Security for AI Systems

A Look at Effective Ways for AI System Security and Privacy

September 2024

Huawei Technologies Co., Ltd.

Contents

01 Purpose

This document shares strategies for addressing the potential cyber security risks of artificial intelligence (AI) systems[1], and provides engineering practices and insights on mitigating security risks at each phase of the AI lifecycle.

02 Scope

As AI technologies continue to evolve and expand into new scenarios, the cyber security risks of AI systems are constantly changing. Therefore, the cyber security protection policies need to be reviewed and updated accordingly. This document aims to share with Huawei's customers and other stakeholders its engineering practices for addressing the emerging cyber security risks of AI systems. It should be noted that this document does not cover AI functional safety issues such as fairness, transparency, and inclusiveness, nor the abuse of AI technologies in cyber attacks.

The cyber security engineering practices for AI systems described here involve protecting critical elements such as data, models, applications, and computing backbone of AI systems throughout the lifecycle, including design, development, deployment, and operation. These practices aim to ensure that AI systems operate reliably in line with design intents and effectively cope with man-made threats.

[1] https://oecd.ai/en/wonk/ai-system-definition-update

03 AI Concerns and Risks

AI[1], in its broadest sense, is intelligence exhibited by machines, particularly computer systems.

Since its inception at the Dartmouth Conference in 1956, AI has developed for nearly 70 years, experiencing many peaks and troughs. In the late 1990s, the reduction in computer costs and the Internet-driven expansion of data scale ushered AI into a period of relatively stable development. As the use of chips such as GPUs has driven increases in computing power and application of deep learning, AI has achieved breakthroughs across various domains, entering a new development period. After 2020, foundation models developed based on deep learning emerged as the predominant research paradigm in AI, driving significant progress of nearly all aspects of AI. Notably, the success of generative AI has greatly expanded the application areas of AI, bringing artificial general intelligence (AGI) within reach.

From the development process, it is evident that algorithms, computing power, and data are the three primary driving forces behind AI development. However, as AI capabilities grow stronger and even approach human intelligence, AI systems have given rise to increasing cyber security issues. This has garnered extensive attention and reflection on the cyber security of AI systems.

At present, ensuring the cyber security of AI systems presents substantial challenges. Without adequate attention, AI may pose greater cyber security risks in the future. As such, Huawei hereby calls on all stakeholders to conduct and comprehensively strengthen the evaluation and supervision of AI systems to address these challenges together. Only through collaboration can we achieve the long-term sustainable and healthy development of AI.

3.1 AI Security Legislation Around the World

In general, factors such as AI technology reserves, business ecosystems, cultural and legal traditions, and governance phases will affect the AI security governance practices of a country or region to different degrees. Therefore, the formulation and implementation of AI laws and regulations vary across countries and regions.

[1] https://en.wikipedia.org/wiki/Artificial_intelligence

3.1.1 EU

Centering on "excellence and trust" in AI governance, the EU has launched industry development policies in step with ethical and regulatory rules to guarantee the safety and fundamental rights[1]. In April 2021, the European Commission proposed the draft Artificial Intelligence Act[2]. Following several negotiations and revisions, the Artificial Intelligence Act was published in the EU Official Journal on July 12, 2024. It entered into force on August 1, 2024, and has been implemented in phases[3].

The Artificial Intelligence Act provides a unified definition of AI, applies to all EU Member States, and establishes a systematic penalty mechanism, laying a legal foundation for comprehensive and systematic supervision and governance. Regarding risk classification, the Artificial Intelligence Act classifies AI systems into different levels based on risks, each managed by a set of predefined regulatory tools. In terms of category-based governance, the EU introduces specific clauses on general-purpose AI regarding GPT-like foundation models, maintains the hierarchical risk management approach, further divides AGI models into two categories based on whether there are systemic risks, and imposes different compliance obligations accordingly.

3.1.2 US

The US has not enacted any systematic AI security law at the federal level. In October 2023, the White House released the Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence[4], outlining a blueprint for AI security supervision in the US. It emphasizes that "it is important to manage the risks from the Federal Government's own use of AI and increase its internal capacity to regulate, govern, and support responsible use of AI to deliver better results for Americans." For national security reasons, the Executive Order mandates, for the first time, that developers of powerful "dual-use foundation models" with substantial security risks must report and share security information, testing information, and the like with the government. Many US states, including Colorado, Florida, Indiana, and California, have started to eye on AI-related legislation, particularly concerning the obligations of high-risk AI system developers. Additionally, the US views AI as a key technical factor in international strategic competition, and has added AI-related technologies to the Critical and Emerging Technologies List[5], highlighting that AI-related technologies are of particular importance to the national security of the US.

