌کارگیری سامانه‌های هوشمند را برای مقابله با تهدیدات نوین مانند پهپادها نشان داده‌اند. نقطه قوت این دسته از پژوهش‌ها، توجه به ضرورت‌های عملیاتی و سیاست‌گذاری کلان است، اما عموماً از تحلیل عمیق مبانی حقوقی مشروعیت‌بخش به این کاربردها، به ویژه در قالب نهادهای کیفری مانند دفاع مشروع، غفلت ورزیده‌اند.

جریان دوم: پژوهش‌های حقوقی مسئولیت‌محور

در مقابل، جریانی از تحقیقات حقوقی صرفاً بر روی مبانی انتساب مسئولیت به هوش مصنوعی متمرکز شده‌اند. در سطح داخلی، نوری‌سوا (۱۴۰۴) امکان‌سنجی مسئولیت کیفری هوش مصنوعی با تکیه بر مفهوم اراده مستقل را بررسی کرده، اما آن را در بستر عینی موقعیت‌های توجیه‌کننده جرم مانند دفاع مشروع به محک آزمون نگذاشته است. ذاکری‌نیا (۱۴۰۲) نیز ضمن بررسی تطبیقی مسئولیت مدنی، بر امکان شکل‌گیری قواعد مشترک تأکید نموده است. پژوهش‌هایی مانند السان و دهستانی (۱۴۰۱) در حوزه جعل عمیق و بنافی (۱۴۰۲) در زمینه حریم خصوصی نظامی، هر یک گوشه‌ای از چالش‌های حقوقی نوظهور را روشن ساخته‌اند، بی‌آنکه راهکار جامعی برای انتساب مسئولیت در شرایط پیچیده اضطراری ارائه دهند.

در عرصه پژوهش‌های خارجی، این جریان به دو شاخه تقسیم می‌شود:

۱. شاخه حقوق بین‌الملل و اخلاق کاربردی: این شاخه با تحلیل موارد عینی، مشروعیت استفاده از هوش مصنوعی در درگیری‌های مسلحانه را می‌سنجد. گزارش‌هایی مانند مجله تایم (۲۰۲۴) از جنگ غزه، آسوشیتدپرس[1] (۲۰۲۳) از اوکراین، و تحلیل‌های پروفسور جورج‌تاون در مؤسسه لیبر (۲۰۲۵)، پروفسور اشلی‌دیکس[2] (۲۰۲۴) و بیزینس اینسایدر[3] (۲۰۲۴) همگی بر دغدغه اخلاقی و ابهام حقوقی حاصل از فقدان نظارت انسانی مستقیم و دشواری تطبیق قواعدی مانند تناسب و ضرورت تأکید دارند.

۲. شاخه حقوق کیفری محض: این شاخه به طور خاص به معمای مسئولیت کیفری می‌پردازد. پاناتونی (۲۰۲۵) و عبدالعزیز (۲۰۲۵) بر دشواری‌های ناشی از «نبود عامل انسانی» در ارتکاب جرم تمرکز کرده و لزوم بازنگری در قوانین را خاطرنشان ساخته‌اند. حموری (۲۰۲۴) و الکدی (۲۰۲۴) نیز با بررسی امکان اعطای «شخصیت حقوقی» به هوش مصنوعی، بحث را به سطحی بنیادین‌تر کشانده‌اند.

با نگاهی کل‌نگر به پیشینه فوق، روشن می‌شود که یک فاصله نظری و عملی چشمگیر بین دو جریان اصلی وجود دارد. پژوهش‌های راهبردی–دفاعی هوش مصنوعی را به عنوان یک ابزار قدرتمند عمل می‌بینند، ولی از پرسش‌های حقوقی بنیادین

1. Associated Press
2. Professor Ashley Deeks
3. Business Insider

درباره مشروعیت و پاسخگویی این عمل غافل‌اند. از سوی دیگر، پژوهش‌های حقوقی محض، اغلب در سطح کلیات انتزاعی مسئولیت یا تحلیل مصادیق مجرمانه خاص باقی مانده و کمتر به اقتضائات عملیاتی و زمانی فشرده موقعیت‌های اضطراری و دفاعی پرداخته‌اند.

برای پر کردن این خلاء، پژوهش حاضر بر ترکیب دو مفهوم کلیدی تأکید دارد: مسئولیت چندلایه[1]و حکمرانی مسئولانه هوش مصنوعی[2] مسئولیت چندلایه، که در تحقیقات تطبیقی و مقررات اتحادیه اروپا مورد تأکید است، پیشنهاد می‌کند که مسئولیت عملکرد سامانه‌های خودمختار در یک زنجیره ارزشی توزیع شود: طراح و برنامه‌نویس، بهره‌بردار و ناظر، و نهایتاً نهادهای تنظیم‌گر و قانون‌گذار(رودریگوئز و همکاران[3]، ۲۰۲۵). این رویکرد امکان تحلیل تصمیمات خودمختار و انتساب مسئولیت انسانی در عین حفظ اصول بنیادین حقوق کیفری، مانند ضرورت عنصر روانی و شخصی بودن مجازات، را فراهم می‌آورد.

همزمان، حکمرانی مسئولانه بر ایجاد چارچوب‌های پیشینی، مبتنی بر اصول عدالت، شفافیت و پاسخگویی، تأکید دارد و نشان می‌دهد که پاسخ به چالش‌های ناشی از سامانه‌های خودمختار تنها با مدل‌های پسینی مسئولیت[4]ممکن نیست(پاپاگیانیدیس و همکاران[5]، ۲۰۲۵) تنظیم مقررات، نهادهای نظارتی و مکانیزم‌های کنترل فنی در کنار یکدیگر عمل می‌کنند تا خطاها و تبعات عملکرد سامانه‌ها در شرایط اضطراری یا دفاع مشروع مدیریت شود.

ترکیب این دو رویکرد، پایه‌ای برای یک الگوی یکپارچه و میان‌رشته‌ای فراهم می‌آورد که هم عملگرایی و هم انطباق‌پذیری حقوقی دارد. این چارچوب، زمینهٔ ارزیابی رفتار سامانه‌های هوش مصنوعی در شرایط اضطراری و دفاع مشروع را فراهم کرده و خلاء موجود میان رویکردهای دفاعی–راهبردی و حقوقی مسئولیت‌محور را پر می‌سازد. به این ترتیب، پژوهش حاضر با بهره‌گیری از مسئولیت چندلایه و حکمرانی مسئولانه، به تکمیل چالش نظری و عملی حکمرانی بر هوش مصنوعی در حساس‌ترین موقعیت‌ها می‌پردازد.

