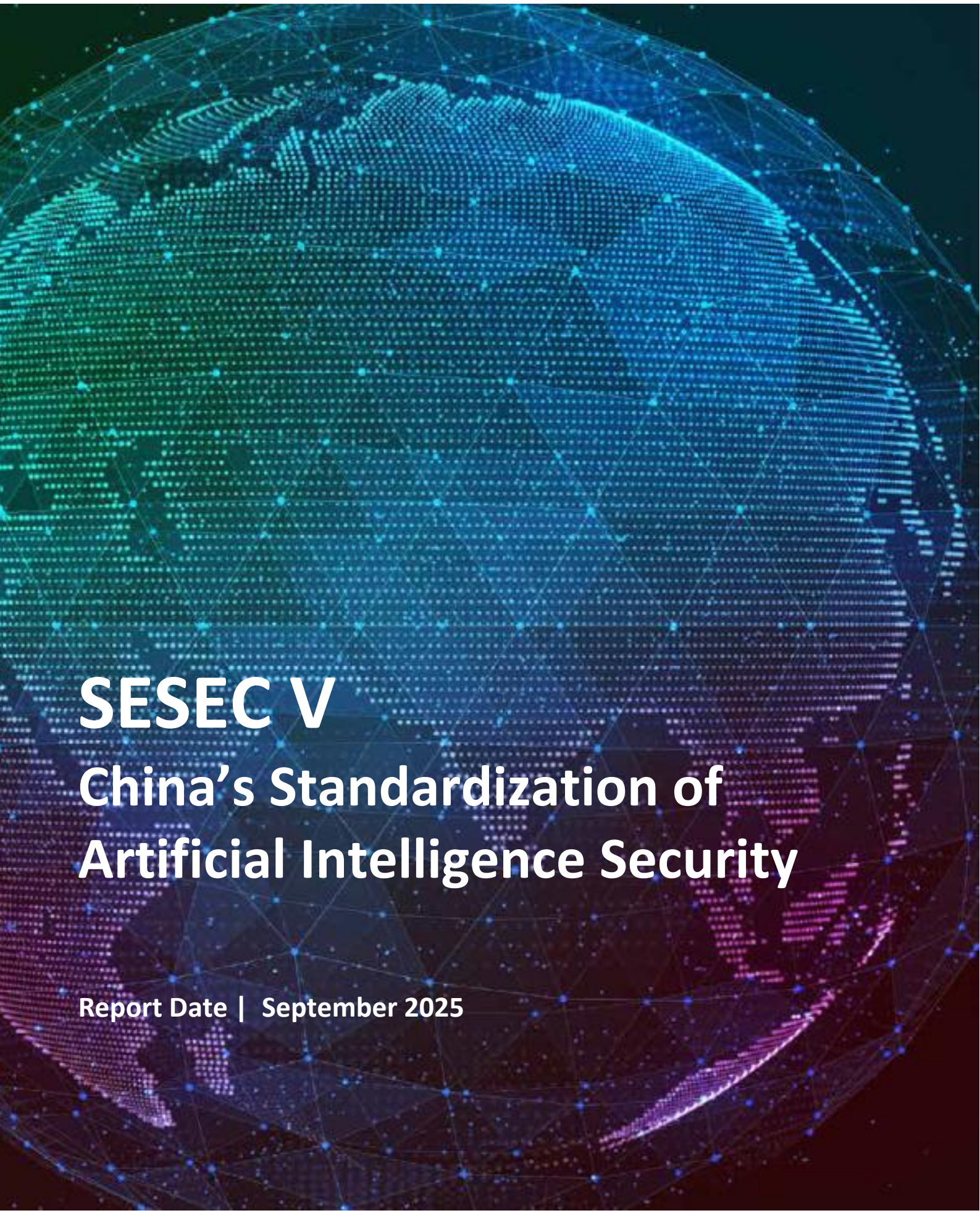


SESEC V

China's Standardization of Artificial Intelligence Security

Report Date | September 2025

Seconded European Standardisation Expert in China
(SESEC)

China's Standardization of Artificial Intelligence Security
September 2025

1. Overview

From 15 September to 18 September 2025, the National Information Security Standardization Technical Committee (**TC260**) under the **State Administration for Market Regulation (SAMR)** and the **Cyberspace Administration of China (CAC)** held its second standards week summarizing its annual standardization work and wrapping up standards development with open standards discussion. Meanwhile, the *Artificial Intelligence Safety Governance Framework (Version 2.0)* was released, outlining China's strategic and operational approach to AI governance.

