Theory and Practice of Science and Technology
2025, VOL. 6, NO. 12, 30-34
DOI: 10.47297/taposatWSP2633-456907.20250612

AI-Driven Security Protection for Public Security Services: Risk Identification and Mitigation

Yulei Lu

Graduate student of Zhejiang Police College, Haiyan County Public Security Bureau, Jiaxing, Zhejiang, 314300, China

ABSTRACT
With the deep integration of artificial intelligence (AI) into public security work, public security organs across China have actively explored the application of large AI model technologies to develop intelligent infrastructure including training datasets, knowledge bases, and public security-specific large models, injecting technological momentum into the continuous enhancement of the new-type combat capabilities of public security organs. However, in the process of improving policing efficiency, issues such as data security breaches, algorithmic bias, and adversarial attacks have occurred frequently. The problem of black-box algorithms has aroused widespread public concern, and ensuring the safety, reliability, and controllability of AI has become a prominent issue, which is increasingly threatening national security. This study systematically analyzes various security risks in AI-driven public security services, and proposes responsive strategies to strengthen confidential security management and data security protection from the perspectives of ideological risks, policy and legal norms, technical construction risks, and regulatory governance mechanisms. It provides insights for public security organs at all levels to seize the opportunities of AI empowerment while using AI technologies in a safe manner.

KEYWORDS
Artificial intelligence technology; New-type combat capabilities of public security organs; Black-box algorithms; Confidential security management; Data security protection

1 Introduction

The deepening application of artificial intelligence technology in public security is reshaping the new landscape of public security work. With the integration of AI tools such as Deepseek and Tongyi Qianwen into various public security business areas, massive volumes of public security operational data have unleashed immense potential by virtue of algorithmic logic and the advantages of large models. Various types of unstructured public security data have also been efficiently integrated and intelligently applied. The application of multimodal fusion technology has significantly reduced the cost and time for public security organs in analyzing unstructured data. From intelligent AI police officers and AI policy consultation assistants to intelligent applications such as police incident analysis and public opinion monitoring, the implementation of the "AI + Public Security" initiative has been continuously deepened. Public security organs' development of intelligent infrastructure using large AI model technologies is a strategic measure to safeguard national security. By building a multimodal integrated policing AI platform, intelligent research and judgment of massive data and risk early warning can be realized, which greatly enhances the ability to predict, warn against, and prevent threats such as violent and terrorist activities, cyber infiltration, and cross-border crimes.

However, the vigorous development of new technologies is accompanied by corresponding risks and challenges. In the process of building large AI models, public security organs handle highly sensitive operational data, which contains a great deal of information related to public security and citizens' privacy. Once such data is leaked, it may lead to severe public security incidents and economic losses, and even encroach on national security. On the other hand, the issue of black-box algorithms has sparked intense public discussion, making it difficult to explain the fairness and accuracy of law enforcement and case handling assisted by large AI models in public security work. If public security organs improperly apply technologies or manage data during the construction of large AI models, it may pose potential risks to national security. For example, large model training relies on massive datasets; if these datasets contain sensitive policing information or citizens' privacy data, a cyber-attack or internal data leakage could result in the data being exploited by hostile forces, endangering the security of critical infrastructure or social order.

In September 2024, to strengthen the standardization of AI security governance, the Ministry of Public Security established the Technical Committee for the Standardization of Artificial Intelligence Security Governance, which is specifically responsible for the overall coordination of standardization work in the field of AI security technologies. However, current research on AI security construction in public security mainly focuses on technical aspects, and there remains a gap in systematically exploring the secure application of AI technologies in the public security field from a multi-dimensional perspective. Based on the background of public security organs exploring the era of AI-driven policing, this study re-examines the relationship between AI technology and multi-faceted security, clarifies the major risks in the technological transformation of public security from the dual perspectives of risk identification and prevention, and further explores strategies for addressing security risks in AI-driven policing.

2 Main Risks Faced by Public Security Organs in AI Construction

2.1 Ideological Security Boundary Issues

(1) Artificial intelligence technology is not neutral when processing public security data. According to its built-in algorithmic logic, it performs inherent tasks such as data simplification, structured integration, and correlation construction, and the resulting outputs are influenced by its pre-set value orientations and ideological tendencies. As some scholars have pointed out, generative artificial intelligence represented by ChatGPT actually has strong capitalist ideological attributes. In essence, large AI models are products of human ideology, so it is necessary to guard against the negative impacts of the ideologies of specific interest groups on the ideological construction of public security through hidden algorithm embedding and hardware facilities, which may cause the ideological deviation of public security police officers during human-computer interaction. On the other hand, the construction of public security large AI models is based on the results of data processing and analysis, while public security police officers' law enforcement needs to comprehensively consider complex factors such as politics, law, and social benefits. Relying on AI to assist law enforcement decision-making may lead some police officers to over-rely on technology, neglect the improvement of their own law enforcement and service literacy, and instead prioritize the outputs of AI algorithms, allowing AI processing results to replace their own judgments, which may lead to law enforcement loopholes. In the long run, this may cause the overall quality of the police force to decline, the business processing model to become rigid, and ultimately lead to the serious consequence of "dehumanization" in law enforcement.