فرج‌پور (۱۴۰۴) در مقاله‌ای با عنوان تحلیل حقوقی و تطبیقی مسئولیت مدنی هوش مصنوعی در تصمیم‌گیری خودکار، منتشر در مجله هوش مصنوعی و فناوری در علوم رفتاری و اجتماعی، با روش تحلیلی–توصیفی و تطبیقی، چالش‌های مسئولیت مدنی ناشی از تصمیم‌گیری خودکار هوش مصنوعی را بررسی می‌کند. نویسنده رویکردهای ایالات متحده (مسئولیت محصول و جانشینی)، اتحادیه اروپا (مسئولیت سختگیرانه و بیمه اجباری) و ایران (عدم شخصیت حقوقی مستقل برای AI و ابهامات حقوقی) را مقایسه کرده و مدل‌هایی مانند مسئولیت توسعه‌دهنده، اپراتور، مسئولیت سختگیرانه و بیمه را تحلیل می‌نماید. پیشنهاد اصلی، تدوین مقررات جدید همسو با استانداردهای بین‌المللی برای رفع چالش‌های اجرایی در ایران است. این منبع در تمایز مسئولیت مدنی از کیفری و تحلیل تطبیقی رویکردها ارزشمند است. فرج‌پور و همکاران (۱۴۰۴). در مقاله‌ای با عنوان نقش هوش مصنوعی در داوری ورزشی و چالش‌های حقوقی ناشی از تصمیمات خودکار، منتشر در مجله هوش مصنوعی و فناوری در علوم رفتاری و اجتماعی، کاربرد AI در داوری ورزشی (مانند VAR و Hawk-Eye) را واکاوی کرده و چالش‌های حقوقی آن مانند مسئولیت در خطاهای الگوریتمی، عدم شفافیت (جعبه سیاه) و ناتوانی در ارزیابی عوامل انسانی را برجسته می‌سازد. نویسندگان پیشنهاد چارچوب‌های نظارتی برای پاسخگویی و عدالت در تصمیم‌گیری‌های خودمختار ارائه می‌دهند. این منبع در بحث چالش‌های عملی تصمیمات حساس AI و مسائل اخلاقی–حقوقی مفید است. فرج‌پور (۱۴۰۴). در مقاله‌ای دیگر با عنوان نقش مسئولیت مدنی در قوانین هوش مصنوعی از منظر نظام‌های حقوقی عمده جهانی، منتشر در مجله مطالعات حقوق عمومی، با رویکرد تحلیلی–استنتاجی و تطبیقی، ناکارآمدی مکانیسم‌های سنتی مبتنی بر تقصیر انسانی در برابر استقلال و تاریک‌خانه‌ای (black box) هوش مصنوعی را واکاوی می‌کند. نویسنده نظام‌های حقوقی آلمان، ایالات متحده، چین، ژاپن،

1. Multi-layered liability
2. Responsible AI Governance
3. Rodríguez et al.
4. Ex Post Liability
5. Papagiannidis et al.

کانادا، اتحادیه اروپا و ایران را مقایسه کرده و نشان می‌دهد که در هیچ‌کدام AI شخصیت حقوقی مستقل ندارد و مسئولیت به توسعه‌دهندگان یا بهره‌برداران منتسب می‌شود. پیشنهاد اصلی، اتخاذ مسئولیت مطلق برای سیستم‌های پرریسک و بیمه اجباری مدنی در ایران، همسو با استانداردهای بین‌المللی است. این منبع در تقویت مباحث تطبیقی مسئولیت مدنی و پیشنهادهای تقنینی برای ایران ارزشمند است.

چارچوب نظری پژوهش

چهارچوب نظری این پژوهش در چهار گام به هم پیوسته تنظیم شده است. نخست، مبانی کلاسیک مسئولیت کیفری و تنگناهای آن در مواجهه با فناوری هوش مصنوعی تشریح می‌شود. سپس، طیف رویکردهای نوین پیشنهادی در ادبیات برای عبور از این تنگناها، نقد و تحلیل می‌گردد. در گام نهایی، چالش‌های عملی تطبیق این رویکردها همراه با شفاف‌سازی مرز مسئولیت کیفری و مدنی در چارچوب هوش مصنوعی با موقعیت‌های ویژه دفاع مشروع و اضطرار، و نیز با نظام حقوقی ایران مورد بحث قرار می‌گیرد.

مبانی کلاسیک مسئولیت کیفری و بن‌بست نظری در عصر هوش مصنوعی

پارادایم مسلط در حقوق کیفری، مبتنی بر یک انسان‌محوری اخلاقی است که مسئولیت را به موجودی مختار، عاقل و صاحب اراده و نیت گره می‌زند. این پارادیم بر سه رکن استوار است:

عنصر مادی (رفتار مجرمانه): این رکن به فعل یا ترک فعل ملموس و مخاطره‌آمیز اشاره دارد. در مورد هوش مصنوعی، این پرسش مطرح است که آیا خروجی یک سامانه (مانند یک مانور رانندگی یا پرتاب موشک) را می‌توان «فعل» مجرمانه دانست یا صرفاً یک «نتیجه محصولی».

عنصر روانی (نیّت و تقصیر): به عنوان «روح جرم»، شرط اخلاقی مجازات است. این عنصر طیفی از قصد مستقیم تا بی‌احتیاطی را دربر می‌گیرد و بر پایه آگاهی و انتخاب انسانی بنا شده است. چالش محوری اینجاست که چگونه می‌توان حالت ذهنی سرزنش‌پذیر را به الگوریتمی که فاقد شعور، آگاهی و ادراک انسانی است، منتسب کرد؟ فقدان پاسخ به این سوال، اساس نظریه سنتی مسئولیت را در مورد هوش مصنوعی متزلزل می‌سازد.

عنصر شخصی (قابلیت انتساب): این رکن تأکید می‌کند که تنها فاعلی می‌تواند مسئول شناخته شود که توانایی تمیز خوب از بد و انتخاب آزادانه را داشته باشد. اصل «لا عقاب بلا انتساب» سنگ بنای این دیدگاه است. هوش مصنوعی، به دلیل نداشتن اراده آزاد و مسئولیت‌پذیری اخلاقی، در این چارچوب نمی‌گنجد.

حقوق کیفری کلاسیک، در یک سه‌گانه پارادوکسیکال در برابر هوش مصنوعی قرار می‌گیرد: هوش مصنوعی قادر به انجام «فعل مادی» زیان‌بار است، اما فاقد «عنصر روانی» لازم برای سرزنش اخلاقی است و در نتیجه، «قابلیت انتساب» مستقیم کیفری به آن ناممکن به نظر می‌رسد. این بن‌بست نظری، ضرورت جست‌وجو برای الگوهای جایگزین را ایجاب می‌کند.

رویکردهای نوین به مسئولیت: از انسان‌محوری تا سامانه‌محوری

برای خروج از این بن‌بست، چهارچوب‌های نظری متنوعی پیشنهاد شده‌اند که بر یک طیف از تمرکز کامل بر انسان تا پذیرش نسبی عاملیت سامانه قرار می‌گیرند.

رویکردهای انسان‌محور و مبتنی بر ریسک

این دسته، مسئولیت نهایی را همواره بر دوش انسان می‌گذارد.

مسئولیت مبتنی بر تقصیر: با استناد به قواعد عمومی، تقصیر طراح، برنامه‌نویس یا کاربر در طراحی ناقص، آموزش نادرست یا نظارت غفلت‌آمیز بررسی می‌شود.

مسئولیت مبتنی بر ریسک (مسئولیت محض): این نظریه، با الهام از مسئولیت دارندگان اشیاء یا فعالیت‌های پرخطر، استدلال می‌کند که بهره‌برداری از فناوری پیشرفته و ذاتاً غیرقابل پیش‌بینی هوش مصنوعی، به خودی‌خود بار مسئولیت را بر عهده بهره‌بردار یا تولیدکننده می‌گذارد، حتی در غیاب تقصیر. مکمل عملی این دیدگاه، ایجاد نظام بیمه اجباری برای پوشش خسارات است.

رویکردهای کارکردگرا (جایگزینی معیارهای عینی)

این رویکردها به جای جست‌وجوی نیت درونی، بر معیارهای بیرونی و عینی تمرکز می‌کنند.

مسئولیت مبتنی بر پیش‌بینی‌پذیری: معیار مسئولیت، امکان پیش‌بینی پیامدهای زیان‌بار توسط یک متخصص معقول در زمان طراحی یا استقرار سیستم است. این معیار شبیه استاندارد «بی‌احتیاطی» در حقوق سنتی است.

مسئولیت مبتنی بر شفافیت: اگر عملکرد سامانه به گونه‌ای باشد که توجیه تصمیم آن برای انسان قابل درک نباشد (پدیده جعبه سیاه)، این عدم شفافیت به نفع افزایش بار مسئولیت طراح یا بهره‌بردار تفسیر می‌شود.