The framework and accompanying standards aim to ensure that AI technologies develop in a safe, controllable, and trustworthy manner, supporting innovation while safeguarding public interests. TC260's activities during 2025 focused on developing standards for AI security, generative AI, and AI agents, reflecting China's ambition to address next-generation risks arising from intelligent, self-directed systems.

Within TC260, the **Special Working Group on Emerging Technology Standards (SWG-ETS)** is responsible for cybersecurity standardization related to artificial intelligence. This SESEC report provides an overview of TC260's ongoing work on AI as reported by SWG-ETS. It examines the structure of AI safety standardization, highlights key ongoing and emerging standards projects, and outlines the main challenges in this area. This report also reviews the legal framework governing cybersecurity and AI-related laws to provide an overall view of China's regulatory landscape.

Note:

Before reading, please bear in mind that the term **AI safety** came from TC260 context is equivalent to what is commonly known as the **AI security**. This report may use two terms - '**AI safety**' and '**AI security**' - interchangeably to respect TC260's original translation and prevent confusion from readers at the same time.

2. China's Current AI Security Standardization Landscape

2.1 Laws and Regulations

China has established an AI governance framework that is "secure, controllable, and innovation-oriented," built upon the pillars of **data sovereignty**, **content identification**, and **algorithmic transparency**, aiming to strictly control risks while guiding technology toward positive and ethical development.

The table below outlines China's key developments in AI security governance from 2021 to 2025. Over this period, China established a comprehensive framework linking laws, administrative measures, and technical standards. Starting with data and algorithm regulations as a foundation, the focus expanded to deep synthesis and generative AI governance, global cooperation, and content labeling. By 2025, CAC and TC260 had transitioned from policy to implementation, introducing national standards supporting secure and traceable AI-generated content management.

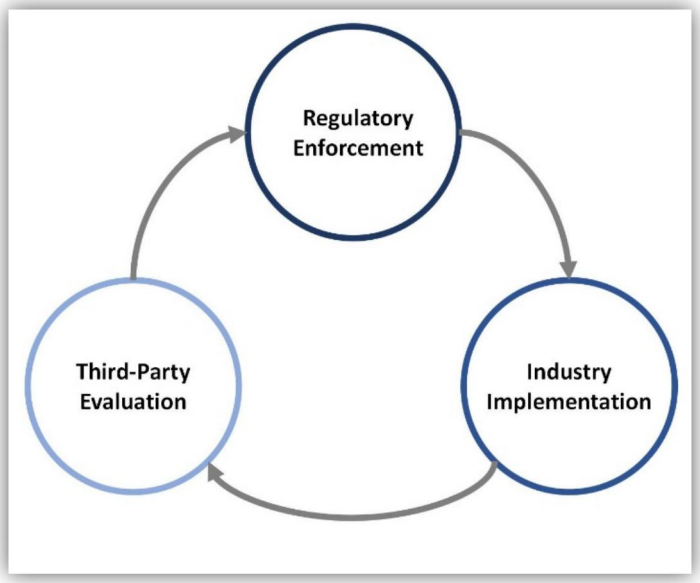
Table 1. Timeline of Key Policies on AI safety Governance in China (2021-2025)

Year	Laws and Regulations	Issuing Body	Key Focus
2021	***Data Security Law; Personal Information Protection Law***	Standing Committee Member of the National People's Congress	Established the legal foundation for data and privacy governance.

