

Artificial Intelligence in National Security: Applications, Risks, and Governance Pathways

Qinghan Ji

*Faculty of Humanities and Social Sciences, University of Nottingham Ningbo China, Ningbo, China
hiyqj1@nottingham.edu.cn*

Abstract. With the rapid development of artificial intelligence (AI), its applications have gradually expanded from the commercial and civilian sectors to national security fields such as military security, cybersecurity, and critical infrastructure protection. Using a literature review and case analysis approach, this article examines the main applications of AI in different security scenarios and the risks they bring. The study finds that AI can improve the information processing and response capabilities of the national security system by enhancing the efficiency of multi-source information processing, strengthening threat identification and prediction capabilities, and increasing the autonomy of military systems. However, enhanced AI capabilities do not necessarily lead to higher levels of security. Its national security applications face three interconnected risks: technical unreliability, malicious use, and strategic instability arising from the security dilemma. Therefore, this article argues that AI governance cannot rely solely on technological improvements. Instead, it requires a governance path comprised of technical safeguards, risk-based regulation, and international cooperation. This should be achieved through testing, auditing, human-in-the-loop mechanisms, and confidence-building measures (CBMs) to mitigate the risks posed by technological errors, malicious misuse, and strategic miscalculations. Overall, the key to AI governance is not restricting AI's entry into the national security domain, but rather establishing corresponding control, oversight, and international coordination mechanisms while enhancing its security capabilities.

Keywords: Artificial Intelligence, National Security, AI Governance, Security dilemma

1. Introduction

With the rapid development of AI technology, its application has gradually expanded from the commercial and civilian fields to the national security field. Traditional national security fields such as the military, cybersecurity, and critical infrastructure are increasingly introducing AI to improve information processing, threat identification, risk forecasting and autonomous execution capabilities, such as target recognition, uncrewed navigation, and mission coordination. For example, AI can improve the efficiency of Intelligence, Surveillance, and Reconnaissance (ISR) by processing large volumes of multi-source intelligence, and can also be applied to network threat detection, critical infrastructure monitoring and the autonomy of the military system. Therefore, AI has not merely

become an important technical tool to improve national security capabilities but also is progressively affecting the operation of the national security system.

However, the advancement of AI capabilities does not necessarily mean that national security will be more secure. The wide application of AI in the national security domain also brings risks that are different from those associated with traditional technologies. First, the AI system itself has technical limitations such as AI hallucinations, unreliability and brittleness, which may produce misjudgment when the data is insufficient or beyond the training range [1, 2]. They may generate erroneous outputs, which could lead to severe consequences in high-consequence applications such as ISR, early warning, and military decision-making. Secondly, AI can not only be used for defense and security protection, but also may be used by malicious subjects for network attacks, information manipulation and other behaviors, thus forming Malicious Use Risks [3, 4]. Furthermore, the development of AI capabilities may also affect the strategic interaction between countries. A country's development of AI to reinforce its own security may be regarded as a potential threat by other countries and prompt other countries to further improve their AI capabilities, thus forming an action-reaction cycle and strengthening the security dilemma [3]. Therefore, the national security problems brought about by AI are no longer limited to the reliability of a single technology system, but involve the three levels of technology, behavior subjects and strategic interaction between countries.

Existing studies have separately focused on the application of AI in military and cybersecurity, the technical risks inherent in AI systems, and the potential impact of AI on international strategic stability [2, 5, 6]. But these discussions often examined a single application scenario or specific risks. In contrast, it is necessary to put the application of AI in different national security fields, along with its technical, behavioral and strategic risks in the integrated analytical framework, and further discuss how to use technical safeguards, regulatory governance and international cooperation to respond to these risks. Especially when the AI risk is transnational, and there is obvious strategic competition between countries, it is difficult to fully solve the relevant problems by relying solely on technological improvement or the supervision of an individual country. Therefore, a comprehensive analysis of the application, risks and governance path of AI in the field of national security not only contributes to understanding how AI is reshaping the traditional national security system but also provides reference for the governance and international cooperation of high-risk AI.

