November 21, 2024

China Publishes the AI Security Governance Framework

Authored by: Liza L.S. Mark and Tianyun (Joyce) Ji

On Sept. 9, 2024, the National Technical Committee 260 on Cybersecurity of Standardization Administration of China (the "**Committee**") released the Artificial Intelligence Security Governance Framework 1.0 *(《人工智能安全治理框架》1.0版)* (the "**AI Framework**"). It is enacted in response to President Xi Jinping's Global AI Governance Initiative *(《全球人工智能治理倡议》)* in October 2023. The AI Framework acknowledges that artificial intelligence (AI) is "a new area of human development" that "presents significant opportunities to the world while posing various risks and challenges." It provides non-binding yet helpful guidance for AI developers, service providers as well as users in dealing with AI-security risks.

China's AI Framework is not the first of its kind. In January 2023, the United States National Institute of Standards and Technology (NIST) published the Artificial Intelligence Risk Management Framework (the "**NIST AI Framework**"). As of Aug. 1, 2024, the EU Artificial Intelligence Act also came into force and is the most comprehensive legal framework governing AI.

This article summarizes highlights of the AI Framework that China just adopted and compares certain of its key concepts against the NIST AI Framework.

1. **Governance principles.**

Unlike the NIST AI Framework, which emphasizes AI trustworthiness and related risks[1], the AI Framework calls for equal attention to AI development as well as security that are mutual, comprehensive, collaborative and sustainable under the following core principles:

- Be inclusive and prudent to ensure safety. This is to encourage development and innovation and take an inclusive approach to AI research, development and application. At the same time, it ensures AI safety, and takes timely measures to address any risks threatening national security, public interest or legitimate rights and interests of individuals.

- Identify risks with agile governance Closely track trends in AI research, development and application to identify AI safety risks from two aspects: the technology itself and the application. Preventive measures are proposed to mitigate these risks.

- Integrate technology and management for coordinated response to adopt a comprehensive safety governance approach that integrates technology and management to prevent and address various safety risks throughout the entire lifecycle of AI research, development and application. It is essential to ensure that all relevant parties (including model and algorithm researchers and developers, service providers and users) assume their respective responsibilities for AI safety.

[1] Under the NIST AI Framework (Chapter 3 AI Risks and Trustworthiness), trustworthy AI shall be valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced and fair with harmful bias managed.

- Promote openness and cooperation for joint governance and shared benefits by sharing best practices, advocating for establishing open platforms and advancing efforts to build global consensus on AI governance.

2. Security Risks and Corresponding Technical Solutions

The AI Framework classifies AI-related risks into two categories: (i) inherent security risks and (ii) AI application security risks. It proposes corresponding technical solutions for each of the identified risks.

While the NIST AI Framework similarly identifies those security-related risks (e.g., risks to transparency, explainability and interpretability, system risks, real-world risks, etc.), it provides solutions (technical or otherwise) in separate contexts and structure by introducing the "RMF Core" (as further discussed in section 3 below). Other NIST frameworks, such as the NIST Cybersecurity Framework, the NIST Privacy Framework, the NIST Risk Management Framework and the Secure Software Development Framework, are also helpful in informing security and privacy considerations in the RMF Core.

The **inherent security risks** capture modeling and algorithm security risks, data security risks and system security risks. According to the AI Framework:

a. **Modeling and algorithm risks** include the risks of explainability, bias and discrimination, robustness, stealing and tampering, unreliable output and hostile attacks. To address those risks, AI developers should (i) constantly improve AI's explainability and predictability, and (ii) establish secure development standards in design, R&D, and deployment to enhance robustness.

b. **Data risks** include the risks of illegal collection and use of data, improper content in training data, unregulated training data annotation and data leakage. To address those risks, AI developers should: (i) follow security rules on data collection, storage, usage, processing, transmission, provision, publication and deletion to ensure AI users' rights under laws and regulations (such as right to control, right to be informed, right to choose etc.) are safeguarded; (ii) strictly select training data to exclude sensitive data in high-risk fields such as nuclear, biological and chemical weapons; (iii) strengthen data security management to comply with relevant data security and personal information protection standards and regulations; (iv) use truthful, accurate, objective and diverse training data from legitimate sources, and timely filter out ineffective, incorrect and biased data; and (v) comply with regulations on cross-border data flow and any applicable export-control requirements.

c. **System risks** include the risks of exploitation through defects and backdoors, computing infrastructure security and supply-chain security. To address those risks, AI developers should (i) properly disclose the principles, capacities, application scenarios and safety risks of AI technologies and products to clearly label outputs and to constantly make AI systems more transparent; (ii) enhance the risk identification, detection and mitigation of platforms where multiple AI models or systems congregate to prevent malicious attacks or invasions; (iii) strengthen the capacity of constructing, managing and operating AI computing platforms and AI system services safely to ensure uninterrupted infrastructure operation and service provision; and (iv) consider the supply-chain security of the chips, software, tools, computing infrastructure and data sources, and timely track any vulnerabilities and flaws of both software and hardware products and make timely repairs and reinforcement to ensure system security.