	Provisions on the Administration of Algorithm-generated Recommendations for Internet Information Services	CAC, MIIT, MPS, SAMR	Introduced requirements for algorithmic transparency, fairness, and accountability.
2022	***Provisions on the Administration of Deep Synthesis of Internet-based Information Service***	CAC, MIIT, MPS	Regulated AI-generated and synthetic content to ensure authenticity and prevent misuse.
2023	***Global AI governance Initiative***	CAC	Launched global AI governance principles.
	Interim Measures for Administration of Generative Artificial Intelligence Services	CAC, NDRC, MOE, MOST, MIIT, MPS, NRTA	Established China's first regulatory framework for generative AI.
2024	***Guidelines on the Construction of Standards System for AI safety Governance in the Industry and Information Technology Field***	MIIT	Outlined the roadmap for AI safety standardization and implementation within the industrial and ICT sectors.
	AI Safety Governance Framework 1.0	TC260	Established the preliminary framework for AI safety governance for the first time, providing a foundation for subsequent refinement and upgrading.
2025	***Global AI governance Action Plan***	World Artificial Intelligence Conference in Shanghai	Promoted international cooperation.
	Measures for the Labeling of Content Generated by Artificial Intelligence	CAC, MIIT, MPS, NRTA	Enhanced traceability and labeling of AI-generated content.
	AI Safety Governance Framework 2.0	TC260	Built upon Framework 1.0, Framework 2.0 focuses on exploring risk classification and grading, and revises the governance approach to reduce regulatory breaks throughout life cycles.
	GB 45438-2025 Cybersecurity technology – Labeling method for content generated by artificial intelligence ***GB/T 45654-2025 Cybersecurity technology – Basic security requirements for generative***	TC260	Provided technical requirements for the secure management and evaluation of AI-generated content.

	artificial intelligence service		

In implementing AI network and data security standards, China has established a closed-loop mechanism integrating regulatory enforcement, industry implementation, and third-party evaluation.



Multiple ministries, including CAC and MIIT, have jointly issued governance and technical guidelines. Third-party organizations have developed trustworthiness evaluation systems for large models, enabling a complete process cycle of enterprise testing, improvement, and re-evaluation.

2.2 Standards System

TC260 has been developing and improving China's AI security standardization system under the guidance of CAC. Drawing on the technical characteristics and industrial development trends of the AI, TC260 has proposed an overall framework for AI security standards covering five dimensions: **foundational and generic standards, safety management, key technologies, testing and evaluation, and product and applications**. This framework provides systematic standardization support for China's ongoing work in ensuring secure and responsible use of AI and the industry consistently meeting regulatory and technical requirements.