[1] https://ec.europa.eu/commission/presscorner/detail/en/IP_21_1682

[2] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

[3] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689

[4] https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

[5] https://www.whitehouse.gov/wp-content/uploads/2022/02/02-2022-Critical-and-Emerging-Technologies-List-Update.pdf

3.1.3 China

The Next-Generation Artificial Intelligence Development Plan[1], released by the State Council in 2017, lays a solid foundation for the law-based AI development in China. It emphasizes that "while vigorously developing AI, we must attach great importance to the potential security risks and challenges arising therein, minimize risks, and ensure the secure, reliable, and controllable development of AI." It sets two strategic objectives for AI governance: (1) By 2025, China will have seen the initial establishment of AI laws and regulations, ethical norms, and policy systems, and the formation of AI security evaluation and control capabilities. (2) By 2030, China will have constructed more comprehensive AI laws and regulations, ethical norms, and policy systems.

Currently, China's Cybersecurity Law, Data Security Law, and Personal Information Protection Law are applicable to AI-related scenarios. Besides, the Cyberspace Administration of China (CAC), in collaboration with relevant departments, has enacted targeted legislation for various application fields, such as algorithmic recommendation services, deep synthesis, and generative AI services, in order to ensure the healthy development and standardized application of AI.

3.2 AI Security Standards and Certifications

Global governments and the industry are actively collaborating to establish international consensus on AI governance, including the OECD AI Principles, the EU's Ethics Guidelines for Trustworthy AI, the UNESCO's Recommendation on the Ethics of Artificial Intelligence, China's Global AI Governance Initiative, and China's Shanghai Declaration on Global AI Governance. In November 2023, 29 countries and regions, including China, the US, Germany, France, the UK, Japan, and the EU, signed the Bletchley Declaration in the UK, underscoring the importance of international cooperation in AI governance. Key points of these international consensus documents include AI cyber security governance and management, risk management, and data protection, which are also the focus of related international standardization efforts.

3.2.1 International Standards: ISO/IEC Standards on AI Security

In 2017, ISO/IEC JTC 1 established SC 42, successor to WG 9 (Big Data Working Group), to take charge of AI standardization. SC 42 has released and is developing several standards on AI security, mainly including:

1. ISO/IEC 42001: 2023 Information Technology — Artificial Intelligence — Management System (released; available for AI management system certification)

[1] https://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm

2. ISO/IEC 23894:2023 Information Technology — Artificial Intelligence — Guidance on Risk Management (released; used with ISO/IEC 42001)

3. ISO/IEC DIS 42005 Information Technology — Artificial Intelligence — AI System Impact Assessment (under development; used with ISO/IEC 42001)

Notably, ISO/IEC 42001:2023 systematically defines 38 controls and guidelines of nine domains in its normative annex, with ISO/IEC 27002:2022 serving as the primary reference.

International standards on AI security released by ISO/IEC JTC 1/SC 42 and SC 27 have been widely adopted by the EU, China, and many others, playing an irreplaceable role in facilitating international trade. Systematically using standards such as ISO/IEC 42001, ISO/IEC 23894, ISO/IEC 42005, and ISO/IEC 27002 helps continuously improve AI system security and foster communication and trust building with customers and other stakeholders.

3.2.2 EU Standards: CEN-CENELEC/JTC 21 Standards on AI Security

The European standardization organizations CEN and CENELEC established JTC 21 to develop AI standards. JTC 21 systematically develops European AI security and trustworthiness standards independently or by building on ISO/IEC standards. JTC 21 has released and is developing several standards on AI security, mainly including:

1. EN ISO/IEC 23894:2024 Information Technology — Artificial Intelligence — Guidance on Risk Management (released; equivalent to the ISO/IEC standard)

2. prEN ISO/IEC 42001 Information Technology — Artificial Intelligence — Management System (under development; equivalent to the ISO/IEC standard)

3. prCEN/CLC/TR AI Risks — Check List for AI Risks Management (under development; reference to ISO/IEC 23894)

3.2.3 US Standards: NIST Standards on AI Security

To better support AI development and internationalization of the US and implementation of the Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, the Information Technology Laboratory (ITL) of the National Institute of Standards and Technology (NIST) has identified AI as a key research focus. It has also outlined several specific tasks aimed at advancing fundamental research on trustworthy AI technologies, AI standardization, and AI technology application innovation.

The NIST has released several standards on AI security and risk management, mainly including:

1. AI 100-1 Artificial Intelligence Risk Management Framework

2. AI 100-2 Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations

3. SP 800-218A Secure Software Development Practices for Generative AI and Dual-Use Foundation Models: An SSDF Community Profile

The standards and practices developed under the leadership of the US NIST are technically practical and conducive to promoting AI security management and security engineering.

3.2.4 China's Standards: TC 260 Standards on AI Security

China attaches great importance to AI development, with standards playing a crucial role in implementing industry policies. The National Technical Committee 260 (TC 260) on Cybersecurity of Standardization Administration of China released the Artificial Intelligence Security Standardization White Paper (2023). This white paper, compiled by 20 organizations including the China Electronics Standardization Institute, proposes a series of recommendations on AI security standardization, aiming to promote the healthy development of AI. These recommendations include continuously enhancing the AI security standards system, researching fundamental common security standards, and releasing security standards that are urgently needed for industry development.