This paper takes Artificial Intelligence in National Security as the research theme, and adopts the method of literature review and case analysis to systematically examine the major applications of AI in the national security domain, the resulting core risks and the corresponding governance path.

2. AI applications in national security

2.1. Military security

In the military field, AI has gradually entered different types of military application scenarios. Modern military operations will produce large volumes of multi-source data, including SIGINT, IMINT, HUMINT, that are difficult for traditional manual analysis to process quickly and accurately [7]. In contrast, AI can alleviate the limitations of the speed and scale of traditional manual analysis when processing data. For instance, Project Maven of the U.S. Department of Defense (DoD) employs computer vision and algorithms to analyze a substantial amount of drone footage, which can reduce the time required to process complex intelligence data and support faster identification of potential threats. By identifying patterns and locating areas of interest, it transforms the raw video data into information that can be utilized for intelligence assessment [7, 8]. Similarly, SitaWare and

JIMAPS in the Danish defense system use AI to integrate multi-source data from sensors, satellites and intelligence databases to enhance situational awareness and decision support [8].

Another significant application of AI in the military is autonomous systems. AI is strengthening the autonomous navigation, target recognition and mission execution capabilities of military platforms. For example, the U.S. Navy's Ghost Fleet Overlord project explores highly autonomous unmanned surface ships to reduce dependence on human crew; Britain's Brimstone air-launched missiles can independently search designated areas and identify predefined target types [8]. In general, these applications reflect two aspects of the role of AI in the military field: on the one hand, it improves the efficiency of information acquisition and analysis, and on the other hand, it enhances the autonomy of the military system, thus providing intelligent technical support for military operations.

2.2. Cybersecurity

Cybersecurity has become an essential component of national security because the country relies increasingly on digital networks, information systems and Critical infrastructure, which are also facing escalating cyber threats [4]. With the rapid growth of network-connected devices and the volume of security data, traditional methods that rely on manual monitoring and analysis are increasingly unable to process the wealth of information in real time. AI is thus gradually applied to threat detection, security monitoring and incident response [9].

AI can use machine learning and other technologies to analyze network traffic, system logs and diverse security data, find potential threats by identifying the differences between normal and abnormal behaviour, and detect unknown attacks that are difficult to identify with traditional security tools, including anomalous patterns that may signal zero-day threats [10]. For example, in endpoint security, AI can identify threats such as malware, ransomware and phishing through behavior-based threat detection [11].

Additionally, AI is promoting the automation of Cybersecurity operations. AI-powered systems can classify and prioritise numerous security alerts and automatically execute certain incident responses, which alleviate the burden on security analysts associated with repetitive tasks, thereby enabling them to allocate more attention to the analysis and remediation of complex threats [12].

This kind of application has gradually entered the practice of national security. For instance, the Israel Defense Forces (IDF) has applied AI to Cybersecurity and through the Artificial Intelligence and Big Data Research Centre develops AI algorithms for threat detection and response [8]. Therefore, the role of AI in Cybersecurity is no longer limited to single threat detection but gradually covers multiple links extending from continuous monitoring, threat identification to automated response.

2.3. Critical infrastructure protection

Critical infrastructure is a significant foundation for the country's social and economic operation, including energy, finance, transportation, communication and other systems. With the continuous digitization of these infrastructures, their operation generates a large amount of real-time data, so AI is used for infrastructure monitoring, prediction and risk management at an accelerating rate.

In the field of energy, AI can be applied to smart grids to carry out power load balancing, fault diagnosis and predictive maintenance by analyzing power grid operation data, so as to help operators identify potential faults in advance and optimize power supply [13]. Meanwhile, AI has also begun to be utilized for protecting smart power grids and smart meters from network attacks.