رویکردهای سامانه‌محور (اعطای شخصیت حقوقی)

این نگاه رادیکال‌تر، درجاتی از عاملیت مستقل را برای هوش مصنوعی به رسمیت می‌شناسد.

مسئولیت نمادین یا صوری[1]: با اعطای «شخصیت الکترونیکی»، سامانه می‌تواند به صورت نمادین مسئول شناخته شود. هدف اصلی، بازدارندگی عمومی و تسهیل جبران خسارت از طریق صندوق‌های خاص است، نه سرزنش اخلاقی.

مسئولیت کیفری محدود: مشابه شخصیت حقوقی شرکت‌ها، این رویکرد قائل به دارایی‌ها و پاسخگویی‌های جداگانه برای سامانه‌های پیشرفته است.

دیدگاه‌های تطبیقی نشان می دهد که دراتحادیه اروپا: گزارش ۲۰۱۷ کمیسیون حقوقی پارلمان اروپا ایده «شخصیت الکترونیکی» را مطرح کرد. هرچند این پیشنهاد به‌شدت مورد انتقاد قرار گرفت، اما نشان می‌دهد نظام‌های حقوقی آماده‌اند حتی به‌طور نمادین، مسئولیت مستقلی برای هوش مصنوعی تعریف کنند؛ برخی نظریه‌پردازان بر این باورند که مسئولیت هوش مصنوعی باید بیشتر جنبه کارکردی داشته باشد و به‌عنوان یک شخصیت حقوقی محدود، صرفاً به دلایل عملی و حمایتی پذیرفته شود(هلگیندرفورد[2]، ۲۰۱۸)؛ مزیت مهم این دیدگاه تقویت اعتماد عمومی و تسهیل جبران خسارت بوده ولی صوری‌گرایی و تعارض با اصل شخصی بودن مسئولیت کیفری در این نوع از مسئولیت همچنان مغفول می‌ماند.

الگوی تلفیقی پیشنهادی: مسئولیت چندلایه

از نقد نقاط ضعف و قوت رویکردهای فوق، الگوی مسئولیت چندلایه به عنوان چارچوب جامع این پژوهش سر برمی‌آورد. این الگو با رد دوگانه انحصاری «یا انسان یا ماشین»، مسئولیت را در یک زنجیره ارزشی توزیع‌شده میان کلیه ذی‌نفعان می‌بیند:

۱. لایه طراحی و تولید: مسئولیت نهادهای توسعه‌دهنده در قبال امنیت ذاتی، آموزش اخلاق‌مدار و شفافیت الگوریتم.

۲. لایه نظارت و بهره‌برداری: مسئولیت کاربر، اپراتور یا نهاد ناظر در قبال استفاده مطابق با دستورالعمل و مداخله در شرایط اضطراری.

۳. لایه تنظیم‌گری و جبران: مسئولیت قانون‌گذار و نهادهای تنظیم‌گر در ایجاد چارچوب‌های شفاف، نظام‌های بیمه و صندوق‌های جبران خسارت.

این مدل، با توزیع مسئولیت، هم از «فرار از پاسخگویی» جلوگیری می‌کند، هم انگیزه نوآوری ایمن را حفظ می‌نماید و هم مسیر جبران را برای بزه‌دیده هموار می‌سازد.

1. Functional / Symbolic Liability
2. Hilgendorf

به این ترتیب، مسئولیت کیفری دیگر یک خط مستقیم از «عامل» به «نتیجه» نیست، بلکه مجموعه‌ای از زنجیره‌های مسئولیت است که در کنار هم چارچوب پاسخگویی را شکل می‌دهند.

انعکاس واقعیت فناورانه بر اساس این دیدگاه، رفتار مجرمانه ناشی از هوش مصنوعی معمولاً محصول یک تصمیم فردی ساده نمی‌داند، بلکه حاصل تعامل کدها، داده‌ها، کاربر و محیط است. مسئولیت چندلایه این پیچیدگی را منعکس می‌کند؛ اما بهره برداری از این دیدگاه سبب جلوگیری از فرار از پاسخگویی مسئولیت می‌شود به این تریب که، اگر تنها یکی از عوامل مسئول شناخته شود، دیگران می‌توانند از مسئولیت شانه خالی کنند. تقسیم مسئولیت میان لایه‌ها مانع این امر می‌شود؛ بنابراین، قربانی به جای گیر کردن در دعوای تقصیر یا اثبات نیت، می‌تواند از مجموعه مسئولان جبران خسارت دریافت کند؛ خصوصیت بسیار مهم این بینش توازن میان نوآوری و عدالت کیفری است که، با تقسیم مسئولیت، فشار مطلق بر یک گروه (مثلاً طراحان یا کاربران) برداشته می‌شود و در نتیجه هم توسعه فناوری و هم حمایت حقوقی تقویت می‌گردد.

بر همین اساس اتحادیه اروپا، در پیش‌نویس‌های تقنینی مربوط به هوش مصنوعی (AIAct)، مسئولیت میان سازنده، واردکننده، توزیع کننده و کاربر تقسیم شده است؛ این نمونه بارز رویکرد چندلایه است.

بر مبنای این نوع از مسئولیت حقوق آمریکا، در پرونده‌های مرتبط با خودروهای خودران، دادگاه‌ها گاه شرکت سازنده، گاه کاربر، و گاه هر دو را به درجات مختلف مسئول شناخته‌اند، و در حقوق آلمان نظریه‌پردازان پیشنهاد کرده‌اند که هوش مصنوعی به‌طور نمادین در لایه سوم مسئول باشد تا نظام پاسخگویی از «تک‌عاملی» به «چندعاملی» تغییر کند؛ مزیت واقع‌گرایی، حمایت گسترده از بزه‌دیدگان، و ایجاد توازن میان نوآوری و عدالت کیفری مواردی است که این نظریه به همراه دارد اماچالش پیچیدگی در اجرا، نیازمند قوانین دقیق و نهادهای تخصصی، است.

چالش‌های عملی تطبیق در موقعیت‌های دفاع مشروع و اضطرار

کاربست هر یک از رویکردهای فوق در موقعیت‌های حساس دفاعی و اضطراری، با چالش‌های عینی ویژه‌ای روبروست که این پژوهش به آن می‌پردازد.

چالش در انطباق عوامل موجهه جرم

در بسیاری از نظام‌های کیفری از جمله ایران، دفاع مشروع و اضطرار تنها در صورتی پذیرفته می‌شوند که شرایطی همچون ضرورت، تناسب و فوریت احراز شوند. اما پرسش مهم آن است که آیا این معیارها را می‌توان در تصمیم‌های الگوریتمی نیز ارزیابی کرد؟یک خودروی خودران ممکن است میان دو خطر، کم‌خطرتر را انتخاب کند؛ اما آیا این انتخاب دقیقاً معادل مفهوم حقوقی «تناسب» است؟سامانه دفاعی خودکار ممکن است در برابر تهدید فوری واکنش نشان دهد؛ ولی آیا می‌توان آن را واجد معیار «فوریت» و «ضرورت» دانست؟ برخی پژوهشگران معتقدند چون این واکنش‌ها مبتنی بر محاسبه الگوریتمی است و نه بر اراده انسانی، امکان انطباق کامل با مفاهیم سنتی وجود ندارد(پاگالو[1]، ۲۰۱۳).

بنابراین پرسش این است که چگونه می‌توان این معیارهای مبتنی بر قضاوت انسانی آنی را به تصمیم ازپیش‌برنامه‌ریزی‌شده یا خودآموخته یک سامانه نسبت داد؟ آیا محاسبه سرد الگوریتمی می‌تواند جایگزین ارزیابی عاطفی و اخلاقی انسان از «تناسب» شود؟

که در پاسخ، سه رویکرد شکل گرفته است:

بر اساس نظریه شخصیت مستقل، هوش مصنوعی را می‌توان مشابه اشخاص حقوقی دارای شخصیت مستقل دانست و بدین‌وسیله عنصر «انتساب شخصی» را از طریق یک اعتبار حقوقی بازسازی کرد؛ مبنای فلسفی این با پذیرش این فرض که نهادهای غیرانسانی نیز می‌توانند به‌طور قراردادی یا اعتباری «فاعل حقوقی» محسوب شوند (مانند شرکت‌ها).

