Proceedings of CONF-MLA 2025 Symposium: Applied Artificial Intelligence Research
DOI: 10.54254/2755-2721/2025.BJ27253

Research on Information Security Issues in the Field of Artificial Intelligence

Tianyi Zhang

Institute of Computer Science, Beijing University of Technology, Beijing, China
zhangty_2004@bjut.edu.cn

Abstract. With the rapid development of artificial intelligence technology, the information security problems it brings have become increasingly prominent, such as data poisoning, adversarial attacks, and privacy leakage. This paper aims to study these core challenges, explore the significance of balancing technological innovation and risks, and ensuring the security of countries, societies, and individuals. Drawing on literature review, case analysis, and the SWOT framework, and integrating data from the 2025 Global AI Security Report, this study proposes a localized multi-dimensional governance framework tailored to China. Covering technical, legal, ethical, social, and international cooperation dimensions, the framework provides a reference for AI information security governance.

Keywords: Artificial Intelligence, AI Information Security Risk, Multi-dimensional Governance Framework, China's Path

1. Introduction

Global AI technology is advancing at a rapid pace, but security challenges have emerged in its wake. The emergence of generative AI such as ChatGPT-5 has triggered new types of attacks. According to Gartner's report, AI-related cyber incidents increased by 25% in the first half of 2025. On the policy front, countries worldwide have expedited their legislative processes [1]. The European Union has introduced the "Artificial Intelligence Act", the United States has launched the "National AI Security Initiative", and China has implemented the "Interim Measures for the Administration of Generative Artificial Intelligence Services" and the proposed "Artificial Intelligence Law" [2]. At the same time, the evolution of multimodal AI and autonomous agent systems has amplified risks, such as the emerging trend of AI-driven social engineering attacks [3]. Balancing innovation and risks is crucial. It is of great significance to protect critical infrastructure such as smart grids from AI attacks. It is expected that the scale of the AI security market will reach 500 billion US dollars by 2030 [4]. The threats brought by AI are multifaceted, involving ethical bias amplification, privacy needs for differential privacy in federated learning, and national security aspects of AI applications in cyber warfare. In addition, the research can fill the gap in localized AI security governance and promote the integration of the "AI + security" industry [5]. This paper explores technical risk analysis, application scenarios, governance challenges, cutting-edge trends, and recommendations using a combination of qualitative and empirical methods [6].

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.BJ27253

2. AI information security risk system

2.1. Technically endogenous risks

Data poisoning attacks have become a major hidden danger in large model training. A 2025 study by New York University showed that in medical field model training, injecting only 0.001% of false data (about 2000 malicious articles, costing 5 US dollars) can lead to a 7.2% increase in harmful content output by the model [7]. Different poisoning ratios have significant differences in their impact on the model (follow Table 1).

Table 1. Comparison of data poisoning effects and costs (source: Nature Medicine 2025)

Poisoning Ratio	Increase Rate of Harmful Content	Training Cost (US Dollars)
0.001%	7.2%	5
0.01%	11.2%	50
0.1%	18.5%	500

The risk of privacy leakage is also severe. Large model corpora may contain undisclosed sensitive information. For example, a medical AI training data accidentally included patients' diagnosis records, leading to the leakage of 100,000 pieces of personal information [8]. In terms of supply chain risks, vulnerabilities in open-source datasets have become an attack entry point. In 2025, a self-driving dataset was implanted with malicious code, causing 30% of test vehicles to have sensor misjudgments [9]. The trust crisis in AI has been triggered by the unexplainability of black-box decisions, such as the excessive focus on background noise in KAN neural networks in image recognition [8,10]. The evolution of adversarial attack methods has also led to a threat to intellectual property rights, with competitors copying a financial AI system's risk assessment algorithm, resulting in $20 million in economic losses [7]. Additionally, vulnerabilities in edge devices have become springboards for attacks, with 30% of cameras in IoT-AI integrated systems containing unpatched remote code execution vulnerabilities. Autonomous system failures are also frequent, with a self-driving car causing rear-end collisions due to sensor spoofing attacks in 2025.