For example, AI-enabled Intrusion Detection Systems (IDS) can use the Machine Learning model to analyze attack data and identify Distributed Denial of Service (DDoS) attacks and abnormal activities [5]. In this way, artificial intelligence can detect faults and threats earlier, thereby enhancing the capabilities of infrastructure, enabling operators to respond more quickly, and reducing the likelihood that local outages will develop into wider infrastructure failures.

In the financial field, AI is mainly used to improve the security and risk identification ability of digital financial systems. For example, AI-supported biometric authentication can reduce the risk of unauthorized access, while real-time risk scoring enables financial institutions to assess transaction risks on an ongoing basis [9]. Furthermore, AI can also be used for fraud detection and prevention [13]. These applications show that AI is being used to improve the operational reliability and risk management capabilities of different types of critical infrastructure.

3. Risks of AI integration into national security

3.1. AI hallucinations

AI hallucinations are an innate structural defect of machine learning models, which is distinct from conventional software bugs that can be rapidly repaired. It means that the model output has high confidence, yet even without external malicious manipulation, it can still contain factual inaccuracies, fabricated references, or erroneous reasoning [2]. The underlying triggers of this problem include training dataset deficiencies and bias, model overfitting, and the inexplicable characteristics of deep learning black-box [1]. At the same time, the system is highly susceptible to being deliberately induced by adversarial examples to produce misjudgment.

When AI is deployed in high-risk national security systems such as military ISR platforms, autonomous weapons, and nuclear C2, AI Hallucinations will form security hazards of chain transmission. ISR-type AI may misjudge civilian facilities as hostile force deployment; autonomous killing systems will launch strikes based on distorted images; command and control AI will be enough to fabricate enemy mobilization and network attack clues out of thin air to output biased action plans to commanders; key industrial infrastructure monitoring AI There will also be problems such as false alarms and intrusions [1, 2]. When artificial intelligence is deployed in the field of high consequences, this kind of error is particularly worrying. In military and nuclear-related applications, the misidentification and lack of explainability of AI may weaken decision-makers' confidence in its output and increase the risk of strategic misjudgment [1].

3.2. Malicious use risks

AI can not only be used in legitimate scenarios such as network defense and people's livelihood services, but it will also be maliciously used by various subjects [3]. The malicious abuse of AI refers to hacker organizations, national hostile forces, criminal gangs misappropriating AI algorithms, deliberately creating network destruction, information misleading and other harmful behaviors.

For example, smart cities are facing increasing AI-based cyberattacks, including telecommunications fraud using deepfake-generated voice spoofing and breaching infrastructure security protocols using machine-learning-generated adversarial attacks [14]. Because of the highly interconnected nature of telecommunications networks and smart cities, cybercriminals and state-sponsored actors may target them, using artificial intelligence technologies and AI-enabled

malicious code to find security vulnerabilities, steal sensitive information, or suspend smart city services [9, 14].

The malicious use of AI may also extend to information operations. AI can reduce the cost of generating and spreading misinformation and expand social differentiation by generating content for specific groups. After Microsoft's Tay chatbot was launched in 2016, a large amount of racist, sexist and extremist content was entered by users, and then it began to generate similar harmful remarks, which were finally taken offline by Microsoft [3]. This incident shows that AI may be manipulated by malicious actors and used to spread harmful information. In the political environment, if similar capabilities are used on a large scale, it may further expand the impact of information manipulation, exacerbate polarization and social instability, thus creating national security risks.

3.3. The security dilemma

The security dilemma refers to the measures taken by a country to enhance its own security, which may be regarded as a threat by other countries, thus prompting the other party to take further security measures and eventually forming mutually enhanced competition. The Application of AI in national security is exacerbating this mechanism [3]. The United States has regarded China as an important competitor in shaping its national defense and security strategy and has turned its concern over China's AI development into accelerating AI investment, research, development, acquisition and deployment [8]. Simultaneously, some studies estimate that the annual investment of the Chinese military in AI is about 1.6–2.7 billion USD, which is roughly equivalent to the AI investment of the U.S. military [15]. This implies that the growth of one country's AI capacity may prompt another country to further develop its own capabilities and form an action-reaction cycle.