1. Pagallo

در نگاه مسئولیت محور، این رویکرد عنصر شخصی را به افراد انسانی پیرامون هوش مصنوعی منتقل می‌کند؛ طراحان، کاربران یا نهادهای بهره‌بردار. در اینجا عنصر روانی و اراده انسانی جایگزین فقدان قصد و اراده در هوش مصنوعی می‌شود؛ مبنای فلسفی این رویکرد را باید در مسئولیت بر مبنای «قابلیت پیش‌بینی و کنترل انسانی» و اصل سرزنش‌پذیری انسان مختار هموار نمود؛ اما، گاه ممکن است رابطه میان رفتار سامانه و اراده انسانی آن‌قدر دور باشد که انتساب اخلاقی و سرزنش حقوقی را دشوار می‌سازد.

همچنین در رویکرد انتساب غیرمستقیم، در این نظریه، قانون‌گذار به‌طور اعتباری قصد و اراده را به سامانه نسبت می‌دهد، اما مسئولیت نهایی به اشخاص انسانی بازگردانده می‌شود. در واقع، عنصر روانی جعل و سپس بر انسان‌های مرتبط تحمیل می‌شود. دیدگاه مسئولیت غیرمستقیم، که پیامدهای کیفری را نهایتاً به مالک یا طراح منتسب می‌سازد؛ دیدگاه شخصیت حقوقی محدود، که قائل به اعطای نوعی شخصیت حقوقی خاص به سامانه‌های پیشرفته و امکان مسئولیت مستقل آن‌هاست(برتولینی و اپیسکوپو[1]، ۲۰۲۰).

مبنای فلسفی این دیدگاه، استفاده از ابزار «انتساب فرضی» برای پر کردن خلأ ناشی از فقدان نیت واقعی است؛ اما این مدل نیز از حیث فلسفی و اخلاقی ضعیف است؛ زیرا در واقعیت، قصد و نیتی در سامانه وجود ندارد، بلکه صرفاً یک ساختار حقوقی اعتباری ساخته می‌شود.

با توجه به این مساله که سامانه هوش مصنوعی فاقد قصد و آگاهی انسانی است، بنابراین امکان انتساب «عنصر معنوی جرم» به آن وجود ندارد(لایگنیا و سارتور[2]، ۲۰۲۰)؛ مالک یا کاربر ممکن است در زمان وقوع حادثه هیچ دخالت مستقیمی نداشته باشد، خصوصاً در مواردی که تصمیم توسط الگوریتم به‌صورت مستقل اتخاذ شده است؛ طراح یا برنامه‌نویس نیز در همه موارد نمی‌تواند مسئول شناخته شود، چرا که بخشی از تصمیم‌ها حاصل فرایند یادگیری و تعامل سیستم با محیط است(برایسون[3]، ۲۰۱۹).

با تمامی مفروضات بیان شده؛ همچنان چالش اصلی این رویکردها، فقدان اراده و نیت حقیقی در سامانه است، که پذیرش مسئولیت کیفری اصیل را دشوار می‌سازد؛ در واقع پاسخ این سوال در چارچوب چندلایه است: تعیین معیارهای الگوریتمیک قابل برنامه‌ریزی برای این مفاهیم در لایه طراحی و نظارت فعال انسانی در لایه بهره‌برداری برای موقعیت‌های غیرمنتظره؛ می‌تواند انطباق عوامل موجهه جرم را قابل ارزیابی نماید.

چالش در تعیین ذی‌نفع/ذی‌صلاح دفاع

اگر فرض شود که رفتار الگوریتمی را می‌توان ذیل دفاع مشروع یا اضطرار بررسی کرد، پرسش اساسی آن است که چه کسی ذی‌حق استفاده از این دفاع خواهد بود؟ سیستم هوشمند، نمی‌تواند منتفع از دفاع باشد، زیرا شخصیت کیفری مستقل ندارد؛ همچنین، مالک یا کاربردر این بین، محتمل‌ترین ذی‌نفع است، ولی همچنان پرسش‌هایی درباره میزان آگاهی و نقش او باقی خواهد ماند، و اگر ذی‌نفع دفاع طراح یا شرکت سازنده باشد؛ در برخی اسناد بین‌المللی گرایش بر این است که در صورت اثبات نقص یا سهل‌انگاری، مسئولیت متوجه آنان باشد (پارلمان اروپا[4]، ۲۰۲۱).

البته پاسخ به این چالش‌ها، در چارچوب چندلایه نهفته است، بنابراین پاسخ به این پرسش مستلزم تعریف دقیق نقش‌ها، حدود دسترسی و مسئولیت‌های هر یک از بازیگران در قراردادها و مقررات از پیش تعیین شده است.

1. Bertolini & Episcopo
2. Lagioia & Sartor
3. Bryson
4. European Parliament

چالش در بستر حقوق ایران و ضرورت تقنین پیش‌نگر

مواد ۱۵۶ و ۱۵۹ قانون مجازات اسلامی که به دفاع مشروع و اضطرار اختصاص یافته‌اند، آشکارا با فرض انسان به‌عنوان فاعل جرم تدوین شده‌اند. بنابراین در مواردی همچون خودروهای خودران یا ربات‌های پزشکی، قضات ناگزیر خواهند بود با تفسیر موسع این مواد تصمیم بگیرند؛ چنین تفسیری می‌تواند زمینه صدور آرای متفاوت و حتی متناقض را فراهم کند و امنیت حقوقی را کاهش دهد (شاهیده و قوانلو، ۱۴۰۳: ۹۸).

بررسی تطبیقی نشان می‌دهد که، در آلمان و انگلستان، بحث‌هایی درباره امکان اعمال دفاع مشروع در سامانه‌های نظامی خودکار در حال شکل‌گیری است (گلس، سیلورمن و وایگن[1]، ۲۰۱۶)؛ در اتحادیه اروپا، توصیه‌هایی برای تدوین مقررات خاص در خصوص تصمیم‌های الگوریتمی مطرح شده است (شورای اروپا[2]، ۲۰۲۲).

در ایران، هنوز مقررات اختصاصی وجود ندارد و تنها می‌توان با توسل به قواعد کلی موضوع را بررسی کرد. نتیجه این بخش آن است که حقوق کیفری ایران در وضعیت کنونی برای پاسخگویی به مسائل ناشی از عملکرد سامانه‌های هوش مصنوعی آمادگی کافی ندارد. ابهام در تعیین فاعل جرم، دشواری در انطباق عوامل موجهه، نامشخص بودن ذی‌نفع دفاع و نبود قوانین اختصاصی، همگی چالش‌های تقنینی و نظری را آشکار می‌سازد؛ که عدم توجه به، مقنن را در بهت حقوقی قرار خواهد داد.

این خلأ می‌تواند به تفسیرهای قضایی ناهمگون و نقض اصل امنیت حقوقی بینجامد. تطبیق رویکردهای نوین (به ویژه الگوی چندلایه) با اصول فقهی و حقوقی ایران نیازمند بازاندیشی است. برای نمونه، می‌توان با تأکید بر مسئولیت تضمین‌کننده ایمنی و مسئولیت ناشی از تخلف از شرط مراقبت، بخشی از بار مسئولیت را در چارچوبی مشابه الگوی چندلایه، بر عهده تولیدکنندگان و بهره‌برداران نهاد.

