An Artificial Intelligence Security Framework

Huiyun Jing¹, Wei Wei¹, Chuan Zhou² and Xin He³*

¹ China Academy of Information and Communications Technology, Beijing, China
² Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
³ National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing, China

* Corresponding Author: hexin@cert.org.cn

Abstract. With the overall acceleration of the scale construction and application of artificial intelligence worldwide, the security risks of artificial intelligence infrastructure, design and development, and integration applications are becoming increasingly prominent. Major countries have developed AI security governance by formulating AI ethical norms and improving laws and regulations and industry management. Artificial intelligence security technology system is an essential part of artificial intelligence security governance, critical support for implementing artificial intelligence ethical norms, meeting legal and regulatory requirements. It is also an important guarantee for the artificial intelligence industry's healthy and orderly development. This article in view of the problem of lacking artificial intelligence security framework worldwide, focusing on the current artificial intelligence prominent security risks, proposes an AI security framework that covers AI security goals, graded capabilities of AI security, and AI security technologies and management systems. We look forward to providing useful references for the community to improve the safety and protection capabilities of artificial intelligence.

1. Artificial Intelligence Security Risks

Combined with the whole process of AI system design and operation, we analyze the security risks AI system faces in each life cycle stage in detail, which will help diagnose and locate risk sources, research and deploy targeted security defense theories and technologies. The International Organization for Standardization (ISO) has launched a standard project called "Artificial Intelligence System Lifecycle Process," which summarises the whole life cycle of artificial intelligence system into eight stages: Inception, Design and Development, Verification and Validation, Deployment, Operation and Monitoring, Continuous Validation, Re-evaluation, and Retirement. Based on the ISO division of the artificial intelligence system's full life cycle, we draw the security risk map of artificial intelligence, as shown in Figure 1.
The Security Risk Map of Artificial Intelligence

Figure 1. The Security Risk Map of Artificial Intelligence.
Inception stage security risks. The Inception stage refers to transforming ideas into tangible systems, mainly including task analysis, requirement definition, risk management design, and other parts. In this stage, the security risk is manifested primarily that the application goals of artificial intelligence violate national laws and regulations and social ethics.

Design and Development stage security risks. The Design and Development stage refers to completing the creation of a deployable artificial intelligence system, including the determination of design methods, the definition of system framework, the writing of software code, risk management, and other processes. In this stage, the security risks are mainly manifested as the security risks caused by the defective artificial intelligence infrastructure, technical vulnerability[1,2], and the wrong design and development.

Verification and Validation stage security risks. The Verification and Validation stage checks whether the AI system works according to the expected requirements and fulfills the predetermined goals. In this stage, the security risks are mainly manifested as insufficient test verification, and failure to timely find and repair the security risks in the pre-sequence stage.

Deployment stage security risks. The Deployment stage refers to the process of installing and configuring an AI system in the target environment. The security risks at this stage are mainly manifested in the untrustworthy hardware and software environment of the deployed artificial intelligence system, and the system may be subject to unauthorized access and use[3,4].

Operation and Monitoring stage security risks. The Operation and Monitoring stage mainly includes the operation and monitoring, maintenance, and upgrading of the artificial intelligence system. The security risks in this stage are mainly manifested as the security attacks launched by malicious attackers against artificial intelligence systems, such as adversarial attack[5], backdoor attack[6], model steal[7], model feedback misguidance, model inversion attack[8], member inference attack[9], code vulnerability exploitation, as well as the abuse or malicious use of artificial intelligence system.

Continuous Validation stage security risks. In the Continuous Validation stage, continuous validation is carried out for the constant learning artificial intelligence system. The security risks are mainly reflected in the delay in updating the test verification data and the failure to timely discover and repair the misguidance of model feedback introduced by continuous learning.

Re-evaluation stage security risks. When initial goals are not met or need to be modified, the Re-evaluation stage is started. This stage mainly includes design redefinition, requirements redesign, risk management redesign, and other processes. Its security risks are similar to those in the inception stage: artificial intelligence's application goals violate national laws and regulations and social ethics.