The following standards are under development by TC 260 as of September 2024:

1. Cybersecurity Technology — Security Specification for Generative Artificial Intelligence Pre-training and Fine-tuning Data

2. Cybersecurity Technology — Basic Security Requirements for Generative Artificial Intelligence Service

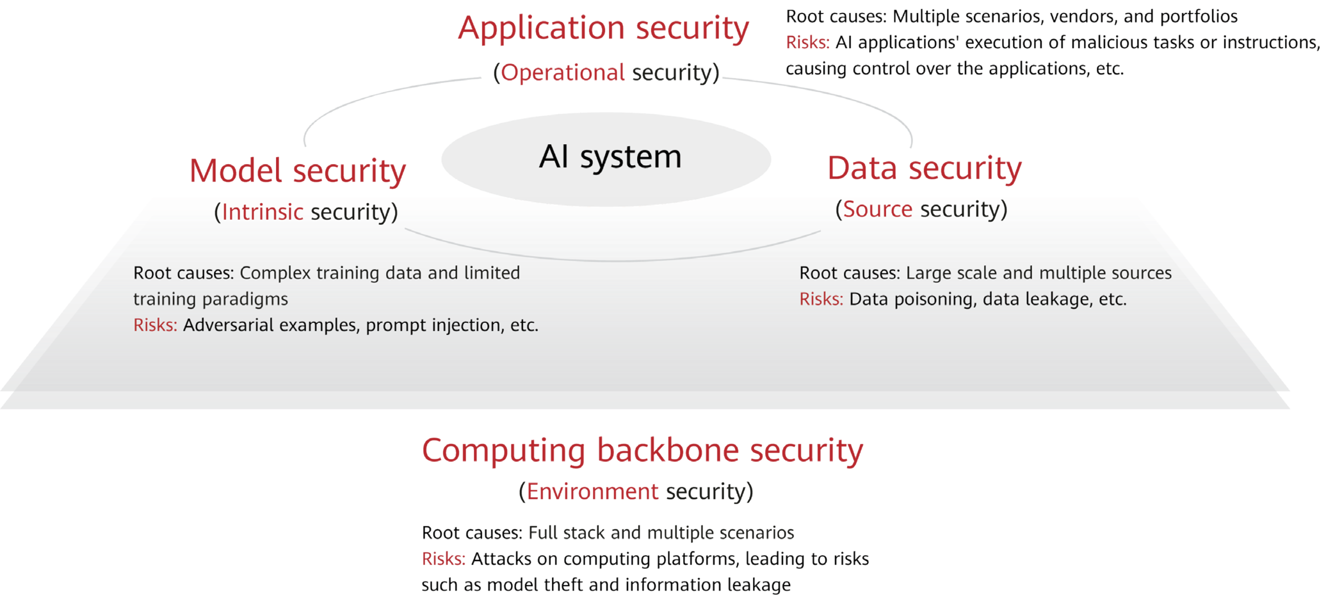
3. Information Security Technology — Security Specification for Deep Synthesis of Internet Information Services

4. Cybersecurity Technology — Generative Artificial Intelligence Data Annotation Security Specification

These standards have effectively supported the implementation of the Interim Measures to Regulate Generative AI Services, the Provisions on the Administration of Internet Information Service Algorithm Recommendation, and the Provisions on the Administration of Deep Synthesis of Internet Information Services, and will continue to guide and promote the healthy and orderly application of AI technologies across various sectors.

3.3 AI Security Risks

As AI technologies become more widely used in social production and daily life, new risks keep emerging and are attracting increasing research and attention.



Root causes: Multiple scenarios, vendors, and portfolios
Risks: AI applications' execution of malicious tasks or instructions, causing control over the applications, etc.

Root causes: Complex training data and limited training paradigms
Risks: Adversarial examples, prompt injection, etc.

Root causes: Large scale and multiple sources
Risks: Data poisoning, data leakage, etc.

Root causes: Full stack and multiple scenarios
Risks: Attacks on computing platforms, leading to risks such as model theft and information leakage

3.3.1 Data Security Risks

Data Poisoning

Data poisoning occurs when malicious actors inject malicious or carefully crafted data samples during model training that will make a model exhibit specific incorrect behaviors or backdoors after training. Consequently, the model performs properly with normal input, but generates unexpected behaviors under specific triggers.

The challenge of data poisoning arises from the vast amount and diverse sources of data that complicate data review, and the black-box nature of models further makes detecting poisoning more difficult. At the same time, the pursuit of profits by malicious actors and weak data security awareness among some developers provide opportunities for such threats. Data poisoning poses serious security risks to models. For example, in application scenarios such as medical diagnosis, financial transaction, and autonomous driving, data positioning may cause severe security accidents or economic losses. In addition, once a model is poisoned, substantial resources may be required to re-collect data and re-train the model, affecting the development efficiency and trustworthiness of the model provider.