شفاف‌سازی مرز مسئولیت کیفری و مدنی در چارچوب هوش مصنوعی

یکی از چالش‌های بنیادین در تحلیل مسئولیت هوش مصنوعی، ابهام میان حوزه‌های کیفری، مدنی و شبه‌کیفری است. در نظام‌های کلاسیک حقوقی، مسئولیت کیفری مستلزم تحقق سه رکن اساسی است: رفتار مجرمانه، عنصر روانی (نیّت یا تقصیر) و عنصر شخصی (قابلیت انتساب به مرتکب). در حالی که مسئولیت مدنی یا شبه‌کیفری می‌تواند مبتنی بر نتیجه یا ریسک باشد و نیازمند عنصر روانی یا اراده آزاد نیست. با ظهور سامانه‌های هوشمند، مرز این دو حوزه دچار هم‌پوشانی و ابهام شده است، زیرا برخی رویکردهای نوین، به ویژه «مسئولیت مبتنی بر ریسک» و «نظام‌های بیمه‌ای اجباری»، اصولاً بر کنترل پیامدها و جبران خسارت تمرکز دارند و نه بر سرزنش اخلاقی یا مجازات کیفری.

به منظور شفاف‌سازی این مرز، می‌توان چارچوبی تفکیکی پیشنهاد کرد که شامل سه سطح عملیاتی است. نخست، مسئولیت کیفری محض، که مبتنی بر تحقق رفتار زیان‌بار همراه با عنصر روانی قابل سرزنش و قابلیت انتساب به یک فاعل انسانی مختار است. در حوزه هوش مصنوعی، این سطح شامل خطاها یا نقض‌هایی است که ناشی از تصمیم یا غفلت انسان-طراح، اپراتور یا کاربر باشد، به شرطی که عنصر اراده یا تقصیر انسانی قابل اثبات باشد. نمونه‌هایی از این سطح شامل استفاده ناصحیح از سیستم تسلیحاتی خودکار یا عدم توقف خودرو خودران در شرایط اضطراری علی‌رغم امکان پیش‌بینی خطا است.

سطح دوم، مسئولیت شبه کیفری یا مدنی، است که بر جبران خسارات ناشی از رفتار سامانه بدون نیاز به عنصر روانی یا قصد مجرمانه متمرکز است. این سطح مبتنی بر ریسک، بی‌احتیاطی یا نقص طراحی است و در زمینه هوش مصنوعی شامل خسارات ناشی از تصمیم‌های الگوریتمی غیرقابل پیش‌بینی، خطای سیستم یا نقص طراحی است که عنصر روانی انسانی قابل اثبات نیست. نمونه‌هایی از این سطح شامل تصادف خودرو خودران ناشی از اختلال الگوریتم و آسیب جانبی در پهپادهای خودمختار بدون تقصیر مستقیم اپراتور است.

1. Gless, Silverman & Weigend
2. Council of Europe

سطح سوم، الگوی تلفیقی چندلایه، این امکان را فراهم می‌کند که مسئولیت در یک زنجیره ارزشی توزیع شود و مرز کیفری و مدنی با وضوح بیشتری مشخص گردد. در این چارچوب، لایه طراحی و توسعه شامل مسئولیت کیفری ناشی از تقصیر یا نیت خطای انسانی و مسئولیت مدنی ناشی از نقص‌های قابل پیش‌بینی یا خطر ذاتی سیستم است. لایه بهره‌برداری و نظارت، مسئولیت کیفری را زمانی دربر می‌گیرد که اپراتور از دستورالعمل‌ها یا کنترل‌های اضطراری تخطی کند و مسئولیت شبه‌کیفری زمانی تحقق می‌یابد که رعایت دقیق دستورالعمل‌ها امکان‌پذیر نبوده و نتیجه زیان‌بار رخ دهد.

در نهایت، لایه تنظیم‌گری و جبران شامل مسئولیت مدنی یا شبه‌کیفری ناشی از نبود چارچوب‌های شفاف قانونی یا بیمه‌ای است و مسئولیت کیفری محدود به مواردی می‌شود که فقدان مقررات مستقیم به ارتکاب جرم منجر شود.

این شفاف‌سازی، ضمن مشخص کردن جایگاه هر رویکرد نوین، امکان تلفیق اصول سنتی مسئولیت کیفری با نوآوری‌های فناورانه را فراهم می‌آورد و مسیر جبران خسارت را برای قربانیان هموار می‌کند. با روشن شدن مرز میان مسئولیت کیفری، مدنی و شبه‌کیفری، داوران و نهادهای قانون‌گذاری می‌توانند رویکردهای نوین را به شکلی دقیق و سازگار با اصول سنتی حقوقی به کار گیرند و از هم‌پوشانی یا سوءتفاهم در تبیین مسئولیت جلوگیری نمایند.

پیشنهادات

۱. اولویت‌بخشی به مسئولیت مالک یا بهره‌بردار: قانون‌گذار می‌تواند همانند نظام حاکم بر وسایل نقلیه موتوری، مسئولیت کیفری و مدنی مالک یا بهره‌بردار سامانه هوش مصنوعی را در اولویت قرار دهد، زیرا این اشخاص بیشترین نقش در بهره‌برداری و نظارت عملی بر سامانه دارند.

۲. تدوین لایحه جامع ویژه هوش مصنوعی: تدوین یک لایحه جامع در حوزه مسئولیت کیفری و مدنی هوش مصنوعی ضروری است تا موضوعاتی مانند تعریف «فاعل جرم»، بازنگری در عوامل موجهه، و تعیین حدود مسئولیت در حوزه‌های مختلف روشن گردد و از برداشت‌های متعارض قضایی جلوگیری شود.

۳. ایجاد دستورالعمل‌های بخشی برای حوزه‌های پرخطر: در زمینه‌هایی مانند خودروهای خودران، ربات‌های پزشکی و سامانه‌های دفاعی، باید دستورالعمل‌های تخصصی تدوین شود که معیارهای ضرورت، تناسب و فوریت در تصمیم‌گیری‌های الگوریتمی را به زبان حقوقی روشن سازد.

٤. پیش‌بینی نظام بیمه‌ای اجباری: مشابه بیمه شخص ثالث در رانندگی، ایجاد بیمه اجباری برای جبران خسارات ناشی از خطاهای الگوریتمی می‌تواند بار مالی و حقوقی را کاهش دهد و اعتماد عمومی به استفاده از فناوری‌های نوین را افزایش دهد.

٥. ترجمه مفاهیم حقوقی به زبان الگوریتمی: یکی از خلأهای جدی، فقدان تلاش نظام‌مند برای بازتعریف مفاهیمی مانند ضرورت و تناسب در قالب معیارهای قابل‌برنامه‌ریزی است. توسعه «زبان مشترک میان حقوق و فناوری» می‌تواند امکان انطباق تصمیمات الگوریتمی با اصول بنیادین حقوق کیفری را فراهم کند.

٦. انجام مطالعات موردی تخصصی: تا کنون پژوهش‌های داخلی بیشتر کلی و نظری باقی مانده‌اند. انجام مطالعات عمیق و میان‌رشته‌ای در خصوص مسئولیت کیفری خودروهای خودران، ربات‌های پزشکی یا سامانه‌های نظامی خودکار می‌تواند مسیر سیاست‌گذاری دقیق‌تر را هموار کند.

۷. بررسی ایده شخصیت حقوقی محدود: هرچند در برخی منابع خارجی، ایده اعطای «شخصیت حقوقی محدود» به سامانه‌های هوش مصنوعی مطرح شده است، اما در حقوق ایران این موضوع هنوز به‌صورت منسجم بررسی نشده است. پرداختن به این بحث می‌تواند به غنای علمی و شفافیت تقنینی کمک کند، هرچند همچنان باید انسان به‌عنوان مسئول نهایی باقی بماند.