Retirement stage security risks. In the Retirement stage, AI systems whose usage purpose no longer exists or have a better alternative are destroyed, including data, algorithm model, and the whole system. The security risks at this stage are mainly reflected in incomplete destruction and disclosure of personal privacy[10].

2. Artificial Intelligence Security Framework

2.1. Design Thought

2.1.1. Scope of Application
This framework focuses on artificial intelligence endogenous security, mainly to solve security risks faced by artificial intelligence infrastructure, design and development. This framework also focuses on solving security risks which make the artificial intelligence application behavior decision-making out of control directly caused by the security problems mentioned above. For the physical world and social security risks caused by the abuse or malicious use of artificial intelligence applications, the users are mainly regulated by national laws and regulations and industry regulatory policies.
2.1.2. Core elements

Based on the principles of practicability, foresight, and integrity that the artificial intelligence security framework should follow, the artificial intelligence security framework is constructed from the following three aspects.

First, clearing AI security goals are a prerequisite. This framework puts forward the artificial intelligence security goals by comprehensively analyzing the security risks faced by artificial intelligence applications. It points out the direction for implementing artificial intelligence security protection work.

Second, building AI security capabilities is key. To achieve the artificial intelligence security goals, we take the construction of artificial intelligence security capability as the guidance and refer to the sliding scale of cyber security[11], and put forward the graded superposition evolution model of artificial intelligence security capability.

Third, the deployment of safety technical measures and the implementation of safety management are important guarantees. To help artificial intelligence application R&D and operation enterprises effectively form and continuously improve artificial intelligence security capabilities, we propose an artificial intelligence security technology system and management system supporting artificial intelligence security capabilities.

In summary, the construction of an artificial intelligence security framework includes four dimensions: security goals, security capabilities, security technology, and security management, and guides companies to carry out artificial intelligence security protection work from four different levels.

2.2. Security framework

The artificial intelligence security framework includes four dimensions of security goals, security capabilities, security technologies and security management, as shown in Figure 2. These four protection dimensions guide enterprises to build artificial intelligence security protection system in a top-down, hierarchical way. Among them, setting reasonable security goals is the starting point and foundation to guarantee the security of artificial intelligence applications, security ability is the effective guarantee to realize the security goals, and security technology and safety management are the support and embodiment of security ability.
Figure 2. The Artificial Intelligence security framework.
2.2.1. **security goals** By systematically analyzing the security risks faced by artificial intelligence and their causes, we put forward the security requirements and goals of artificial intelligence from six aspects: application, function, data, decision-making, behavior, and incident.

2.2.2. **Security capabilities** According to the stepwise increasing difficulty of security capability construction and the stepwise decreasing input-output ratio of security resources, and referring to the sliding scale of cyber security, we propose five AI security capabilities: architecture security, passive defense, active defense, threat intelligence, and offense. The former level of security capability is the foundation of building the subsequent level of security capability. Among them, Architecture Security aims to guide enterprises to build the ability to plan, design, establish and upkeep artificial intelligence applications with security in mind. Passive Defense is designed to guide enterprises to deploy static, passive security capabilities beyond AI applications. Active Defense aims at driving enterprises to strengthen artificial intelligence security teams and realize dynamic, adaptive, and self-growing security capabilities. Threat Intelligence is designed to guide enterprises in acquiring and using AI security threat intelligence to enable AI security systems, equipment, and personnel. Offense aims at teaching enterprises to build a legal offense capability against artificial intelligence malicious attackers.

2.2.3. **Security technologies** Artificial intelligence application, algorithm, training data and framework platform are the four core components of artificial intelligence application construction, and are also the key objects of artificial intelligence security protection. Therefore, this framework provides security protection technology means for applications, algorithms, data, and platforms.