In addition, AI has obvious dual-use characteristics. The same technology can be used for defence, intelligence analysis and infrastructure protection, and may also be used for surveillance, cyber operations or military engagement. Therefore, even if a country positions the development of AI as a defensive capability, it may be difficult for other countries to judge its real use and regard it as a potential offensive capability and then take corresponding preventive and development measures [1]. This mutual misjudgment of ability and intention means that a country's effort to strengthen its own security through AI development may reduce the perceived security of other countries. This dynamic can further intensify the security dilemma.

4. Governance challenges and responses

4.1. Enhancing AI reliability and technical safeguards

In response to AI Hallucinations and their possible wrong decisions, the primary goal of governance is not to pursue the "zero error" of AI, but to improve its reliability, robustness and controllability in complex and uncertain environments. In the process of training AI (reducing Hallucinations), it is necessary to enter a large amount of data to help AI establish a correct machine learning system. However, the lack of data has made the process difficult, especially in the military field, which is important for national security. Unlike many commercial AI applications that can draw on large volumes of continuously generated data, military AI often operates in data-constrained environments, such as nuclear attack warning, enemy behaviour tracking and other rare events, themselves are difficult to provide sufficient training data; when the model is applied to a dynamic environment beyond the original training situation, its performance may also decline [6]. That is, AI's machine learning system is easy to show unreliability and brittleness when the data is

insufficient or beyond the design range, resulting in the situation of "good performance in the laboratory and serious failure in the real environment" [6]. Accordingly, the AI system should be improved through more rigorous testing and evaluation. In accordance with the NIST AI RMF, risk management should be implemented throughout the AI lifecycle, following the GOVERN, MAP, MEASURE, and MANAGE framework [16].

4.2. Strengthening regulatory and institutional governance

Relying only on AI developers' own technical safeguards and voluntary security commitments, it is difficult to cope with the high-risk consequences that AI may have in the field of national security. Thus, it is necessary to establish a regulatory framework with risk-based regulation as the core, and to impose more stringent testing, auditing, and safety standards for high-risk AI systems. Amodoi pointed out that with the rapid development of AI capabilities, supervision needs to consider the future development of technology, not solely current capabilities, and suggests the establishment of a testing and auditing regime for advanced AI systems [17]. For high-risk applications in the national security domain, this suggests that they should be strictly evaluated before the system is deployed, and the regular checkpoints should be established to continuously monitor potential risks in areas such as cybersecurity and nuclear technology [17]. Risks, so as to move from developer self-regulation to institutionalized external oversight.

On this basis, public institutions such as NIST should promote the measurement, evaluation and testing research of AI to provide a more reliable technical foundation for regulatory standards [17]. Meanwhile, regulation should not only focus on the technical performance of the AI system itself but also clarify the accountability and control relationships between AI and human decision-makers, the AI risk governance also necessitates determining the appropriate human-AI configuration according to specific application scenarios [16]. For high-consequence applications such as ISR, early warning and military decision-making, AI should not completely replace human judgment [6]. Considering the possible vulnerability of automation bias and AI under novel conditions, the appropriate human-in-the-loop mechanism should be retained so that human decision-makers can verify, interpret, and override AI outputs. Thus reducing the risk of direct transformation of technical errors into national security decision-making [6].

4.3. Promoting international cooperation and confidence-building measures

AI risks are obviously transnational, and it is difficult for a single country's regulation to cope with the risks brought about by cross-border technology diffusion, malicious use and AI competition among major countries. Zeng pointed out that AI security problems cannot be fully solved through unilateral governance, and the continuous AI capability competition may further strengthen the security dilemma [18]. Therefore, countries need to reduce the uncertainty and strategic misjudgment in the development of AI through international cooperation.

International cooperation can start with specific risk assessment and information sharing. The AI Safety Institutes network promoted by the British government provides a cooperation platform for countries to share AI safety research, testing methods and risk information [19]. On this basis, countries can further carry out joint risk assessment, information sharing and crisis communication to improve the transparency of high-risk AI capabilities and their potential risks and reduce the possibility of misjudgment and unexpected escalation through confidence-building measures (CBMs).