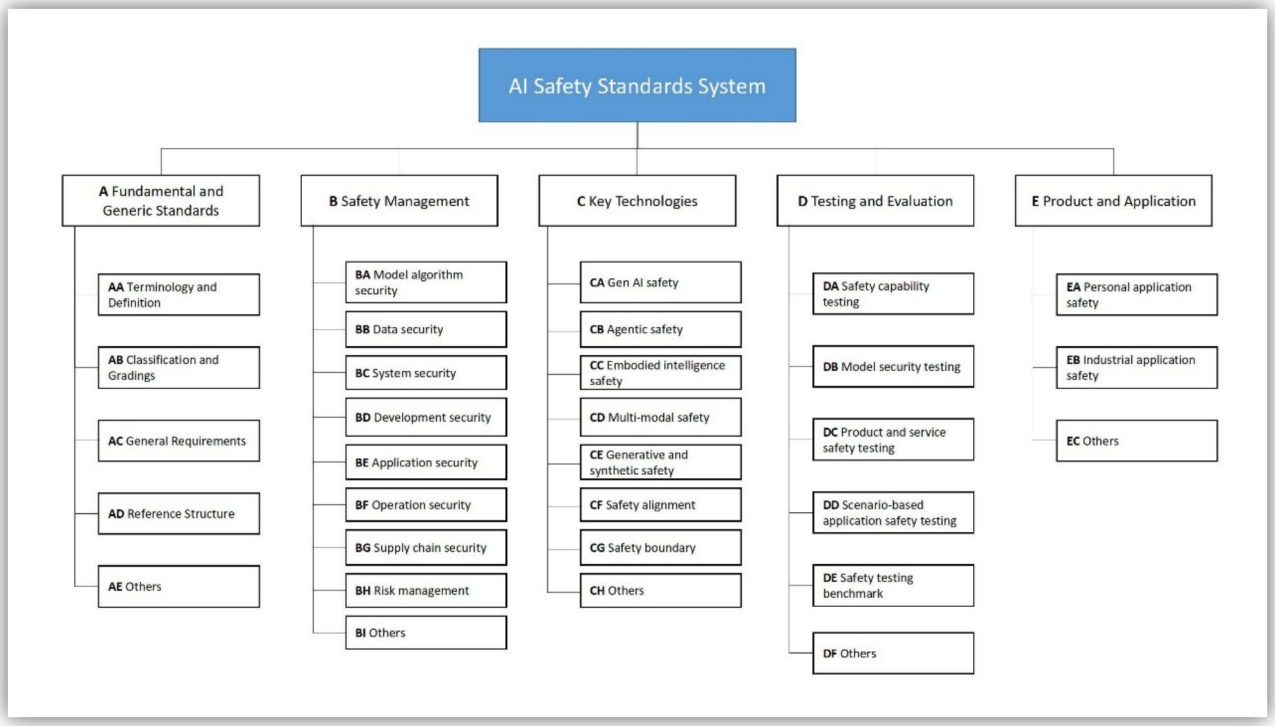


Figure 1. TC260 AI Safety Standards System as of September 2025

2.3 Generative AI Security Under the Spotlight

To in line with ***Interim Measures for Administration of Generative Artificial Intelligence Services*** and support the registration and filling of large model, TC260 has published:

- ***GB/T 45654-2025 Cybersecurity technology – Basic security requirements for generative artificial intelligence service***

Meanwhile, another two national standards regarding security specification of training data and data labeling have been published:

- ***GB/T 45652-2025 Cybersecurity technology – Security specification for generative artificial intelligence pre-training and fine-tuning data***
- ***GB/T 45674-2025 Cybersecurity technology – Generative artificial intelligence data annotation security specification***

The Key Mandatory Standard of the Year

In February 2025, TC260 officially published mandatory national standards ***GB 45438-2025 Cybersecurity technology - Labeling method for content generated by artificial intelligence*** to support the ***Measures for the Labeling of Content Generated by Artificial Intelligence*** issued in March 2025***.*** The mandatory standard came into effect in September 2025. To support providers of online information content dissemination service (e.g., AIGC

tools) in achieving a smooth and effective transition to the new regulatory requirements, TC260 published **6 practice guidelines** detailing the step-by-step instructions for different types of AI-generated content and for building an internal compliance framework. These guidelines also provide third-party testing institutes an important reference for their operations.

In addition, TC260 is developing **another 11 practice guidelines** with 3 dedicated to technical pathway and 8 left for different application scenarios such as intelligent search, voice-based AI customer service and intelligent terminal assistant.

2.4 New AI Governance Framework

In September 2025, TC260 officially published ***AI Safety Governance Framework 2.0*** which is a revision based on the framework in 2024 (here for more details and to download its original document). Framework 1.0 was an initial attempt exploring AI security governance model. It compiled a list of risks to guide future exploration and the establishment of governance paradigms. Framework 2.0 builds on that foundation as a more systematic governance framework, refined within a year through practical experience and lessons learned.

As the new framework came into public domain, TC260's standard development and system refinement will gradually shift towards the structure of the new framework. The standards currently under development or already published cover four key areas:

1. foundational and general standards,
2. technological intrinsic risks,
3. technological application risks, and
4. Application-derived risks.

In the previous framework, non-technical risks arising from the use of technology (i.e., collateral damages from using AI) were arranged under technological application risks.

The new framework reorganizes the negligence of mixing technical and non-technical risks, expanding the classification of security risks by dividing them into two dimensions:

- Technological risks and
- Application-derived risks (i.e. the collateral damage).