2.2.4. **Security management** Based on national and industrial artificial intelligence security laws and regulations, industry policies, ethical norms, technical standards, and other requirements, this paper puts forward the implementation requirements of enterprises in the aspects of the artificial intelligence security organization, personnel, and system.

3. **Artificial Intelligence Framework Analysis**

3.1. **security goals**

At present, the European Union, the United States, China, and other major countries in the world and technology giants such as Microsoft and Google have proposed ethical guidelines for AI. Among them, legality, reliability, controllability, fairness, traceability, privacy security, and other security goals become the focus of AI ethical standards. Based on fully drawing lessons from the ethical standards of artificial intelligence, this security framework puts forward the following six security goals based on the security risks and challenges faced by artificial intelligence and the actual needs of artificial intelligence application.

- **Legal and compliant applications**: Artificial intelligence has shown excellent capabilities in areas such as transportation and health care. The misuse or malicious use of AI applications will substantially negatively impact the physical world and the nation and society. Therefore, it is necessary to ensure that the artificial intelligence system's application goal conforms to the requirements of national laws and regulations and social ethics.

- **Reliable and controllable functions**: Artificial intelligence technology is gradually applied to intelligent medical care, autonomous driving and other safety-critical scenarios, and the robustness and reliability of artificial intelligence is increasingly important. However, new security attacks such as adversarial attack[1], and backdoor attack[2] can induce artificial intelligence applications to produce unexpected false outputs by modifying input data in real time. Therefore, it is necessary to ensure that the artificial intelligence system's functions always produce the expected behaviors and results within the specified operating conditions and time cycle and are always under the control of human operators.

- **Safe and reliable data**: Data is the cornerstone of artificial intelligence. While AI learns knowledge from data, it is also faced with many security risks such as data leakage[12], data bias, data
poisoning[13], etc. Therefore, it is necessary to ensure that the data collected, used, and stored by artificial intelligence applications will not be stolen, will not leak users' privacy, and will not be tampered with, which can truly reflect the physical world and human society.

**Fair and just decision-making:** Artificial intelligence applications such as intelligent risk control and intelligent recruitment are gradually assisting or even replacing humans in making critical decisions. Reasons such as unbalanced training data and wrong algorithm design may lead to biased decisions in AI applications, which will damage national social equity and justice. Therefore, it is necessary to ensure that various groups' characteristics should be taken into account in the artificial intelligence application. No biased decisions should be made against specific people or groups.

**Explanable behavior:** The "inexplicable" nature of AI algorithms such as deep neural networks makes it impossible to explain why an algorithm makes a decision and to understand its internal operating principles and position the problems. The interpretability of artificial intelligence guides diagnosing, discovering, and repairing the inherent defects of the algorithm model, which is the basis of artificial intelligence security. Therefore, it is essential to ensure that AI applications provide rational and accurate explanations of their actions and results in a way that humans can understand.

**Traceable incidents:** The "inexplicable" nature of AI algorithms brings challenges to analyzing traceability factors such as the causes of artificial intelligence security incidents and the relevant entities, and traditional security audit methods are not competent. Therefore, AI applications should be tailored to business scenarios, improve the traceability system, and deploy technical measures to effectively track and trace the causes, links, and relevant entities of security incidents.

### 3.2. Security capabilities

At present, enterprises mainly focus on AI application technology research and development and business operation. With little investment and a weak foundation in AI security, they cannot complete all security capacity building in a short period. Moreover, AI security still belongs to the frontier innovation field, and the systematic reduction of AI application security risk still needs the continuous breakthrough of security theory and technology. To effectively guide enterprises to improve AI security capability step by step, this security framework proposes a graded capability model of AI security.

Ensuring AI applications' endogenous security and effectively preventing new security risks of AI are the main scope and core objectives of this AI security framework. The sliding scale of cyber security[4] referred to in this framework is mainly oriented to traditional information systems. The specific security capabilities specified at each level do not apply to artificial intelligence applications with significant differences in technical characteristics and security risks from traditional information systems. Therefore, based on the core ideas of the sliding scale of cyber security, this framework puts forward a graded and superimposed evolution model of artificial intelligence security capabilities and systematically plans the artificial intelligence security capabilities at all levels.