However, current AI governance is still limited by differences in geopolitical rivalry and governance concepts, so it is unrealistic to establish unified global mandatory rules in the short term. The Global Artificial Intelligence Governance Initiative emphasizes that more open and inclusive multilateral governance should be used to urge countries to take part in AI governance [20]. Therefore, a more feasible path is to adopt a gradual cooperation method for specific issues, giving priority to high-risk areas such as military AI, nuclear-related AI and autonomous systems. Carry out cooperation such as risk assessment, information sharing and crisis communication in these fields. This would gradually build mutual trust and reduce the risks of security dilemmas and conflict escalation induced by AI.

5. Conclusion

The analysis in this article shows that the changes brought about by AI entering the field of national security are not merely the addition of a new technological tool, but also that it has changed the way national security capabilities are formed, risks and strategic interaction between countries to a certain extent. AI can improve the efficiency of information processing and threat identification and expand the autonomy of the military system. But these capabilities also increase the dependence of the national security system on the reliability of the AI system. The resulting problem is not a simple technical error, but the technical output may further enter into intelligence assessment, security response and strategic decision-making, and further forming broader safety consequences.

The unreliability of the technical system may first cause misinformation or misjudgment; AI may also be used by different subjects for malicious purposes; when these capabilities are placed in a competitive environment between nations, the mutual assessments of AI capabilities and strategic intentions among different countries may further amplify the security dilemma. Thus, the risk of AI in national security cannot be completely solved by simply improving the performance of the model. Instead, it is necessary to pay attention to the three levels of technology operation, behavior subject and interaction between countries. This article believes that the focus of AI governance is not to simply limit the application of AI in the field of national security, but to establish a control, supervision and coordination mechanism that matches its potential risks while allowing it to play its security value.

From the perspective of governance practice, technical safeguards can reduce the possibility of errors and failures in AI systems, while regulatory governance can institutionalize testing, auditing, and human oversight. International cooperation can further reduce information asymmetry and strategic miscalculation between countries. Especially for high-risk fields such as military AI, nuclear-related AI and autonomous systems, incremental and issue-specific cooperation is more realistic and feasible than pursuing unified global rules in the short term. This means that future AI governance may need to gradually shift from singular technology security management to a multi-layered governance approach that combines technology, institutions and international coordination.

The significance of this article is to link the technical applications and security risks of AI with the government responses, demonstrating that the impact of AI on national security cannot be judged solely from the perspective of technological capabilities; it also requires understanding in conjunction with the actions of actors and strategic interactions between states. This analysis may provide some reference for the risk management of high-risk AI and related international cooperation. However, this article mainly relies on literature review and case analysis, lacks empirical testing of specific AI system performance and governance measures, and the comparison of governance models of different countries is still limited. Future research could combine empirical data, comparative case studies, and testing results of specific AI systems to further evaluate the

actual effects of different governance mechanisms and explore the impact of AI capability competition on long-term strategic stability and international security governance.