2.2 Issues in Defining Technological Application Risks

The "algorithmic black-box" problem of public security large AI models poses challenges to the transparency of law enforcement and public trust. Since large AI models are usually tested using test datasets, their operational logic rules and internal prompt words are generally set by model developers. Once the test datasets contain "biased data" from historical law enforcement practices—such as labeling migrant populations in a certain area as a high-risk group for intellectual property infringement—large AI models are likely to reduce complex social issues to statistical problems, and such biases will be amplified through algorithmic reinforcement, leading to output bias and ultimately triggering a crisis of public trust in fair law enforcement. At the same time, frontline police officers only operate through the "input query-output result" mode, with the internal decision-making process remaining invisible. Due to technical limitations in complex environments, large AI models cannot perform complex identification tasks like "real police officers". Some large AI models integrate and push relevant information based on user demand-oriented data integration logic, which cannot guarantee the authenticity of the information and may lead to the distortion of final conclusions. If police officers adopt distorted information, it will to a certain extent have a negative impact on the public security image and public trust in its work. Moreover, the "algorithmic black-box" feature makes it unclear to identify the rules responsible for law enforcement errors; in the event of law enforcement mistakes, it is difficult to define whether the responsibility lies with the model developers, the police, or the model algorithms themselves.

2.3 Risks of Lack of Autonomy

Public security organs often face the problem of lack of autonomy in the construction and operation of large AI models. When using open-source platforms for privatized deployment or orchestration tools to invoke models, public security organs usually adopt universal commercial platforms rather than establishing their own internal tool platforms. If the default configurations of third-party tools have vulnerabilities such as unauthorized access and model theft, it is very easy to trigger public security network and data security incidents. For example, when using commonly used tools like Ollama to deploy large models such as Deepseek locally, a web service will be launched locally, exposing port 11434 to the public network and causing problems such as data leakage and computing power theft. When using orchestration tools such as Dify, if third-party vendors' API-based large model services are selected instead of independent deployment and design, the token consumption of each conversation may lead to the leakage of public security document data to third-party vendors. If the construction of public security large AI models is undertaken by technology companies, it is easy to be bound by third-party services and lack independent operation capabilities. In the process of building the underlying logic of the model, it may be impossible to independently modify, adjust, and upgrade due to restrictions on the source code of technology companies or the setting of encrypted interfaces. Meanwhile, in the case of public cloud or hybrid cloud construction, data ownership rights may be held by cloud service providers, making public security organs prone to losing control over the storage and processing of public security data. The training of public security large AI models requires high-performance GPUs; domestic chips currently cannot meet the computing power demand, so most chips used are foreign-made (such as NVIDIA). Additionally, some underlying AI frameworks used by public security organs (such as TensorFlow and PyTorch) are also developed overseas, which may contain "backdoors". If public security

large AI models are controlled by third parties, it will pose significant hidden dangers to public security data security.

2.4 Legal and Regulatory Policy Issues

The construction of large AI models by public security organs not only requires technical specifications but also is a systematic project involving comprehensive legal, ethical, and social supervision. The currently implemented Interim Measures for the Administration of Generative Artificial Intelligence Services regulates the healthy development of artificial intelligence from aspects such as technological development and governance, service specifications, supervision and inspection, and legal liability. However, on the whole, these measures are relatively macro, with unclear specific applicable management requirements for internal use in the public security system, and they cannot effectively guide the resolution of problems encountered in specific practices. The Measures for the Labeling of AI-Generated Synthetic Content, which will be implemented in September, only regulates the labeling of AI-generated synthetic content. Due to the complexity of generative artificial intelligence, there are still deficiencies in legal norms regarding how to construct and use large AI models. In terms of personal information protection, compliant data usage, and sensitive data control, specific regulatory measures have not yet been formed. In particular, since large AI models involve cross-departmental and multi-domain applications, joint supervision by multiple departments is required, but in actual supervision, inter-departmental collaborative governance has not formed an effective regulatory synergy. On the other hand, AI technology is developing rapidly, and public security organs will encounter new problems beyond the scope of current policy norms during the construction of large models. Legal and regulatory policies usually have a lagging nature and are difficult to adapt to the new needs of current technical specifications.