The **security risks in AI applications** capture risks in the fields of cyberspace, real world, cognitive and ethics.

a. **Cybersecurity risks** include the risks of information and content safety, confusing facts, misleading users and bypassing authentication, information leakage due to improper usage, abuse for cyberattacks and security flaw transmission caused by model reuse. To address those risks, AI developers should (i) establish security protection mechanisms to prevent models from being interfered and tampered and (ii) set up data safeguard to ensure that AI systems comply with applicable laws and regulations when outputting sensitive personal information and important data.

b. **Real-world risks** include the risks of economic and social security, using AI in illegal and criminal activities, and misuse of dual-use items and technologies. To address those risks, AI developers should (i) establish service limitations according to users' actual application scenarios and cut AI systems' features that might be abused beyond the present scope and (ii) improve the ability to trace the end use of AI systems to prevent high-risk application scenarios, such as manufacturing weapons of mass destruction.

c. **Cognitive risks** include the risks of amplifying the effects of "information cocoons," and the usage in launching cognitive warfare. To address those risks, AI developers should (i) identify unexpected, untruthful and inaccurate outputs according to laws and regulations; (ii) take strict measures to prevent abuse of AI systems that collect, connect, gather, analyze and dig into users' inquiries to profile their identity, preference and personal mindset and (iii) intensify R&D of AI-generated content (AIGC) testing technologies to better prevent, detect and navigate cognitive warfare.

d. **Ethical risks** include the risks of exacerbating social discrimination and bias, challenging traditional social order and AI becoming uncontrollable in the future. To address those risks, AI developers should (i) screen training data and verify outputs during algorithm design, model training and optimization, service provision and other processes to prevent discrimination based on ethnicities, beliefs, nationalities, region, gender, age, occupation and health etc. and (ii) equip emergency management and control measures for AI systems applied in key sectors, such as government departments, critical information infrastructure and areas directly affecting public safety and people's health and safety.

3. Governance Measures

In addition to having technological solutions, the AI Framework emphasizes the need for sound governance and control from collaborative efforts by various stakeholders (including AI developers, service providers, users, government authorities etc.). The AI Framework suggests the following best practices for non-mandatory governance measures:

- To implement a tiered and categorical management for AI application, by classifying and grading AI systems based on their features, functions and application scenarios, and setting up a testing and assessment system according to risk levels (applicable to security risks in AI applications).

- To develop a traceability management system for AI services by using digital certificates to label the AI systems serving the public (applicable to security risks in AI applications).

- To improve AI data security and personal information protection regulations by explicating the requirements for data security and personal information protection in various stages, such as AI

training, labeling, utilization and output based on the features of AI technologies and applications (applicable to data risks).

- To create a responsible AI R&D and application system by proposing pragmatic instructions and best practices to uphold the people-centered approach and adhere to the principle of developing AI, exploring the copyright protection, development and utilization systems, and establishing AI-related ethical review standards, norms and guidelines (applicable to modeling and algorithms risks).

- To strengthen AI supply chain security by promoting knowledge sharing in AI, making AI technologies available to the public under open-source terms and jointly developing AI chips, frameworks and software (applicable to system risks).

- To advance research on AI explainability regarding transparency, trustworthiness and error-correction mechanisms in AI decision-making from the perspectives of machine learning theory, training methods and human-computer interaction (applicable to modeling and algorithms risks).

- To share information and emergency response of AI safety risks and threats by constantly tracking and analyzing security vulnerabilities, defects, risks, threats and safety incidents related to AI technologies, software and hardware products, services and other aspects (applicable to system risks).

- To enhance the training of AI safety talents (applicable to security risks in AI applications).

- To establish and improve the mechanisms for AI safety education, industry self-regulation and social supervision (applicable to security risks in AI applications).

- To promote international exchange and cooperation on AI safety governance (applicable to security risks in AI applications).

The AI Framework also emphasizes that it needs to align AI governance with global standards and practices, and it recognizes that cross-border collaboration is necessary for addressing global challenges around AI (e.g., cybersecurity, ethics, security etc.).

As for the NIST AI Framework, its core is composed of four functions: govern, map, measure and manage (the "**RMF Core**"). Under the NIST AI Framework, "govern" is a cross-cutting function that is infused throughout AI risk management, which includes: (i) cultivating and implementing a culture of risk management; (ii) outlining processes, documents and organizational schemes that anticipate, identify and manage risks; (iii) incorporating processes to assess potential impacts and (iv) providing a structure by which AI risk-management functions can align with organizational principles, policies and strategic priorities etc.

4. Safety Guidelines for AI development

The AI Framework provides stakeholder-specific safety guidelines for AI development and application as follows:

- **Model algorithm developers** should adhere to ethics, strengthen data security and protection, guarantee the security of training environments, assess potential biases, evaluate readiness of products and services, regularly conduct safety and security evaluations, and generate and analyze testing reports.

- **AI Service providers** should publicize information and disclosures related to their AI use, obtain user consent, establish and improve real-time risk monitoring and management systems, report safety and security incidents and vulnerabilities, and assess the impact of AI products on users.