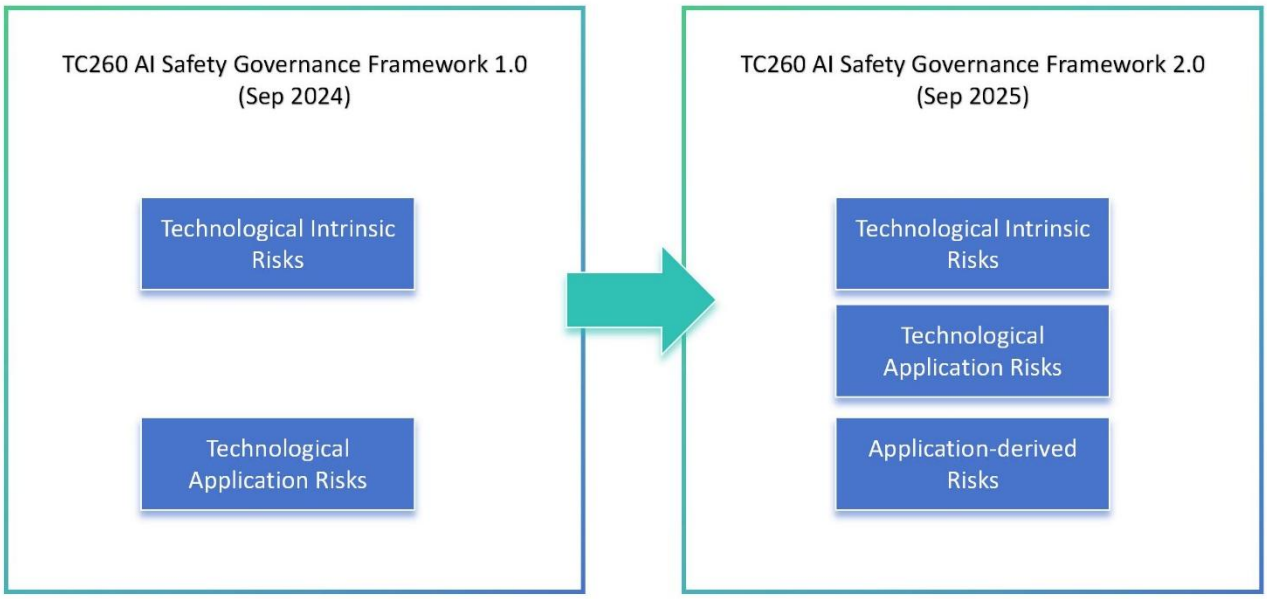


Figure 2.0 Old Framework vs. New Framework

Within the technological dimension, it further distinguishes between **technological intrinsic risks** and **technological application risks**, while creating a separate major category for **non-technical risks triggered by the use of AI technologies**, such as disruptions to employment structures or excessive emotional dependence on AI. The separate classification of application-derived risks reflects China's growing emphasis on this area, recognizing its critical role in accelerating the deployment and adoption of AI across real-world application scenarios and in advancing the goals of the AI+ Initiative more swiftly.

The following sections summarize TC260's completed AI safety standards, ongoing standard development projects, and ongoing practice guide development projects.

Note:

1) Items with such code ***20240395-T-469*** are ongoing standards projects. Standards are not yet to be published.
2) Practice guides are abbreviated as **PG** and usually coded in such format: TC260-PG-YYYYNA. Published practice guides will have year information (e.g. TC260-PG-202512A） , while the practice guide in developing and unpublished are simply prefixed with TC260-PG-YYYYNA.
3) In addition, some of the items with no project code are new standard projects launched in 2024 and 2025. They have not been assigned official codes at time of drafting this report. It should be noted that several of the projects currently under development are still awaiting their official project designations.

Foundational and Generic Standards

No.	Standards category	Standards name
1	Terminology	• ***Vocabulary for Artificial Intelligence Safety***
2	Classification and grading management	• ***Classification and grading method for the security of artificial intelligence applications***
3	General safety requirements	• ***GB/T 45654-2025 Cybersecurity technology—Basic security requirements for generative artificial intelligence service*** • ***GB/T 42888-2023 Information security technology—Assessment specification for security of machine learning algorithms*** • ***Basic security requirements for generative artificial intelligence services invoking third-party foundation models***
4	Testing and evaluation	• *TC260-PG-YYYYNA Guideline for security evaluation methods and model selection of artificial intelligence model*

Technological Intrinsic Risk

No.	Standards category	Standards name
1	Training Safety	• ***GB/T 45652-2025 Cybersecurity technology—Security specification for generative artificial intelligence pre-tra-ining and fine-tuning data*** • ***GB/T 45674-2025 Cybersecurity technology—Generative artificial intelligence data annotation security specification*** • ***Security guide for artificial intelligence model development and customization***

		• *TC260-PG-YYYYNA Security guide for artificial intelligence training data cleansing*
2	Key Safety Characteristics	• ***Methods for robustness evaluation and enhancement of artificial intelligence models*** • ***Methods for interpretability evaluation and enhancement of artificial intelligence systems***