Data Leakage

Data leakage refers to the unauthorized access, use, or disclosure of sensitive information related to models. This can occur through various means, such as reversing engineering models to leak training data, making models produce incorrect outputs to expose sensitive information such as personal identities, and stealing or misusing the models' structures and parameters.

The root cause of data leakage lies in the models' powerful feature extraction and memory capabilities. Such leakage can lead to personal privacy violations, identity theft, and disclosure of business secrets. In sensitive sectors such as healthcare and finance, these risks can lead to serious legal issues.

3.3.2 Model Security Risks

Adversarial Example

An adversarial example is a typical threat to models. It involves adding tiny perturbations to the inputs to make models generate incorrect outputs. For example, adding noise to an image of a panda may cause the model to misidentify it as a gibbon.

The main reason for adversarial examples is that the complexity of high-dimensional data makes the model's decision boundary highly nonlinear and discontinuous, making the model sensitive to tiny perturbations. Additionally, limited training data makes the model unstable when encountering out-of-distribution examples. Traditional optimization objectives mainly focus on minimizing training errors, while ignoring the threat posed by adversarial examples. These factors collectively render the model prone to misjudgment when faced with carefully crafted perturbations, which exposes the shortcomings of current models in terms of robustness and generalization capability.

Adversarial examples pose a serious threat to the secure application of models, affecting multimodal fields such as visual, auditory, and text processing systems. In the visual field, autonomous vehicles may misinterpret tampered traffic signs, and security check systems may overlook carefully processed images of dangerous goods. In the auditory field, voice assistants may be manipulated by adversarial instructions hidden in the background sound. Facial recognition and biometric systems are also at risk of being deceived by carefully crafted images or videos.

In the era of AI-generated content (AIGC), multimodal adversarial examples may cause foundation models to produce harmful content that is controllable to malicious actors. For example, a processed image input may make the model generate a textual description containing guidance on illegal or dangerous behaviors. A video generation model may be manipulated to generate disturbing or misleading deepfake content. Similarly, an audio generation system may be induced to create false voice messages or mimic the voice of a specific person.

If such content generated by multimodal adversarial examples becomes widespread, it will cause information confusion. In cross-modal content understanding and generation, the threats posed by adversarial examples are more covert and difficult to detect, offering new avenues for abuse and fraud. These multi-dimensional security risks highlight the urgency and importance of enhancing defense against multimodal adversarial examples in the development of AI.

Prompt Injection

The rapid development of AIGC introduces new security threats, with prompt injection attacks being the most typical. Prompt injection is a type of malicious behavior where malicious actors manipulate foundation models to perform unexpected or potentially harmful operations by subtly crafting input text. Prompt injection is classified into direct injection and indirect injection. Direct injection, also known as jailbreaking attack[1], enables foundation models to output insecure content, while indirect injection means that malicious actors hijack the original tasks of models and manipulate model outputs.

The main reason for prompt injection is the high sensitivity and flexibility of foundation models to instructions. While foundation models perform well in natural language understanding (NLU) and can adapt to a wide range of inputs and instructions, this flexibility can be maliciously exploited. Malicious actors may subtly craft prompts to manipulate the foundation models by exploiting its context sensitivity and faithful execution of instructions.

In addition, models may have challenges in dealing with complex, multi-layered, or potentially contradictory instructions, leaving them vulnerable to exploitation by malicious actors. Despite their strong NLU capabilities, models still face difficulties in distinguishing between normal requests and malicious instructions, especially when these instructions are skillfully embedded within normal text. Malicious actors can subtly craft malicious text to manipulate foundation models.

Prompt injection poses a wide range of security risks to foundation models, involving multimodal content such as text, image, and audio. This may lead to sensitive information leakage, unauthorized system operations, generation of misleading or harmful content, and fraud by exploiting vulnerabilities in third-party applications. In multimodal scenarios, malicious actors may hide text in images, embed instructions in audio, or manipulate video content to launch attacks, making it harder to detect and prevent attacks. Malicious actors can inject malicious prompts directly or indirectly, and even trigger harmful behavior without the users' knowledge.

[1] https://securiti.ai/owasp-top-10-for-llms/

3.3.3 Application Security Risks

Agent Security Risks

An agent is an intelligent entity that can perceive its environment, make decisions, and take actions. As the capabilities of foundation models improve, they can act as the brain of agents and complete various complex tasks through components such as planning, memory, and tool execution.

The security risks of foundation model agents arise when foundation models are manipulated to plan malicious task sequences or generate and execute malicious instructions through prompt injection or adversarial examples. The main cause of the risks lies in the foundation models' inability to distinguish between data and instructions in the input prompts. For example, if a foundation model is tasked with reading and summarizing web page comments (which are data in this context) and these comments contain malicious prompts, such as sending photos to an email address, the foundation model may plan and execute a malicious task when processing such comments.