References

- [1] Johnson, J. (2023). *AI and the bomb: Nuclear strategy and risk in the digital age*. Oxford University Press.
- [2] Frunzeti, T., Tomoiaga, T., & Coşoreanu, L. (2025). Effects of AI hallucinations on military systems. *Ann. Ser. Mil. Sci.*, 17(2), 13–37. <https://doi.org/10.56082/annalsarscimilit.2025.2.13>
- [3] Burton, J., & Soare, S. R. (2019). Understanding the strategic implications of the weaponization of artificial intelligence. In *2019 11th International Conference on Cyber Conflict (CyCon)*, Tallinn, Estonia, 1–17. <https://doi.org/10.23919/CYCON.2019.8756866>
- [4] Kim, G., & Park, K. (2024). Effect of AI: The future landscape of national cybersecurity strategies. *Tech. Gaz.*, 18(1), 29–36. <https://doi.org/10.31803/tg-20230218142012>
- [5] Garcia, M. G., & Kumar, S. (2025). AI-enabled intrusion detection of cyber security flood attacks on smart meters. In *2025 IEEE 6th Annual World AI IoT Congress (AIIoT 2025)*, Seattle, WA, USA, 567–573. <https://doi.org/10.1109/AIIoT65859.2025.11105336>
- [6] Horowitz, M. C., & Scharre, P. (2021). AI and international stability: Risks and confidence-building measures. Center for a New American Security. <https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures>
- [7] Mikhailov, D. I. (2023). Optimizing national security strategies through LLM-driven artificial intelligence integration. <https://doi.org/10.48550/arXiv.2305.13927>
- [8] Borchert, H., Schütz, T., & Verbovszky, J. (2024). *The very long game: 25 case studies on the global state of defense AI*. Springer. <https://ebookcentral.proquest.com/lib/nottingham/detail.action?docID=31554433>
- [9] Goffer, M. A., et al. (2025). AI-enhanced cyber threat detection and response advancing national security in critical infrastructure. *J. Posthumanism*, 5(3), 1667–1689. <https://doi.org/10.63332/joph.v5i3.965>
- [10] Sarker, I. H., Furhad, M. H., & Nowrozy, R. (2021). AI-driven cybersecurity: An overview, security intelligence modeling and research directions. *SN Comput. Sci.*, 2(3), 173. <https://doi.org/10.1007/s42979-021-00557-0>
- [11] Khan, M. I., Arif, A., & Khan, A. R. A. (2024). The most recent advances and uses of AI in cybersecurity. *Bullet*, 3(4), 566–578. <https://journal.mediapublikasi.id/index.php/bullet/article/view/4540>
- [12] Hofstetter, M., Riedl, R., Gees, T., Schaberreiter, T., Quirchmayr, G., et al. (2020). Applications of AI in cybersecurity. In *2020 Second International Conference on Transdisciplinary AI (TransAI)* (pp. 85–92). IEEE. <https://doi.org/10.1109/TransAI49837.2020.00031>
- [13] Laplante, P., Milojicic, D., Serebryakov, S., & Bennett, D. (2020). Artificial intelligence and critical systems: From hype to reality. *Computer*, 53(11), 9237327. <https://doi.org/10.1109/MC.2020.3006177>
- [14] Al-Turjman, F., Zahmatkesh, H., & Shahroze, R. (2022). An overview of security and privacy in smart cities' IoT communications. *Trans. Emerg. Telecommun. Technol.*, 33, e3677. <https://doi.org/10.1002/ett.3677>
- [15] Fedasiuk, R., Melot, J., & Murphy, B. (2021). Harnessed lightning: How the Chinese military is adopting artificial intelligence. Center for Security and Emerging Technology, Georgetown University. <https://doi.org/10.51593/20200089>
- [16] National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0). U.S. Department of Commerce. <https://doi.org/10.6028/NIST.AI.100-1>
- [17] Amodei, D. (2023). Written testimony for a hearing on oversight of A.I.: Principles for regulation. U.S. Senate Committee on the Judiciary. https://www.judiciary.senate.gov/imo/media/doc/2023-07-26_-_testimony_-_amodei.pdf
- [18] Zeng, Y. (2023). Chinese perspectives on AI safety. *Chinese Perspectives on AI*. <https://chineseperspectives.ai/yi-zeng>
- [19] Department for Science, Innovation and Technology. (2024, May 21). Global leaders agree to launch first international network of AI safety institutes. *GOV.UK*. <https://www.gov.uk/government/news/global-leaders-agree-to-launch-first-international-network-of-ai-safety-institutes-to-boost-understanding-of-ai>
- [20] Embassy of the People's Republic of China. (2023). Global AI governance initiative. https://gd.china-embassy.gov.cn/eng/zxhd_1/202310/t20231024_11167412.htm