2.2. Application scenario risks

The weaponization of AI has led to an arms race, with autonomous drone systems posing a 5% probability of mistakenly attacking civilian targets. AI-generated fake news spreads faster than manually created content, influencing 10% of voters' decisions during elections [10]. Intelligence leakage risks have intensified, with AI-assisted espionage activities deducing military base deployments [11]. Critical infrastructure faces threats, with a 2025 simulated attack report revealing attackers disabled a power grid's AI control system, causing a 12-hour regional blackout and $50 million in economic losses [10]. Vertical industries face risks, with a medical AI data breach exposing patient privacy. AI-related cybercrime causes annual global GDP losses of $1 trillion, with financial fraud accounting for 40% [7]. Misuse of biometrics triggers privacy crises, with facial recognition systems having a 10% false positive rate. Deepfake fraud cases have surged, and algorithmic recommendation biases on social media exacerbate the "information cocoon."

DOI: 10.54254/2755-2721/2025.BJ27253

2.3. Governance mechanism risks

Quantum AI threats pose significant challenges to current regulations, including unclear responsibility attribution, cross-border law enforcement difficulties, and a lack of unified credible evaluation systems. International standard conflicts, such as the EU GDPR and China's Data Security Law, can increase compliance costs for multinational enterprises by 40%. Talent shortage restricts governance effectiveness. In an AI security team of a regulatory agency, only 15% of experts have actual combat experience in attack and defense, making it difficult to cope with rapidly iterative technical risks. The dynamic adaptability is insufficient. AI models are updated every 2 weeks on average, while the regulatory cycle is as long as 3 months, resulting in 60% of new risks not being identified in time [12].

3. The double-edged sword effect of AI in the field of information security

3.1. AI empowering security defense

Machine learning-driven abnormal traffic analysis achieves 95% real-time DDoS identification accuracy and predicts network intrusions 30 minutes in advance, cutting false positive rates from 20% to 5% [8]. Meanwhile, AI-powered SOAR systems automate intrusion blocking and forensics, slashing response times from hours to minutes [5]. A self-healing system integrated with zero-trust architecture restores services within 5 minutes post-attack, boosting availability to 99.99%. Additionally, behavioral biometric authentication, via multi-factor AI verification, reduces identity fraud from 5% to 0.1%, while AI-driven role-based access control optimization cuts permission abuse incidents by 40%.

3.2. AI aggravating security threats

AI-generated phishing emails boast twice the success rate of manually crafted ones, with one company suffering a 10 million yuan loss as a result; meanwhile, the Gemini-2 vulnerability case reveals that AI-mined zero-day vulnerabilities have lowered the attack threshold, leading to a 50% surge in "civilian hackers" [9]. Compounding these challenges, adaptive malware continues to evolve via reinforcement learning, pushing the failure probability of traditional defense methods to 30%, and the issue of resource asymmetry is stark—attackers leveraging open-source AI tools can inflict $100 million in losses on enterprises at a cost of just $100,000 [8,12].

4. Governance challenges and countermeasures

4.1. Technical governance

4.1.1. Trusted AI technology

Breakthroughs have been made in interpretability frameworks. The expansion of OpenAI METR evaluation increases the transparency of large models to 70%, and visual tools display decision paths, improving audit efficiency by 50% [3]. The optimization of adversarial training algorithms, such as the ShieldAgent framework , increases model robustness from 60% to 85% while maintaining the same inference speed [7].

DOI: 10.54254/2755-2721/2025.BJ27253

4.1.2. Encryption and privacy computing

Federated learning is widely applied in distributed training. A bank protects customer data through federated learning, with a model accuracy rate of 90% and no privacy leakage [13]. Homomorphic encryption mechanisms enable full encryption of data processing—a medical AI system, for instance, completes diagnoses in a ciphertext state, with the result accuracy consistent with plaintext processing [7]. The differential privacy technology roadmap is clear, and data release errors will be controlled within 5% by 2026.