The security risks of foundation model agents may cause serious harm, including substantial financial losses and leakage of important personal data. Foundation model agents usually have sensitive permissions to access or operate systems, exacerbating traditional cyber security risks. For example, agents can be tricked into visiting websites embedded with Trojan horses, allowing malicious actors to exploit browser vulnerabilities and intrude into mobile phones or PCs.

Application Framework Security Risks

The AI application framework is an important part of an AI system. If this framework has cyber security vulnerabilities, malicious actors may gain control of the AI system. This is because the input and output of models usually need to be processed by plug-ins and tools. For example, when processing a mathematical operation request, the AI system invokes and executes a scientific computing tool within the AI application framework to obtain accurate results. If related plug-ins and tools have command injection vulnerabilities or improper permission configurations, malicious actors may execute malicious code.

Malicious behaviors on the AI application framework may lead to serious consequences, such as the complete control of the AI system by malicious actors. This can result in the leakage of sensitive information (e.g., model files), theft of user assets, spread of malware, and denial of service (DoS) attacks on the AI system.

3.3.4 Computing Backbone Security Risks

Security Risks at the Hardware Layer

The side channel technology may be used to steal critical model information from hardware environments, including CPUs, GPUs/NPUs, DPUs, and even PCIe hardware for communications. Side channel–based model theft primarily involves inferring confidential attributes of a target model by leveraging additional information from operating systems or hardware during model deployment and operation. Consequently, a major risk posed by side channel threats is the inference of model attributes, with the target model's architecture information being of particular interest to malicious actors. Side channel–based model theft risks involve cache, energy consumption, timing, PCIe, and GPU side channels.

Security Risks at the Operating System Layer

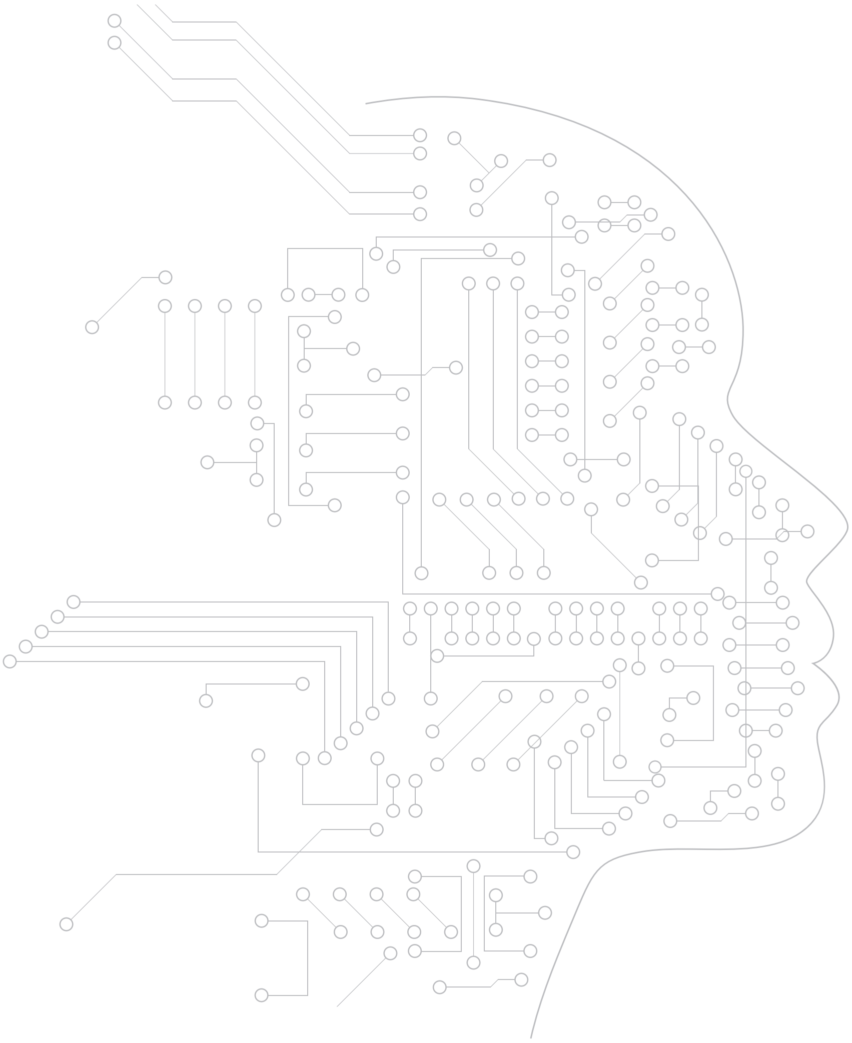
Like other similar software systems, AI systems rely on the underlying software for operation, such as operating systems and drivers. Many application-layer software does not protect their models. Even when models are protected or encrypted by applications, malicious actors can use simple dynamic analysis techniques to extract model structures, training data, and other elements from the application layer. Therefore, when permissions are improperly configured, or malicious actors gain elevated system permissions through vulnerabilities or other techniques, models in many commercial products may be stolen, including scenario-specific modules used for facial recognition, liveness detection, ID card or bank card recognition, and malware detection. In the AI computing environment, GPUs/NPUs are widely used to accelerate model training and inference. Therefore, malicious actors may exploit potential vulnerabilities in specific GPU drivers to implement malicious behaviors such as code execution, privilege escalation, model theft, and data tampering.