۸. ایجاد نهاد تنظیم‌گر: پیشنهاد می‌شود نهادی تخصصی برای نظارت بر توسعه و به‌کارگیری هوش مصنوعی در ایران تأسیس شود که وظیفه تدوین استانداردها، ارزیابی ریسک، و هماهنگی میان بخش‌های مختلف را بر عهده داشته باشد

نتیجه‌گیری

این پژوهش با هدف پاسخ به این پرسش اساسی شکل گرفت که در مواجهه با تصمیمات بالقوه زیان‌بار سامانه‌های هوش مصنوعی در موقعیت‌های دفاع مشروع و اضطرار، چارچوب قانونی و قضایی مناسب برای انتساب مسئولیت کیفری و تضمین عدالت چیست؟ یافته‌ها آشکار ساخت که الگوهای سنتی مسئولیت کیفری، که بر سه‌گانه «عنصر مادی، روانی و شخصی» و محوریت انسان مختار استوارند، در مواجهه با عاملیت غیرانسانی و خودآموز هوش مصنوعی دچار بن‌بست نظری و عملی می‌شوند. رویکردهای جایگزین نیز هر یک به تنهایی ناقص هستند: مسئولیت مبتنی بر پیش‌بینی با مشکل «جعبه سیاه» مواجه است، مسئولیت محض (ریسک‌محور) ممکن است نوآوری را سرکوب کند، و اعطای شخصیت حقوقی مستقل با مبانی فلسفی حقوق کیفری در تعارض قرار می‌گیرد.

در این میان، الگوی مسئولیت چندلایه به عنوان راه‌حلی جامع و عمل‌گرا مطرح می‌گردد. این مدل، با عبور از دوگانه بی‌ثمر «انسان یا ماشین»، مسئولیت را در امتداد یک زنجیره ارزشی توزیع می‌کند: از مسئولیت اخلاقی و تضمینی طراح و تولیدکننده در قبال ایمنی ذاتی و آموزش اخلاق‌مدار سیستم، تا مسئولیت نظارتی و احتیاطی بهره‌بردار در قبال استفاده صحیح و مداخله در شرایط بحران، و در نهایت مسئولیت تضمینی و جبرانی نهادهای تنظیم‌گر از طریق ایجاد چارچوب‌های شفاف و صندوق‌های جبران خسارت. این الگو نه تنها قادر است پیچیدگی فناورانه را مدیریت کند، بلکه با تأکید بر پاسخگویی نهایی انسان، سازگاری خود را با اصول بنیادین حقوق کیفری حفظ می‌نماید؛ گذر از بن‌بست نظری به راه‌حل عملی بر اساس مسئولیت چندلایه در مواجهه با دفاع مشروع و اضطرار الگوریتمی میتواند ازجنبه های کیفری، فنی و مسئولیت افق آینده استفاده از این سامانه های هوشمند را از منظر مسئولیت روشن نماید؛ یافته‌ها نشان می‌دهد که هسته نظریه سنتی مسئولیت کیفری بر محور مفاهیمی می‌چرخد که در دنیای الگوریتم‌های خودآموز فاقد مصداق عینی هستند. «اراده آزاد» به عنوان مبنای انتخاب و «قصد مجرمانه» به عنوان شرط اخلاقی مجازات، در مواجهه با سیستم‌هایی که بر پایه محاسبات احتمالاتی و بهینه‌سازی توابع هدف تصمیم می‌گیرند، دچار نوعی فروپاشی تحلیلی می‌شوند. پرسش محوری این نیست که آیا هوش مصنوعی اراده دارد، بلکه این است: سامانه‌ای که رفتارش محصول زنجیره‌ای از «اگر–آنگاه»های از پیش تعریف‌شده یا خروجی یک مدل آماری پیچیده است، چگونه می‌تواند موضوع سرزنش اخلاقی و مجازات کیفری قرار گیرد؟ در یک سناریوی عینی، یک سامانه پدافند خودکار که طیفی از اهداف را بر اساس الگوهای آموخته‌شده از داده‌های تاریخی شناسایی و منهدم می‌کند، فاقد «قصد دفاع» به معنای انسانی آن است. این عمل، یک واکنش محاسبه‌شده است، نه یک انتخاب اخلاقی. بنابراین، تلاش برای تحمیل چارچوب‌های سنتی به این پدیده نوین، تنها به انتساب‌های فرضی و غیرواقع‌بینانه یا انکار ساده مسئله می‌انجامد. این خلاء، ضرورت جست‌وجوی پارادایم‌های جایگزین مسئولیت را غیرقابل اجتناب می‌سازد.

بررسی هر یک از رویکردهای نوین در بستر واقع‌گرایانه موقعیت‌های دفاع مشروع و حالت اضطرار، محدودیت‌های ذاتی آن‌ها را آشکار می‌سازد:

مسئولیت مبتنی بر پیش‌بینی: در شرایط اضطراری کاملاً نوظهور و بی‌سابقه (مانند یک حمله سایبری–فیزیکی ترکیبی با الگوی ناشناخته)، معیار اساسی «پیش‌بینی‌پذیری برای یک متخصص معقول» خود دچار ابهام می‌شود. مرز بین ریسک غیرقابل پیش‌بینی و خطای قابل پیش‌بینی در این شرایط نامشخص است.

مسئولیت محض (ریسک‌محور): اگرچه در توزیع عادلانه بار جبران خسارت کارآمد است، اما فاقد ظرافت لازم برای تمایز قائل شدن بین عمل مجرمانه و عمل موجه است. این رویکرد نمی‌تواند میان خطای ناشی از نقص فنی و اقدام موجهی که متضمن تلفات جانبی اجتناب‌ناپذیر است (مثل تصمیم یک خودروی خودران برای انحراف به پیاده‌رو به منظور نجات جان سرنشینان) تفاوت بگذارد.

شخصیت حقوقی مستقل: این رویکرد در حل معمای «ذی‌صلاح و ذی‌نفع دفاع» کاملاً ناتوان است. یک سامانه، فاقد غریزه بقا، منفعت شخصی یا توانایی استدلال اخلاقی دربارهٔ عمل خود است. بنابراین، چگونه می‌توان ادعا کرد که به «دفاع از خود» یا دیگری پرداخته است؟ این ناتوانی، بنیان این نظریه را در موقعیت‌های اضطراری سست می‌کند.

الگوی مسئولیت چندلایه با کنار گذاشتن جست‌وجوی اتلاف‌بار برای یافتن یک «مقصر واحد» و تمرکز بر «شبکه علیت توزیع‌شده»، امکان تحلیل دقیق‌تر و عادلانه‌تری از حوادث را فراهم می‌آورد. این مدل، یک رویداد را در سه افق به هم پیوسته بررسی می‌کند:

۱. افق طراحی و توسعه (لایه اول – مسئولیت اخلاقی و تضمینی): آیا الگوریتم با داده‌های جانب‌دارانه یا ناقص آموزش دیده بود؟ آیا ارزش‌های اخلاقی بنیادین مانند اولویت جان غیرنظامیان یا اصل تناسب به درستی در منطق تصمیم‌گیری آن تعبیه شده بود؟ مسئولیت پاسخگویی در این افق متوجه تولیدکننده و توسعه‌دهنده است.

۲. افق بهره‌برداری و نظارت (لایه دوم – مسئولیت نظارتی و احتیاطی): آیا اپراتور یا کاربر نهایی، آموزش کافی دیده بود؟ آیا در لحظه بحران، امکان مداخله معنادار و لغو دستور را داشت؟ آیا پروتکل‌های عملیاتی روشنی برای شرایط غیرمنتظره وجود داشت؟ مسئولیت این افق بر عهده بهره‌بردار و ناظر انسانی است.