Technological Application Risk

No.	Standards category	Standards name
1	Content safety	• ***GB 45438-2025 Cybersecurity technology—Labeling method for content generated by artificial intelligence*** • ***20240395-T-469 Information security technology – Security specification for deep synthesis of internet information services*** • *TC260-PG-202512A Practice guide for cybersecurity standards – Detection of AI-generated and synthetic content Part 1: framework*
2	System safety	• *TC260-PG-YYYYNA Guide for secure development of artificial intelligence systems* • ***Guide for building security protection capabilities of artificial intelligence systems*** • ***Practice guidelines for transparency in artificial intelligence systems*** • *TC260-PG-YYYYNA Guide for security emergency response of generative artificial intelligence artificial intelligence services* • ***Security guide for using artificial intelligence*** • ***Security guide for open-source artificial intelligence***
3	Safe use of data	• *TC260-PG-YYYYNA Specification for personal information protection in artificial intelligence* • ***Guide for secure application of AI-generated and synthetic data***
4	Operational environmental safety	• ***Guide for secure deployment of large models*** • ***Security framework for artificial intelligence computing platform*** • ***Basic security requirements for AI PC products*** • *TC260-PG-YYYYNA Basic security requirements for large-model all-in-one products*
5	Agentic safety	• *TC260-PG-YYYYNA Security requirements for AI agents* • *Security guide for interoperability of AI agents*
6	Vertical application safety	• ***Security requirements for AI code generation services*** • ***Audit guide for code generation for code generation by cybersecurity large model***

		• ***Security guidance for industry applications of artificial intelligence***
7	Supply chain safety	• *TC260-PG-YYYYNA Technical specification for the security of artificial intelligence acceleration chips* • ***Security requirements for AI training and inference frameworks***

Application-Derived Risk

- *TC260-PG-20211A Guide for preventing ethical and security risks of artificial intelligence*
- ***Basic security specifications for embodied AI services***
- ***Cybersecurity requirements for intelligent driving services***
- ***Security guide for application of AI for technology involving minors***

3. Challenges Faced During Standards Implementation

Despite notable progress in building China's AI security standardization framework, TC260 experts highlighted several challenges in implementation during the Standards Week.

At the demand level, the main difficulty lies in aligning diverse industrial needs with uniform standards. Sectors such as healthcare, finance, and manufacturing have very different AI security requirements, yet existing standards are mostly based on general frameworks.

"One-size-fits-all" approach often results in standards that are either too strict or too loose, reducing their practical value. As technical requirements evolve rapidly, small and medium-sized enterprises often lack the resources to fully comply.

Many engage in superficial implementation or ignore standards altogether, leading to uneven adoption and weak links in the industrial value chain. New technologies such as generative AI, multi-modal AI, and edge AI have also introduced new security demands, while the absence of a flexible response mechanism limits the standards system's ability to adapt in time.

At the technical level, most existing standards remain principle-based and lack clear, measurable technical criteria, leaving enterprises without concrete implementation guidance. This gap makes it difficult to ensure consistent compliance on issues such as model interpretability or data anonymization.

Existing testing tools, developed for traditional AI systems, are not fully compatible with emerging technologies and struggle to address risks such as inference security or cross-modal attacks. In addition, differences in encryption methods, privacy-preserving computation protocols, and underlying standards between domestic and international systems complicate integration and increase compliance costs for multinational companies.

At the mechanism level, coordination among stakeholders remains insufficient. Overlapping responsibilities and limited communication between regulators, research institutions, and enterprises slow the translation of research outcomes into standards and result in redundant or fragmented efforts.

The absence of a dynamic update mechanism also means that standards often lag behind rapid technological progress, creating gaps in governance. Moreover, most AI security standards are voluntary, with weak incentives and unclear penalties for non-compliance. High compliance costs and low enforcement risks discourage active

adoption, particularly among SMEs, weakening the overall effectiveness of standards in ensuring AI and data security.

4. Standardization Trend of AI safety

Following these implementation challenges, TC260 experts analysed standardization trend of AI security they expect in the coming future.