Security Risks of Third-Party Components

Currently, the AI field has seen a complex ecosystem. The AI computing environment typically contains numerous software components, which may present potential threats. For example, there are a large number of open-source and third-party components in AI systems. Malicious actors can exploit zero-day or unremediated vulnerabilities in these components to implement malicious behaviors such as code execution and model theft. Furthermore, container technologies such as Docker are widely used in the current AI computing ecosystem to simplify the training, deployment, and inference processes, and the KubeFlow framework is used to deploy machine learning tasks in Kubernetes clusters. However, Docker has long faced security threats related to file system isolation, process and communications isolation, device management and host resource restriction, network isolation, and image transmission. Improper configurations may lead to malicious behaviors such as container escape and permission abuse, directly threatening critical AI models and training data.

3.3.5 Continuous Risk Evolution

With the development of AIGC, AI has gradually become a continuously operated online service. The preceding cyber security risks faced by the data, models, AI applications, and computing backbone are not limited to the development and deployment phases. Instead, these risks persist throughout the operations phase of AI systems, plagued by other challenges such as continuous evolution and fast iteration of malicious behavior methods, as well as dynamic attack and defense.



04 AI Developments at Huawei

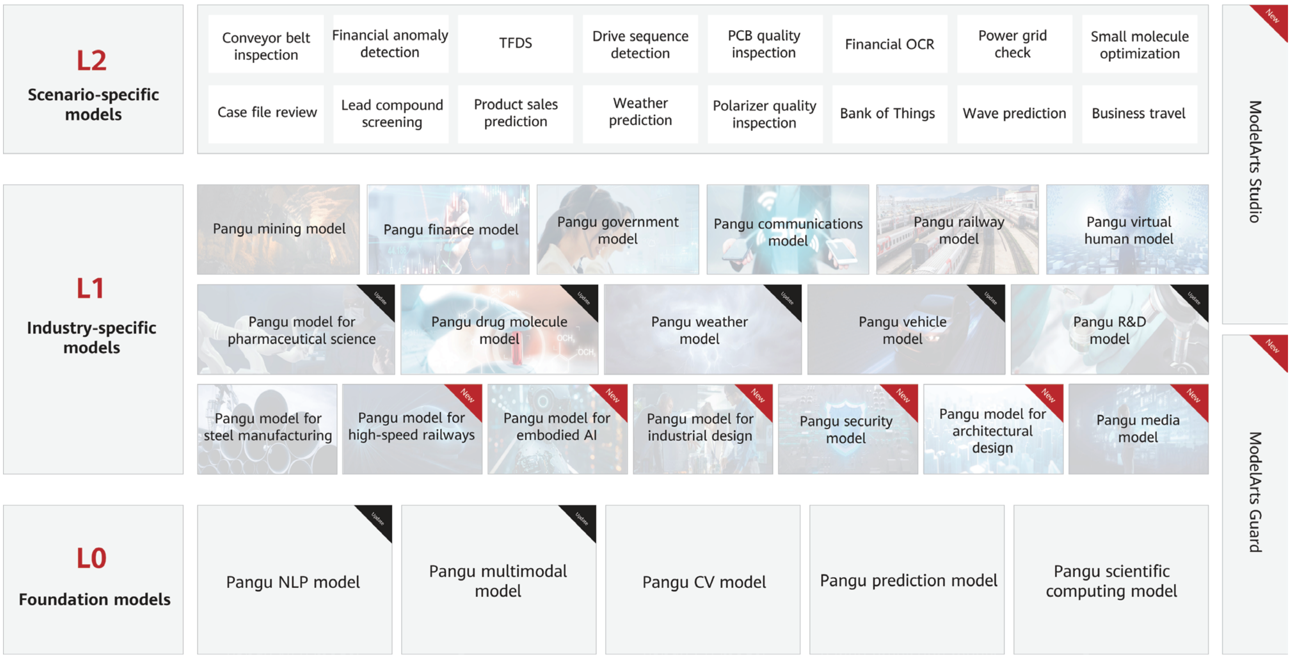
Huawei's AI model development began with the establishment of the Noah's Ark Laboratory in 2012. In 2024, Huawei launched the three-layer Pangu Models 5.0 based on its full-stack, in-house technologies.