#### 3.2.1. Architecture security

Architecture security refers to the ability to plan, design, establish and upkeep artificial intelligence applications with security in mind to improve their endogenous security, which mainly includes the following five aspects.

**Compliance assessment:** In the initial demand analysis stage, combined with specific application scenarios, the goals and implementation method of artificial intelligence application are assessed whether they conform to national laws and regulations, industry regulatory policies, and ethical norms.

**Service security guarantee:** Deploy access control, security isolation, security fuse, security redundancy, security monitoring, and other mechanisms in the application layer of artificial intelligence applications to ensure the safe operation of artificial intelligence applications in the incident of security attacks and other emergencies. For example, we can deploy a hierarchical pullback mechanism in autonomous driving cars to ensure that the vehicle's control is returned to humans on time in a dangerous situation.
Enhance algorithm security: Improve algorithm robustness, interpretability, and fairness by improving algorithm training method, adjusting algorithm model structure, etc. For example, it can enhance the algorithm's robustness by employing adversarial training[5] and model regularization[17].

Data security improvement: Improve the confidentiality and availability of data through data privacy computing, problem data cleaning, etc. For example, data confidentiality can be enhanced through differential privacy[18], homomorphic encryption[19], federated learning[20], and other technologies.

DL framework security detection and repair: Conduct security detection on the third-party pre-training model and open-source platform of machine learning, and timely repair the found security problems to perceive risks in advance and reduce the probability of security incidents. MindSpore[21], for example, has a well-established vulnerability management process that allows it to respond quickly to new security vulnerabilities submitted.

3.2.2. Passive defense
Passive defense refers to deploying static and passive security capabilities on the artificial intelligence application to resist new security attacks against artificial intelligence, mainly including the following three aspects.

Malicious behavior discovery: Through analysis and refinement of new security attacks and malicious application behavior characteristics against artificial intelligence, we conduct real-time detection of external access, input data, behavioral decision-making, etc. of artificial intelligence applications, and timely discover security attacks and malicious behaviors such as adversarial examples[5], algorithm backdoors[6], model steal[7], and deepfake[22].

Algorithm security protection: Deploy security protection components outside the artificial intelligence algorithm model, and help artificial intelligence applications effectively resist algorithm security attacks such as adversarial attack[5] and model steal[7] using algorithm intellectual property protection, problem data reconstruction, algorithm security evaluation, and other measures. For example, using the problem data reconstruction technology, under the condition of retaining the original image semantics as much as possible, destroy the disturbance added by the attacker maliciously to achieve the purpose of defense against the adversarial examples.

Data security protection: Deploy security protection components outside of artificial intelligence applications, and help artificial intelligence applications more effectively resist data security attacks such as training data poisoning, model inversion attack, and membership inference attack[8] through data tracking and traceability, data security evaluation, and other measures. For example, using the data security tag technology, the maliciously tampered data can be found in time to defend against the training data poisoning attack.

3.2.3. Active defense
Artificial intelligence security attack and defense technology are in rapid evolution, and passive security defense is difficult to deal with the continually innovating security attack means effectively. To make up for the limitation of static passive defense, this framework introduces active defense to introduce and strengthen the artificial intelligence security team strength and realize the dynamic, adaptive, and self-growing security capability, which mainly includes the following four aspects.

Continuous safety monitoring: During the operation of an artificial intelligence application, artificial intelligence security experts are used to monitor the application's operation continuously and safety status, give the current security risk level of the application and give a timely warning for abnormal operation of the application.

Security event analysis: In the incidents of data leakage and behavior loss of control in artificial intelligence applications, artificial intelligence security experts are introduced to analyze and judge the impact range, severity, and cause of the incidents in time.
Security incident response: When a security incident occurs, we shall comprehensively utilize various security defense technologies to respond and deal with the incident on time to restore artificial intelligence applications' regular operation.