- **Users in key sectors** (such as government, critical information infrastructure, and areas directly affecting public safety and health) should assess impacts of applying AI technology, conduct risk assessments, regularly perform system audits, fully understand data processing and privacy protection measures, enhance network and supply-chain security, limit data access and avoid complete reliance on AI for decision making without human intervention.

- **General public users** should be on alert for potential safety risks associated with AI, carefully review all terms of service, enhance awareness of personal information protection, become informed about data-processing practices and cybersecurity risks and be aware of the potential impact of AI products on minors.

While the NIST AI Framework does not provide similar stakeholder-specific safety guidelines, under its "manage" function of the RMF Core, framework users shall allocate risk resources to mapped and measured risks on a regular basis and as defined by the "govern" function[2]. Detailed best practices are described under the NIST AI RMF Playbook[3].

5. Takeaways

Unlike the NIST AI Framework, which is more detail-oriented for the primary purpose of increasing AI trustworthiness and to help the responsible design, development, deployment and use of AI systems, China's AI Framework takes a more general risk-based approach to AI governance and ties each risk category to specific mitigation measures.

China's AI Framework acts as a helpful technical guide for AI developers, service providers and users to effectively respond to AI risks. As stated by the speaker of the Committee, the AI Framework is expected to take an important role in promoting AI security governance by all stakeholders in the society. It will also help by promoting global efforts, forming consensus in AI governance and ensuring that AI technology benefits mankind.[4]

For more information, please visit our China Updates page or see the following resources:

China's Data as a Fifth Market Production Factor – an Asset on Your Balance Sheet, September 23, 2024

China Releases New Rules to Ease Burden on Cross-Border Transfer of Data, May 16, 2024

China Increases Filing Thresholds for Antitrust Merger Review, April 2, 2024

China Streamlines Requirements Regarding Data Export in the Greater Bay Area, February 29, 2024

[2] According to "Table 4: Categories and subcategories for the MANAGE function" of the NIST AI Framework.
[3] https://www.nist.gov/itl/ai-risk-management-framework/nist-ai-rmf-playbook
[4] https://www.tc260.org.cn/front/postDetail.html?id=20240906174148, the Committee website.

China Releases Regulation on the Protection of Children in Cyberspace, December 5, 2023

China Publishes Interim Measures for the Management of Generative Artificial Intelligence Services, August 7, 2023

Mexico Nearshoring: Opportunity for Manufacturers in China and the U.S., April 5, 2023

China MIIT Releases Data Security Management Measures for Industrial and Information Technology Sectors, February 20, 2023

A New Guideline Added to China's Data Protection Framework, August 17, 2022

China Revises its Anti-Monopoly Law 14 Years After its Initial Implementation, July 26, 2022

China Releases Judicial Interpretation of Anti-Unfair Competition Law, April 28, 2022

Select Proposed Changes to the Company Law of the People's Republic of China, March 22, 2022

A Snapshot of China's Cyberspace Administration and Data Protection Framework, February 9, 2022

China Intensifies Regulations on Cryptocurrency Trading and Mining, November 2, 2021

China's Amended Administrative Penalty Law Took Effect on July 15, October 8, 2021

China Issues New Rules Regulating Personal Information Collection by Mobile Apps, April 28, 2021

A New Gateway to China – Recent Policy Developments in the Hainan Free Trade Port, April 6, 2021

China Issues Measures for the Security Review of Foreign Investments, February 9, 2021

China Patent Law Fourth Amendment—Impact on Foreign Companies, January 26, 2021

China Regulators Remove Restrictions on Insurance Fund Investment, December 14, 2020

China Adopts Interim Provisions on the Review of Concentrations of Business Operators for the Anti Monopoly Law, November 30, 2020

China Releases Draft Personal Data Protection Law for Comments, November 12, 2020

China Adopts Export Control Law, November 5, 2020

China Releases New QFII/RQFII Rules, October 27, 2020

China Releases Provisions on Strengthening the Supervision of Private Equity Investment Funds (Draft), October 15, 2020

China Releases Provisions on the Unreliable Entity List, October 5, 2020

China Releases Revised Measures on Handling Complaints of Foreign-Invested Enterprises, September 23, 2020

China Releases Administrative Measures for Strategic Investment by Foreign Investors in Listed Companies, September 10, 2020

China Releases Draft Data Security Law, September 8, 2020

China Releases Circular on Further Stabilizing Foreign Trade and Foreign Investment, August 24, 2020

China Releases Draft Measures for the Administration of Imported and Exported Food Safety, August 18, 2020

U.S. Listed Chinese Companies: Regulatory Scrutiny and Strategic Options, July 30, 2020

China Passes Controversial Hong Kong National Security Law, July 9, 2020

China's Relaxed Financial Sector May Aid Foreign Investors, June 18, 2020

Is There a Law in China Similar to the US Defense Production Act?, May 8, 2020

Coronavirus Brings Force Majeure Claims to LNG Contracts, March 4, 2020

The Rise of China, March 4, 2020

Coronavirus Fears Cast Cloud Over Dealmaking, February 27, 2020