۳. افق تنظیم‌گری و جبران (لایه سوم – مسئولیت تضمینی و جبرانی): آیا نهاد ناظر مستقل و متخصصی پیش از استقرار، این سامانه را ارزیابی و تأیید کرده بود؟ آیا سازوکار سریع و منصفانه جبران خسارت برای قربانیان، بدون نیاز به گذر از اثبات تقصیر، پیش‌بینی شده بود؟ این مسئولیت متوجه قانون‌گذار و نهادهای تنظیم‌گر است.

در راستای تاثیر چندلایه مسئولیت باید گفت ، زمان که در حادثه‌ای فرضی یک سیستم پزشکی خودکار تشخیص و درمان در شرایط اضطرار ناشی از قطع برق طولانی‌مدت، مجبور به تخصیص منبع محدود دارو به یک بیمار از بین دو بیمار حیاتی می‌شود، مدل چندلایه پرسش‌های تحلیلی زیر را مطرح می‌کند:

در لایه طراحی: تابع هدف و الگوریتم تخصیص این سیستم بر چه مبنایی (شانس بقا، سن، فایده اجتماعی) تنظیم شده بود؟ آیا این مبانی توسط کمیته‌های اخلاق پزشکی بازبینی شده بود؟ پاسخ‌گویی در این لایه مشروط به شفافیت مبانی اخلاقی الگوریتم (مانند معیار تخصیص) و اخذ تأییدیه از کمیته‌های اخلاق مستقل است. عدم شفافیت یا طراحی بدون پیش‌بینی مکانیزم توقف اضطراری، تقصیر و مسئولیت مدنی/کیفری تولیدکننده را به دنبال دارد.

در لایه بهره‌برداری: آیا پرستار یا پزشک مسئول حاضر در محل، از حادثه مطلع بود و آیا دستورالعمل روشن و اختیار قانونی برای نقض تصمیم سیستم در موارد استثنایی را داشت؟ مسئولیت در این سطح منوط به آموزش کافی پرسنل، وجود پروتکل‌های شفاف برای نقض تصمیم سیستم و امکان واقعی مداخله انسانی است. اگر مدیریت بیمارستان در آموزش یا تدوین دستورالعمل قصور کرده باشد، مسئول اصلی است. غفلت پرسنل آگاه نیز مسئولیت مستقیم ایجاد می‌کند.

در لایه تنظیم‌گری: آیا بیمارستان موظف به داشتن منبع برق اضطراری با پشتیبانی کافی بود؟ آیا نهاد نظارتی سلامت، استانداردهای اجباری برای سیستم‌های خودمختار پزشکی در شرایط بحران تدوین کرده بود؟ مسئولیت نهادهای نظارتی در تدوین استانداردهای اجباری (مثل الزام به برق اضطراری)، اعطای مجوز مشروط پس از ارزیابی و نظارت مستمر پس از استقرار است. سهل‌انگاری در این وظایف، می‌تواند مسئولیت نهاد ناظر را محقق کند.

در نهایت باید گفت که، چالش هوش مصنوعی در حقوق کیفری، یک چالش ساختاری و پارادایمی است، نه یک مورد استثنایی ساده. الگوی مسئولیت چندلایه با پذیرش این واقعیت، به جای تمرکز بر یافتن مقصر، به توانمندسازی و پاسخگو ساختن کل زنجیره ارزش می‌اندیشد. برای نظام حقوقی ایران، پذیرش این الگو نه انتخاب یک مدل خارجی، که یک ضرورت درون‌زای راهبردی است تا بتواند در عین پاسداشت اصول جزمی خود، پاسخی عادلانه، عمل‌گرا و پیش‌گیرانه به یکی از پیچیده‌ترین مسائل حقوقی عصر حاضر ارائه دهد. این تحلیل، بنیان نظری مستحکمی برای پیشنهادهای سیاستی عینی و لایه‌بندی‌شده که در بخش نتیجه‌گیری ارائه گردید، فراهم می‌سازد. با این حال، کاربست این چارچوب در نظام حقوقی ایران، به ویژه در حوزه حساس دفاع مشروع و اضطرار، نیازمند بازاندیشی تقنینی و رویه‌ای فوری است. مواد ۱۵۶ و ۱۵۹ قانون مجازات اسلامی و دیگر مقررات مرتبط، بر فرض «فاعل انسانی» تدوین شده‌اند و خلائی آشکار در قبال تصمیمات خودمختار الگوریتمی وجود دارد. این خلأ می‌تواند به بی‌عدالتی، ناامنی حقوقی و فرار از مسئولیت بینجامد.

گذار از جامعه مبتنی بر انسان به جامعه انسان–ماشین، تنها با اتکا به قواعد دیروز میسر نیست. این پژوهش نشان داد که آینده حقوق کیفری در گرو پذیرش الگوهای انعطاف‌پذیر، توزیع‌شده و میان‌رشته‌ای مانند مسئولیت چندلایه است. تحقق این امر، نه یک انتخاب، که یک ضرورت فوری برای حفاظت از کرامت انسانی، حفظ حاکمیت قانون و تضمین امنیت ملی در عصر فناوری‌های خودمختار است. اقدام امروز قانون‌گذار و سیاست‌گذار، تعیین‌کننده چگونگی پاسخ تاریخ به پرسش از «عدالت در عصر هوش مصنوعی» خواهد بود.

ملاحظات اخلاقی

مشارکت نویسندگان

مشارکت نویسندگان در این مقاله به شکل زیر است:

نویسنده اول: تهیه و آماده سازی نمونه ها، انجام آزمایش و گردآوری داده ها، انجام محاسبات، تجزیه و تحلیل آماری داده ها، تحلیل و تفسیر اطلاعات و نتایج، تهیه پیش نویس مقاله.

نویسنده دوم: استاد راهنمای پایان نامه، طراحی پژوهش، نظارت بر مراحل انجام پژوهش، بررسی و کنترل نتایج، اصلاح، بازبینی.

نویسنده سوم: مشاوره پایان نامه، طراحی پژوهش، نظارت بر مراحل انجام پژوهش، بررسی و کنترل نتایج، اصلاح، بازبینی.

تعارض منافع

بر اساس اظهارات نویسندگان، این مقاله تعارض منافعی ندارد.

حامی مالی

بنابر اظهارات نویسندگان این پژوهش هیچگونه حامی مالی ندارد.

سپاسگزاری

از تمامی مشارکت کنندگان در این پژوهش سپاسگزاری می شود.

منابع

شاهیده، فرهاد و قوانلو، طاهره. (۱۴۰۲). مسئولیت کیفری ربات‌ها. چاپ سوم. تهران: بنیاد حقوقی میزان.

حسین زاده، جواد؛ احمدی، فرید و کلب خانی، هاشم . (۱۴۰۳). ارائه یک مدل جامع برای سنجش آمادگی کشورها در مواجهه با انقلاب صنعتی چهارم. *آینده‌پژوهی دفاعی*، ۹(۳۲)، ۱۶۱-۱۹۰.10.22034/dfsr.2024.2019174.1761

بنافی، فرشته . (۱۴۰۲). حفاظت از حقّ حریم خصوصی اطلاعاتی در مقابل تهدیدات ناشی از هوش مصنوعی نظامی. پژوهش حقوق خصوصی، ۱۲(۴۵)، ۱۴۹-۱۷۶. 10.22054/jplr.2024.68659.2691

ترکی، هادی. (۱۴۰۳). رهیافت جدید قدرت مبتنی بر هوش مصنوعی (مطالعه موردی رقابت آمریکا و چین از ۲۰۱۰ تا ۲۰۲۳). م*طالعات راهبردی آمریکا*، ۴(۲)، ۹۱-۱۱۴. 10.47176/asr.2024.1206

ذاکری‌نیا، حانیه. (۱۴۰۲). ماهیت و مبنای مسئولیت مدنی ناشی از هوش مصنوعی در حقوق ایران و کشورهای اتحادیه اروپا. *فصلنامه حقوق تطبیقی*، ۲۰(۱)، ۱۳۵-۱۵۲. 10.22059/jolt.2023.356703.1007186

رستمی، محسن. (۱۴۰۱). ناسایی و معرفی ظرفیت‌های کاربردی هوش مصنوعی در توسعه مضمون‌های راهبردی در سازمان‌های نظامی. *فصلنامه مطالعات دفاعی و امنیتی*، دانشگاه عالی دفاع ملی،۷۸(۲۰)، ۳۴-۷۳. https://ds.sndu.ac.ir/article_2167.html

السان، مصطفی و دهستانی، سوور . (۱۴۰۱). جنبه‌های حقوقی جعل عمیق. *فصلنامه تحقیقات حقوقی*، ۲۵(ویژه نامه حقوق و فناوری)، ۱۹۳-۲۱۸.