Security threat prediction: Use artificial intelligence, big data analysis, and other technologies to realize the perception and prediction of unknown security threats from historical data.

3.2.4. Threat intelligence
Make full use of threat intelligence information to enhance further and expand the effectiveness of the active defense. Threat intelligence refers to the acquisition and use of artificial intelligence security threat intelligence, enabling artificial intelligence security systems, equipment, and personnel, mainly including the following three aspects.

Intelligence management: Artificial intelligence security experts use various technical measures to complete the comprehensive management such as acquisition, sorting, analysis, rating, and classification of threat intelligence.

Intelligence consumption: Artificial intelligence security experts comprehensively use threat intelligence to mine unknown threats, update system defense strategies and enhance security equipment capabilities.

Intelligence generation: AI security experts use a combination of technical measures to analyze and acquire knowledge about security risks and threats from a variety of public data sources.

3.2.5. Offense
Offense refers to the legal offense capability against artificial intelligence's malicious attack, which mainly includes the following two aspects.

Safety incident tracing: When a security incident occurs, ensure that the incident can be traced back to the relevant entities to support the subsequent legal protection.

Protection of legal rights and interests: For self-defense, use legal means to counter attackers.

3.3. Security technology
Recently, local breakthroughs have been made in artificial intelligence security, such as algorithm robustness enhancement[23], interpretation enhancement[14,15,16], data privacy computing[24], security attack detection, security attack defense, etc. They can support realizing the security capability of the initial two levels of architecture security and passive protection in the graded capability model of artificial intelligence security. However, the AI security theories and technologies required for the three levels of active defense, threat intelligence, and offense still need to be further innovated by academia and industry.

3.3.1. Application security technology
Application security technology refers to the security defense technology deployed in the AI application layer. Application security technology mainly includes four aspects: application compliance assessment, security attack detection, application security mechanism, and malicious use detection.

3.3.2. Algorithm security technology
Algorithmic security technology refers to the security defense technology for the deployment of artificial intelligence algorithms. Algorithm security technology mainly includes five aspects: algorithm robustness enhancement[23], algorithm fairness guarantee, algorithm interpretability improvement[14,15,16], algorithm intellectual property protection, and algorithm security evaluation.

3.3.3. Data security technology
Data security technology refers to the security defense technology for the deployment of artificial intelligence training data. Data security technology mainly includes five aspects: data privacy
computing[24], data provenance, problem data cleaning[25], data fairness enhancement, and data security evaluation.

3.3.4 Platform security technology
Platform security technology refers to the security defense technology deployed against the machine learning framework platform. The platform security technology mainly includes vulnerability mining and repair, model file check, and framework platform security deployment.

3.4 Safety management
In the artificial intelligence safety management system, the national government plays a leading role, leading the establishment of management agencies, developing laws and regulations, regulating regulatory policies, the development of technical standards, and other aspects. Based on fully understanding and abiding by artificial intelligence safety management rules, enterprises should continuously improve the organizational configuration, system process, and personnel capacity of artificial intelligence safety management. This paper focuses on the enterprise-level safety management system.

4 Conclusion
This article has summarized the increasingly severe security risks of artificial intelligence in infrastructure, design and development, and integrated applications worldwide. It has drawn a map of artificial intelligence security risks, which depicts the sources and performance of artificial intelligence security risks in each AI's full life stage. However, the lack of a global artificial intelligence security framework has made it impossible to support artificial intelligence security governance and promote the artificial intelligence industry's development. Focusing on the current artificial intelligence prominent security risks, we have proposed an AI security framework covering AI security goals, graded capabilities of AI security, and AI security technologies and management systems. We hope to provide useful references for the community to improve artificial intelligence's safety and protection capabilities.

Acknowledgements
This work is supported by the National 242 Information Security Program (No. 2018Q39).