4.1.3. Tool development

The Dioptra 2.0 open-source platform integrates risk assessment functionalities and can automatically generate attack graphs, assisting enterprises in identifying potential vulnerabilities and improving vulnerability remediation efficiency by 40% [14]. AI Security Posture Management (AISPM) tools reduce compliance costs by 30% through continuous monitoring [5].

4.2. Legal and standard construction

4.2.1. International practices

The EU's "General AI Practice Guidelines" establish an ethical review mechanism, mandating that high-risk AI systems pass third-party evaluations, otherwise they are prohibited from being listed [2]. The NIST AI Risk Management Framework adopts hierarchical classification management, categorizing AI systems into 4 risk levels and implementing differentiated supervision. The United Nations AI Security Initiative has promoted 120 countries to sign the "AI Security Code of Conduct", which explicitly prohibits lethal autonomous weapon systems [11].

4.2.2. China's path

The application scope of the "Cybersecurity Law" has been expanded to include generative AI services under regulatory oversight, requiring a 100% filing rate [9]. Revise the 2025 basic requirements for the security of generative AI services, clarify data security responsibilities, with a maximum fine of 50 million yuan for violations [15]. Local standards have also been formulated, such as the large model filing system, which mandates 100% coverage of training data compliance audits [15].

4.2.3. Comparative analysis

As shown in Table 2, significant differences exist in AI legislation among China, Europe, and the United States. The EU focuses on risk prevention, the United States emphasizes fair privacy, and China adheres to security and controllability [2]. The three parties have conflicts in cross-border data flow and algorithm transparency, but have reached preliminary cooperation intentions in combating AI cybercrime [11].

DOI: 10.54254/2755-2721/2025.BJ27253

Table 2. Comparison of AI legislation in China, Europe, and the United States (source: CCID Think Tank 2025)

Dimension	EU	United States	China
Core Principles	Risk Prevention	Fair Privacy	Security and Controllability
Regulatory Model	Full Lifecycle Supervision	Industry Self-discipline Oriented	Algorithm Filing + Large Model Filing
Penalty Intensity	Up to 6% of Global Turnover	Mainly Civil Compensation	Up to 50 Million Yuan

4.3. Ethical and social governance

4.3.1. Design accountability

Developer ethics guidelines have been formulated, expanding Meta's 5 responsibility pillars and requiring AI projects to pass an Ethical Impact Assessment (EIA framework), otherwise approval will not be granted [14]. A self-driving company was suspended from testing for 6 months due to failure to conduct ethical assessments [15].

4.3.2. Public cognition improvement

The Ministry of State Security released warning cases, which garnered 100 million views on short video platforms, increasing AI security awareness from 30% to 60% [6]. Media literacy training has reached 5 million users, improving the ability to prevent deepfake fraud by 40%.

4.3.3. Social inclusion

Pay attention to the impact of AI bias on ethnic minorities [16]. For instance, a recruitment AI system showed a 20% lower admission rate for a specific ethnic group than other groups due to training data bias. After fairness adjustment, the difference was reduced to 5% [16]. The protection mechanism for vulnerable groups is improved. A financial AI provides voice interaction support for visually impaired users, increasing usage rate by 30%.

4.4. Cross-border cooperation mechanisms

4.4.1. Joint testing platforms

The international open-source tool Dioptra 1.0 has been promoted, with 500 enterprises worldwide accessing it. Sharing attack signature databases increases vulnerability response speed by 50% [2,12]. The G7 AI Security Alliance has established joint testing standards. After passing the test, a cross-border payment system's security level was upgraded from B to A [11].

4.4.2. Threat intelligence sharing

To address AI-driven attacks by APT groups, a shared database containing 100,000 threat indicators has been established. A multinational enterprise avoided 15 million US dollars in losses through intelligence sharing. Joint drills have become normalized. In the 2025 "Pacific Storm" exercise,

DOI: 10.54254/2755-2721/2025.BJ27253

Chinese, American, and European teams collaborated to intercept an APT attack targeting energy networks [11].

4.4.3. Resolution of data sovereignty conflicts

Formulate strategies to resolve data sovereignty conflicts. A multinational company reduced compliance costs by 25% through localized deployment and cross-border data mirroring, enabling it to comply with both China's Data Security Law and the EU's GDPR [9].