At the regulatory level, the focus is on clarifying cross-border data requirements through standardization to ensure secure and controllable data flows, while promoting algorithm transparency to enhance the auditability and security of AI systems.

Strengthening coordinated oversight that integrates policy and law will improve regulatory efficiency, while international cooperation will help promote mutual recognition of standards and global alignment.

Mechanisms for standard revision and monitoring will also be established to improve adaptability and risk warning capabilities.

At the technical level, efforts focus on defining clear implementation, performance, and verification specifications for key technologies such as federated learning, as well as setting requirements for training data, access control, and adversarial defense in large models and generative AI.

Compatibility between technical specifications and international standards will be ensured to support interoperability across domains. Governance measures will be incorporated throughout the entire AI lifecycle to achieve end-to-end management.

At the industrial level, standards are being adapted to specific sectors and application scenarios to meet diverse security needs. A tiered, differentiated approach helps lower compliance costs for SMEs while maintaining comprehensive safeguards for large enterprises.

Technical benchmarks and certification mechanisms guide fair competition and enhance both market credibility and international competitiveness. Standards are also integrated into industrial and supply chain security management by clarifying cross-entity responsibilities and ensuring ecosystem integrity. In addition, interaction between standardization and technological innovation is being strengthened to support dynamic updates that promote secure and sustainable AI-driven industrial development.

At the international cooperation level, the emphasis is on promoting unified or compatible mechanisms to ensure cross-border data security and mutual recognition of certification. Establishing an international coordination mechanism for AI security incidents will enhance global response capabilities. Multilateral cooperation will strengthen the compatibility of international standards and support consistent cross-domain governance. Broader sharing of technical and regulatory expertise will improve the scientific rigor and practicality of standards, while dynamic multilateral collaboration will ensure that standardization evolves in step with technological progress and innovation.

5. Conclusion

China's standardization of AI security under TC260 has entered a stage of structured and coordinated development. The release of the *AI Safety Governance Framework 2.0* marks a transition from the early exploratory phase of AI security governance toward constructing a systematic framework.

TC260's expanding system of standards now covers key areas such as generative AI, data management, model robustness, interpretability, and content labeling. The combination of mandatory standards and practice guidelines demonstrates an effort to link regulatory requirements with technical implementation and to strengthen the connection between policy and industry practice.

At the same time, challenges persist. The rapid evolution of AI technologies continues to outpace the current standardization cycle, while differing industry capacities and fragmented coordination across departments limit consistent adoption. Ongoing efforts to improve communication among regulators, enterprises, and research bodies reflect an awareness of these systemic issues.

Looking ahead, TC260's emphasis on international compatibility, dynamic revision mechanisms, and coordinated governance indicates an intention to build an adaptive and globally relevant AI security standardization system. This evolution will continue to shape China's broader regulatory and industrial landscape, influencing the future direction of technical governance and global standard-setting in the field of artificial intelligence.

Introduction of SESEC Project

The Seconded European Standardisation Expert in China (SESEC) is a visibility project co-financed by the European Commission (EC), the European Free Trade Association (EFTA) secretariat and the three European Standardisation Organizations (CEN, CENELEC and ETSI). Since 2006, there has been four SESEC projects in China, SESEC I (2006-2009). SESEC II (2009- 2012), SESEC III (2014-2017), SESEC IV (2018- 2022) and SESEC V (2022-2025). Dr. Betty XU is nominated as the SESEC expert and will spend the next 36 months on promoting EU-China standardisation information exchange and EU-China standardisation cooperation.

The SESEC project supports the strategic objectives of the European Union, EFTA and the European Standardisation Organizations (ESOs). The purpose of SESEC project is to:

- Promote European and international standards in China;
- Improve contacts with different levels of the Chinese administration, industry and standardisation bodies;
- Improve the visibility and understanding of the European Standardisation System (ESS) in China;
- Gather regulatory and standardisation intelligence.

The following areas have been identified as sectorial project priorities by the SESEC project partners: Internet of Things (IoT) & Machine-to-Machine(M2M) communication, communication networks & services, cybersecurity & digital identity, Smart Cities (including transport, power grids & metering), electrical & electronic products, general product safety, medical devices, cosmetics, energy management & environmental protection (including eco-design & labeling, as well as environmental performance of buildings).