1. L0: Foundation models, including Pangu NLP model, Pangu multimodal model, Pangu CV model, Pangu prediction model, and Pangu scientific computing model

2. L1: Industry-specific models, trained with public data of each industry based on foundation models to serve industry applications

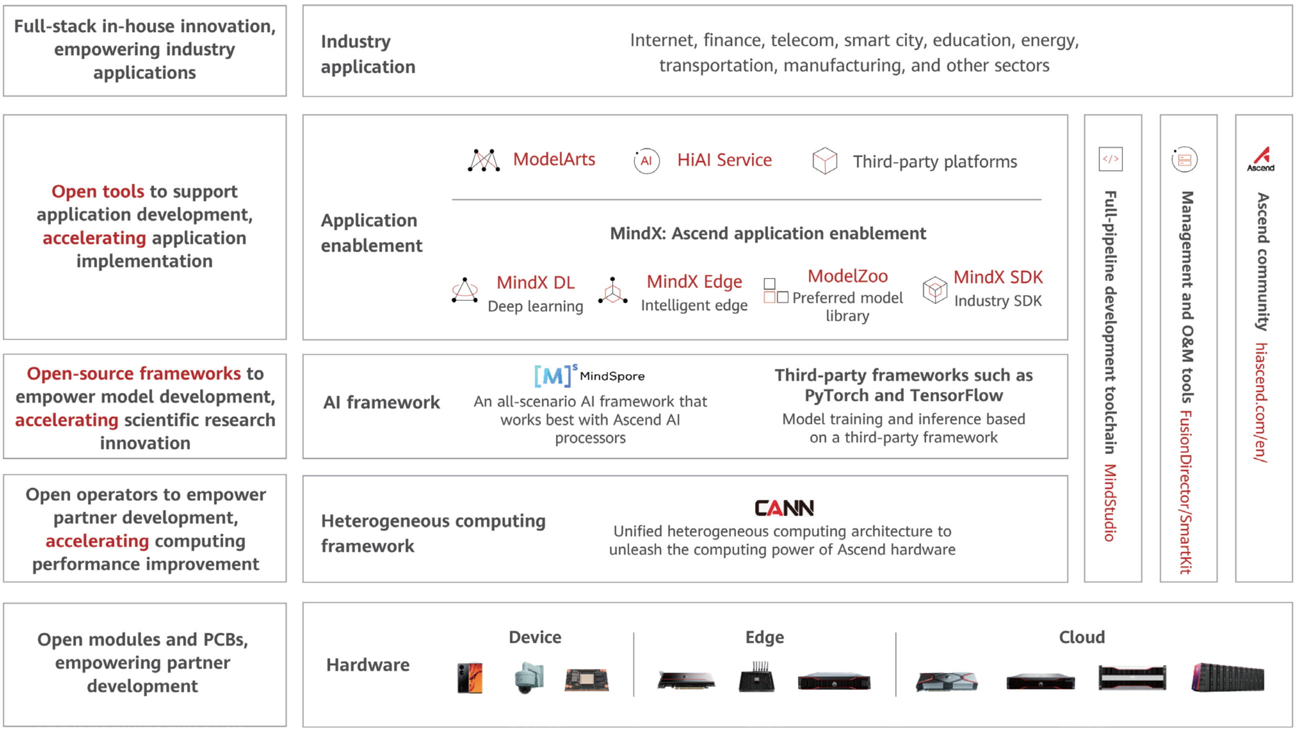
3. L2: Scenario-specific models, trained for specific application scenarios based on industry-specific models and foundation models

The following figure shows the hierarchical architecture of Pangu Models.



Huawei is always committed to developing full-stack AI software and hardware platforms based on in-house technologies. In 2018, Huawei launched the Ascend 310 chip and a full series of inference products, including modules, accelerator cards, and edge stations. In 2020, Huawei released MindSpore, an open-source tool for AI development. In 2023, Huawei unveiled its first AI cluster with over 10,000 GPUs/NPUs. After years of development, Huawei has developed Ascend chips, Ascend hardware, Compute Architecture for Neural Networks (CANN), MindSpore AI model development framework, AI application development enablement tools, and more.

The Ascend AI industry delivers AI computing power based on Huawei's in-house AI software and hardware platforms. Adhering to the strategy of "open hardware, open-source software, partner empowerment, and talent development," the Ascend AI industry works with technical and business partners to build an AI industry featuring "joint contribution, shared benefits, and win-win outcomes." It is dedicated to making AI "affordable, effective, and reliable", empowering social development and industry upgrade, and bringing value to human society.