قاسمی، محمدهادی؛ رحیمی، مهدی و عیوضی، محمدرحیم . (۱۴۰۴). کاربست قابلیت‌های هوش مصنوعی در فرایند آینده‌نگاری راهبردی. *آینده پژوهی ایران*، ۱۰(۱)، ۶۶-۱۰۴. 10.30479/jfs.2023.18401.1463

نوری‌سوا، احمد. (۱۴۰۴). بررسی آثار مسئولیت کیفری هوش مصنوعی. *فصلنامه حقوق جزا و جرم‌شناسی*.

https://civilica.com/doc/2336426

References

Abdelaziz, D. K. A. (2025). Criminal liability for the misuse and crimes committed by AI: A comparative analysis of legislation and international conventions. *Journal of Infrastructure, Policy and Development*, 9(1), 10722. https://doi.org/10.24294/JIPD10722

Associated Press News. (2023). Autonomous robots in the Ukraine war. AP. Retrieved from https://apnews.com News, A. P. (2023). Autonomous robots in the Ukraine war.

Axios. (2024). OpenAI and Anduril partner on Pentagon AI drone defense. Axios. Retrieved from https://axios.com

Bertolini, A., & Episcopo, M. (2020). Artificial intelligence and civil liability. *European Journal of Risk Regulation*, 11(2), 395–415. https://doi.org/10.1017/err.2020.24

Bryson, J. J. (2020). The Artificial Intelligence of the Ethics of Artificial Intelligence: An Introductory Overview for Law and Regulation. The Oxford Handbook of Ethics of AI, 2–25. https://doi.org/10.1093/OXFORDHB/9780190067397.013.1

Buiten, M., de Streel, A., & Peitz, M. (2023). The law and economics of AI liability. Computer Law and Security Review, 48. https://doi.org/10.1016/j.clsr.2023.105794

Business Insider - Latest News in Tech, Markets, Economy & Innovation. (2024). https://www.businessinsider.com/

Council of Europe. (2022). Recommendation on the impacts of algorithmic systems on human rights. Strasbourg: Council of Europe. Retrieved from https://www.coe.int

Dahshan, Y. I. (2021). Criminal Liability for Artificial Intelligence Crimes. UAEU Law Journal, 2020(82), 2. https://scholarworks.uaeu.ac.ae/sharia_and_law/vol2020/iss82/2.

El-Kady, R. (2024). Artificial intelligence from a criminal law perspective. *Al-Zaytoonah University of Jordan Journal for Legal Studies*, 5(2), 168–210.

European Parliament. (2021). Resolution with recommendations to the Commission on a civil liability regime for artificial intelligence (2020/2014(INL)). Official Journal of the European Union. Retrieved from https://www.europarl.europa.eu

Farajpour, R. (2025). Legal and comparative analysis of civil liability of artificial intelligence in automated decision-making. AI and Tech in Behavioral and Social Sciences. https://journals.kmanpub.com/index.php/aitechbesosci/article/view/3682

Farajpour, R., Amerinia, M. B., & Pourjavaheri, A. (2025). The role of artificial intelligence in arbitration and legal challenges arising from automated decisions in sports. AI and Tech in Behavioral and Social Sciences. https://journals.kmanpub.com/index.php/aitechbesosci/article/view/3737

Farajpour, R. (2025). The role of civil liability in artificial intelligence laws from the perspective of major global legal systems. *Journal of Law and Political Studies*, 5(1), 182-196. https://doi.org/10.48309/jlps.2025.518711.1353

Georgetown University. (2023). Necessity in jus ad bellum. Center for Security and International Studies.

German Startup Wants to Regrow Europe's 'Spine' With AI Fighter Pilots, Drone Walls - WSJ. ([s.d.]). Recuperado 28 de setembro de 2025, de https://www.wsj.com/world/europe/german-startup-wants-to-regrow-europes-spine-with-ai-fighter-pilots-drone-walls-09c852f8

Gless, S., & Ligeti, K. (2024). Regulating driving automation in the European Union – criminal liability on the road ahead? *New Journal of European Criminal Law*, 15(1). https://doi.org/10.1177/20322844231213336

Hammouri, J. A. A. (2024). Subjectivity of artificial intelligence in criminal law: New challenges. *Edelweiss Applied Science and Technology*, 8(6), 3832–3842. https://doi.org/10.55214/25768484.V8I6.2835

Hilgendorf, E. (2018). Introduction: Digitization and the Law – a European Perspective. *Digitization and the Law*, 9–20.

Lagioia, F., & Sartor, G. (2020). AI Systems Under Criminal Law: a Legal Analysis and a Regulatory Perspective. *Philosophy and Technology*, 33(3). https://doi.org/10.1007/s13347-019-00362-x

Lieber Institute. (2025). Interpreting the Law of Self-Defense - Lieber Institute West Point. (2025). Lieber Institute. https://lieber.westpoint.edu/interpreting-law-self-defense/

Lieber Institute. (2025). Reconsidering anticipatory self-defense in international law. Lieber Institute Articles, West Point.

Pagallo, U. (2013). The laws of robots: Crimes, contracts, and torts. *The Laws of Robots: Crimes, Contracts, and Torts*, 1–200.

Panattoni, B. (2025). Generative AI and criminal law. Cambridge Forum on AI: Law and Governance, 1, e9. https://doi.org/10.1017/CFL.2024.9

Papagiannidis, E, Mikalef, P, Conboy K. (2025). Responsible artificial intelligence governance: A review and research framework, *The Journal of Strategic Information Systems*, 34(2):101885. https://doi.org/10.1016/j.jsis.2024.101885

Professor Ashley Deeks. ([s.d.]). Using Artificial Intelligence in Warfare Creates Constitutional Concerns, Argues Professor | University of Virginia School of Law.

Recuperado 28 de setembro de 2025, de https://www.law.virginia.edu/news/202410/using-artificial-intelligence-warfare-creates-constitutional-concerns-argues-professorThe Council of Europe: guardian of Human Rights, Democracy and the Rule of Law for 700 million citizens - Portal. ([s.d.]). Recuperado 28 de setembro de 2025, de https://www.coe.int/en/web/portal/home

Rodríguez de Las Heras Ballell T. Mapping Generative AI rules and liability scenarios in the AI Act, and in the proposed EU liability rules for AI liability. Cambridge Forum on AI: Law and Governance. 2025;1:e5.

The Economic Times. (2025). Indian army testing AI machine gun that can detect, decide, and destroy enemy on its own. ET Defense. Retrieved from https://economictimes.indiatimes.com

The Wall Street Journal. (2025). German startup wants to regrow Europe’s spine with AI fighter pilots and drone walls. WSJ. Retrieved from https://wsj.com

TIME. (2024). Israel’s Use of AI in Gaza May Be Setting a New Warfare Norm | TIME. https://time.com/7202584/gaza-ukraine-ai-warfare/