3 Strategies for Addressing Risks in Public Security Organs' AI Construction

3.1 Guiding Technology for Good with Firm Political Conviction

Public security organs must clarify the value orientation of constructing and using large AI models, adhere to consolidating the guiding position of Marxism in the ideological field of artificial intelligence in public security, and "govern algorithms with mainstream value orientations". In the process of building large AI models, it is necessary to integrate core contents such as the theoretical system of socialism with Chinese characteristics and relevant laws and regulations into the underlying logical framework, increase the algorithm priority weight of core values such as serving the people wholeheartedly and promoting social fairness and justice, and ensure that model outputs conform to mainstream ideological content. A full-process supervision and management mechanism should be established, with strict control over the source of training data collection, strengthened political review of various types of data, and the construction of a multi-level content filtering system targeting specific sensitive words and non-mainstream ideologies. At the final stage of training, high-quality data that conforms to ideological requirements should be used for supervised fine-tuning. After application and deployment, real-time monitoring should be carried out to ensure that the output results of large AI models conform to socialist core values. On the other hand, it is necessary to clarify the professional responsibilities and social obligations of public security police officers in constructing and using large AI models, establish the management and usage principle of "whoever is in charge is responsible, whoever operates is responsible, whoever uses is responsible", and adopt the business process of "AI pre-processing + independent police decision-making" to ensure the high efficiency of human-computer collaboration. Public security police officers should regularly carry out quality improvement training, incorporate the correct use of AI into training courses, strengthen talent echelon construction, and promote public security organs in various regions to build internal talent teams with high adaptability and strong autonomy, so that AI technology can exert high-efficiency combat effectiveness in the public security field.

3.2 Promoting Technology Empowerment through Iterative Self-Purification

XAI (Explainable Artificial Intelligence) is an AI technology that can provide visualized, understandable, and traceable decision-making logic. Its core objectives are Transparency, Interpretability, Trustworthiness, and Accountability. To effectively address the "black-box algorithm" problem of public security large AI models, XAI should be embedded into the entire application process of public security large models, ensuring that every decision-making process of the large AI model is deeply integrated with the professional judgment of the police force, forming a transparent, credible, and traceable closed-loop decision-making application system. In the stage of public security data collection and preprocessing, the process of extracting important data features should be explained, and the source of bias and calculation rules should be labeled for each piece of "abnormal data", enabling users to understand the overall model operation rules. In the model analysis stage, not only global explanations (such as correlation weight relationships) should be provided, but also local explanatory bases for individual research and judgment results, and manual review functions

should be available for uncertain results. In the decision-making stage, the original "black-box algorithms" should be replaced with "if-then" rules, such as "after 9 p.m. + carrying bird-catching equipment + appearing in mountain forest areas → push early warning". In the supervision and feedback stage, attention mechanisms should be used to conduct visualized analysis of the results of abnormal attention from AI, and automatic generation of periodic feedback reports should be required, such as "the misjudgment rate of mountain forest bird-catching early warnings has increased by 15% in the past week; please update the rule weights".

3.3 Seizing the Initiative through Independent Research and Development

When innovating and popularizing the in-depth application of artificial intelligence in policing scenarios, public security organs at all levels must seize the initiative in new technologies and methods, and take the path of independent research and development and safe and credible innovation. In terms of technology platform construction, it is necessary to meet the requirements of system integration, data interconnection, application sharing, and global response to research and judgment, develop a flexible and scalable zero-code development platform suitable for application in public security environments and applicable to various police departments. Meanwhile, unify the technical bases of platforms in all provinces, provide one-stop toolchains for model knowledge management, model training, and practical testing, enabling public security organs in various regions and professional police departments to quickly build intelligent agents and workflows. In terms of data technology application, it is necessary to promote mutual recognition and sharing of provincial and municipal resource catalogs, realize the full-domain application of resources such as data, computing power, components, and services across the province, establish vanguard teams for public security large AI model construction focusing on technological innovation, such as "public security research institutes" and "public security brain" laboratories, and jointly build public security AI research projects with more universities, technology enterprises, and social organizations to enhance the exploration and dissemination of advanced technologies in the public security field. In terms of project construction, implement "unified account management" to eliminate decentralized and redundant construction, and avoid resource waste caused by disorderly multi-head development. When building AI models in specific fields, carry out "distillation training" to further compress the model scale and reduce unnecessary computing power demand. In project procurement, continuously explore localization alternatives and adhere to the path of independent innovation, to avoid data leakage caused by public security large models being controlled by third parties.