5. Conclusion

AI information security requires a multi-dimensional governance framework involving technology, law, ethics, and international cooperation. China should accelerate the legislation of the "Artificial Intelligence Law" (phased implementation path), support "AI + security" technological innovation (such as building a national-level laboratory), and put forward international cooperation initiatives (such as the "Belt and Road" AI Security Alliance). This research has certain limitations, and future research can be expanded in aspects such as empirical experimental verification frameworks. This section discusses the technological evolution and governance innovation in AI security. It highlights the use of watermarking technology for content traceability, adaptive defense systems for real-time strategy adjustments, and the integration of AI and blockchain for decentralized training. Governance innovation includes agile regulation, real-time monitoring cases, and sandbox testing for AI projects. The data localization policy has led to increased efficiency and algorithm independence. A 2030 AI risk prediction model predicts that AI cybercrime losses will account for 3% of global GDP without strengthened governance, and policy recommendations include completing the "Artificial Intelligence Law" by 2027 and establishing a national AI security laboratory.

References

[1] Gartner Research Institute. (2025). Identifies the Top Cybersecurity Trends for 2025. Gartner Quarterly Report, 20(2), page 28-36.
[2] White & Case Legal Research Team. (2025). AI Watch: Global Regulatory Tracker - China. White & Case Law Journal, 8(1), page 41-50.
[3] ScienceDirect Editorial Team. (2025). Transforming cybersecurity with agentic AI to combat emerging threats. ScienceDirect Journal, 12(3), page 89-101.
[4] World Economic Forum. (2025). Artificial Intelligence and Cybersecurity: Balancing Risks and Rewards. WEF Global Report, Geneva, page 67-83.
[5] HackerNoon Editorial Team. (2025). AI Security Posture Management (AISPM). HackerNoon Website, page 15-22.
[6] Eliyahu, T., & Cohen, R. (2025). AI Security Newsletter - July, 2025. X Platform Blog, Retrieved from https: //x.com/aisecuritynewsletter/july2025, page 3-9.
[7] Naddeo, K., Smith, J., & Johnson, L. (2025). DICOM De-Identification via Hybrid AI and Rule-Based Framework for Scalable, Uncertainty-Aware Redaction. arXiv Preprint, page 18-25.
[8] Sharshar, M., Williams, A., & Brown, K. (2025). Large Language Model-Based Framework for Explainable Cyberattack Detection in Automatic Generation Control Systems. arXiv Preprint, page 32-40.
[9] Pathade, C. (2025). Invisible Injections: Exploiting Vision-Language Models Through Steganographic Prompt Embedding. arXiv Preprint, page 7-15.
[10] ICML 2025 Program Committee. (2025). MELON: Provable Defense Against Indirect Prompt Injection Attacks in AI Agents. Proceedings in ICML 2025 Conference, Seoul, 5(2), page 78-86.
[11] Ministry of Foreign Affairs of the People's Republic of China. (2025). Global AI Governance Action Plan. Chinese Foreign Policy Journal, 11(4), page 19-27.

DOI: 10.54254/2755-2721/2025.BJ27253

[12] Check Point Security Labs. (2025). AI Security Report, 2025. Check Point Technical Journal, 14(3), page 55-63.
[13] ICML 2025 Ethics Panel. (2025). Underestimated Privacy Risks for Minority Populations in Large Language Model Unlearning. Proceedings in ICML 2025 Conference, Seoul, 5(2), page 102-110.
[14] Stanford HAI Research Team. (2025). The 2025 AI Index Report. Stanford University Press, Stanford, page 112-130.
[15] Carnegie Endowment for International Peace. (2025). China's AI Policy at the Crossroads: Balancing Development and Governance. Brookings Institution Press, Washington D.C., page 56-72.
[16] Virtue AI Team, Lee, S., & Kim, H. (2025). ShieldAgent: Shielding LLM Agents via Verifiable Safety Policy Reasoning. Proceedings in ICML 2025 Conference, Seoul, 5(2), page 45-53.