References
[1] Xiao Q, Li K, Zhang D, Xu W. Security risks in deep learning implementations. In: 2018 IEEE Security and Privacy Workshops (SPW). IEEE; 2018. p. 123–128.
[2] Pei K, Cao Y, Yang J, Jana S. Deepxplore: Automated whitebox testing of deep learning systems. In: proceedings of the 26th Symposium on Operating Systems Principles; 2017. p. 1–18.
[3] Zhang T, He Z, Lee RB. Privacy-preserving machine learning through data obfuscation. arXiv preprint arXiv:180701860. 2018.
[4] Xu K, Cao T, Shah S, Maung C, Schweitzer H. Cleaning the null space: A privacy mechanism for predictors. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 31; 2017.
[5] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, et al. Intriguing properties of neural networks. arXiv preprint arXiv:13126199. 2013.
[6] Geigel A. Neural network trojan. Journal of Computer Security. 2013; 21(2): 191–232.
[7] Tramer F, Zhang F, Juels A, Reiter MK, Ristenpart T. Stealing machine learning models via prediction APIs. In: 25th {USENIX} Security Symposium ({USENIX} Security 16); 2016. p. 601–618.
[8] Shokri R, Stronati M, Song C, Shmatikov V. Membership inference attacks against machine learning models. In: 2017 IEEE Symposium on Security and Privacy (SP). IEEE; 2017. p. 3–18.
[9] Fredrikson M, Lantz E, Jha S, Lin S, Page D, Ristenpart T. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In: 23rd {USENIX} Security Symposium ({USENIX} Security14); 2014. p. 17–32.

[10] Chen D, Zhao H. Data security and privacy protection issues in cloud computing. In: 2012 International Conference on Computer Science and Electronics Engineering. vol. 1. IEEE; 2012. p. 647–651.

[11] RobertM.Lee. The Sliding Scale of Cyber Security [EB/OL]. https://www.sans.org/reading-room/whitepapers/analyst/membership/36240, 2015-8.

[12] Yeom S, Giacomelli I, Fredrikson M, Jha S. Privacy risk in machine learning: Analyzing the connection to overfitting. In: 2018 IEEE 31st Computer Security Foundations Symposium (CSF). IEEE; 2018. p. 268–282.

[13] Jagielski M, Oprea A, Biggio B, Liu C, Nita-Rotaru C, Li B. Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. In: 2018 IEEE Symposium on Security and Privacy (SP). IEEE; 2018. p. 19–35.

[14] Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. Explaining explanations: An overview of interpretability of machine learning. In: 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA). IEEE; 2018. p. 80–89.

[15] HUA Y, ZHANG D, GE S. Research progress in the interpretability of deep learning models. Journal of Cyber Security. 2020;5(3):1–12.

[16] Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: European conference on computer vision. Springer; 2014. p.818–833.

[17] Talvitie E. Model Regularization for Stable Sample Rollouts. In: UAI; 2014. p. 780–789.

[18] Dwork C. Differential privacy: A survey of results. In: International conference on theory and applications of models of computation. Springer; 2008. p. 1–19.

[19] Gentry C, et al. A fully homomorphic encryption scheme. vol. 20. Stanford university Stanford; 2009.

[20] Yang Q, Liu Y, Chen T, Tong Y. Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology (TIST). 2019;10(2):1–19.

[21] MindSpore: https://www.mindspore.cn/security

[22] Porn Producers Offer to Help Hollywood Take Down Deepfake Videos, Janko, Roettgers, https://variety.com/2018/digital/news/deepfakes-porn-adult-industry-1202705749/, Aug. 2019.

[23] Raj A, Bresler Y, Li B. Improving robustness of deep-learning-based image reconstruction. In: International Conference on Machine Learning. PMLR; 2020. p. 7932–7942.

[24] Li F, Li H, Niu B, Chen J. Privacy computing: concept, computing frame-work, and future development trends. Engineering. 2019;5(6):1179–1192.

[25] Rahm E, Do HH. Data cleaning: Problems and current approaches. IEEE Data Eng Bull. 2000;23(4):3–13.