3.4 Ensuring Healthy Development through Strengthened Supervision

Continuous efforts should be made to improve the formation of laws and regulations in the field of policing artificial intelligence construction, further clarify the boundaries of data usage, human-computer collaboration mechanisms, and the legal effect of algorithmic decisions for public security large AI models, and standardize the use and processing of classified public security data. Strictly implement the spirit of relevant documents such as the Regulations on the Management of Internal Networks and Data Security of Public Security Organs, promote the establishment and integration of a zero-trust security system, and especially strengthen the integrated protection of digital terminals, access boundaries, and cloud security. Establish a full-life-cycle large model supervision mechanism, strictly implement the Ministry of Public Security's requirement of "approval before implementation" for large model construction and development, further promote the pre-development filing and review system, and strengthen security supervision. During the application stage, conduct dynamic detection of all applications, network boundaries, and cooperative enterprises, carry out strict security audits of public security data usage, achieve real-time early warning, and issue timely security notifications to ensure the legitimate and compliant use of data. When developing specific large AI models, "sandbox experiments" can be carried out, dividing three risk levels: low-risk areas should first develop low-risk AI policing models such as document generation and official document drafting; medium-risk areas can research AI policing models such as public opinion analysis and video monitoring; after completing the exploration of models in low and medium-risk areas, high-risk policing models such as police incident analysis and crime prediction can be developed. After the completion of application construction, annual audit work should be carried out, emergency plans for unexpected incidents should be formulated, and an emergency control "circuit breaker mechanism" should be established to realize the forced suspension of problematic models in emergency situations. Strengthen the collaborative governance mechanism, especially the implementation of cross-departmental joint governance. At the same time, strictly implement application permission allocation, and grant different model permissions according to data levels, business scenarios, functional levels, and operation content, so as to balance practical needs and security requirements.

4 Conclusion

The in-depth application of artificial intelligence technology has provided strong technical support for public security organs, but it also brings risks and challenges in practical application. Therefore, it is essential to establish rigorous algorithmic auditing, data security protection, and supply chain review mechanisms to ensure that AI development always operates on a safe and controllable track, effectively safeguarding national security interests. Based on the construction of AI-driven policing large models by public security organs across China, this study explores the ideological, technological application, legal, and regulatory risks encountered in practical work from a multi-source perspective, and proposes systematic governance strategies and response plans based on potential risks, providing risk warnings for the construction of AI-driven policing large models in various regions.

As an important functional department responsible for safeguarding national security and protecting people's well-being, public security organs should take proactive actions, make overall plans, adhere to the concept of "security and inclusiveness, practical orientation", and accelerate the creation of high-quality scenarios for the intelligent and practical application of public security characterized by high technology, high efficiency, and high quality. In the future, on the basis of effectively identifying and addressing risks, public security organs will continue to deepen technological integration and institutional innovation, continuously inject technological momentum into the enhancement of the new-type combat capabilities of public security organs, and provide solid public security guarantees for the development of Chinese-style modernization.

References

[1] Liu CY, Yu YC. Ideological security risks of generative artificial intelligence and their governance [J]. J Guangxi Police Coll, 2025, 38(3).

[2] Pu QP, Qi Y. Opportunities, challenges and responses of generative artificial intelligence such as DeepSeek to the security of mainstream ideology [J]. J Chongqing Univ (Soc Sci Ed), 2025, 04(002).

[3] Building High-Quality Partnerships and Opening a New Journey for BRICS Cooperation—Speech at the 14th BRICS Summit [N]. People's Daily, 2022-06-24 (02).

[4] Pu QP, Qi Y. Opportunities, challenges and responses of generative artificial intelligence such as DeepSeek to the security of mainstream ideology [J/OL]. J Chongqing Univ (Soc Sci Ed). [2025-04-13]. http://kns.cnki.net/kcms/detail/50.1023.C.20250410.1029.004.html.

[5] Yang ZW. The ideological attributes and risk regulation of generative artificial intelligence such as ChatGPT [J]. Inner Mongolia Social Sci, 2024 (1): 60.

[6] Zhou Y, Guo LR. The connotation, manifestations and generation logic of implicit security risks in public data opening [J]. Inf Doc Serv, 2024, 45 (3): 70-77.

[7] Chen ZQ. Regulatory sandbox: A new scheme for data element governance [J]. Library Tribune, 2024, 44(4): 138-147.

[8] Kibriya H, Khan WZ, Siddiga A, et al. Privacy issues in large language models: A survey [J]. Comput Electr Eng, 2024.

[9] Recorded Future. ChatGPT Is a "Powerful" Tool for Cybercrime: Recorded Future [R/OL]. 2024-05-20. https://www. crn. com/news/security/chatgpt-is-a-powerful-tool-for-cybercrime-recorded-future.

[10] Choi N, Kim H. Technological convergence of blockchain and artificial intelligence: A review and challenges [J]. Electronics, 2025, 14(1): 84.