Full-stack in-house innovation, empowering industry applications

Industry application

Internet, finance, telecom, smart city, education, energy, transportation, manufacturing, and other sectors

Open tools to support application development, **accelerating** application implementation

Application enablement

ModelArts | HiAI Service | Third-party platforms

MindX: Ascend application enablement

MindX DL Deep learning | MindX Edge Intelligent edge | ModelZoo Preferred model library | MindX SDK Industry SDK

Open-source frameworks to empower model development, **accelerating** scientific research innovation

AI framework

MindSpore: An all-scenario AI framework that works best with Ascend AI processors

Third-party frameworks such as PyTorch and TensorFlow: Model training and inference based on a third-party framework

Open operators to empower partner development, **accelerating** computing performance improvement

Heterogeneous computing framework

CANN: Unified heterogeneous computing architecture to unleash the computing power of Ascend hardware

Full-pipeline development toolchain MindStudio

Management and O&M tools FusionDirector/SmartKit

Ascend community hiascend.com/en/

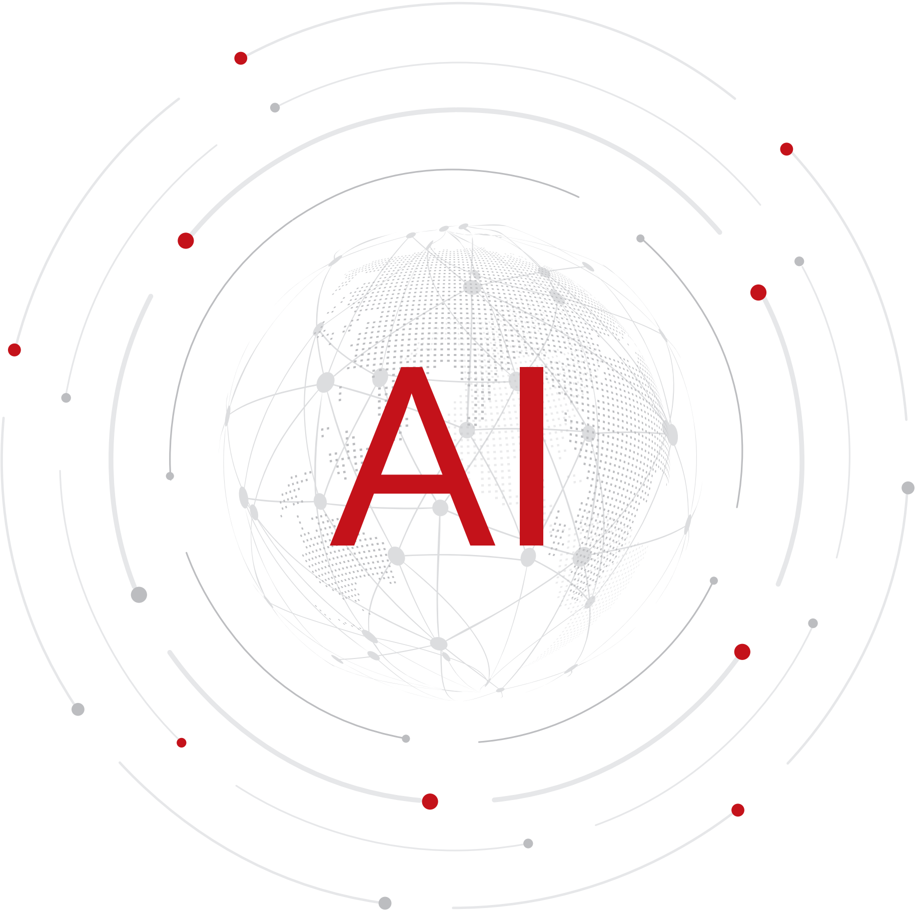
Open modules and PCBs, empowering partner development

Hardware

Device | Edge | Cloud

Huawei actively contributes its expertise to the society. Every year, Huawei publishes 200 to 300 papers at international top conferences such as NeurIPS, ICML, CVPR, and ACL. Recently, Huawei has received the Best and Excellent Paper Awards in areas including optimization theory, model quantization, optimizers, and generative model theory.

As AI grows stronger, it brings both significant opportunities/benefits and increasing cyber security issues/challenges. Huawei's AI systems span multiple business domains. To address the issues in these domains, Huawei continuously explores effective methods to protect AI systems, ensure compliance with market admission rules, and guarantee the operational security of AI systems. As such, we have proposed practical governance practices.



05 Huawei's AI Cyber Security Governance Practices

The considerations and approaches to cyber security risk management are applicable to the design, development, deployment, evaluation, and use of AI systems. Cyber security and privacy risks, including those specific to AI, are also integrated into broader enterprise risk management strategies.

5.1 Overall Principles and Responsibilities of AI Cyber Security and Privacy Governance

Huawei has a mature cyber security and privacy system responsible for AI cyber security and privacy governance. The main governance principles are as follows:

Security-oriented: In Huawei, supporting and ensuring cyber security and privacy of customers and Huawei-operated businesses is a basic business requirement. Huawei ensures that sufficient resources are invested in meeting cyber security and privacy requirements.

Legal and regulatory compliance: Huawei complies with national laws and regulations and industry standards to ensure legitimate and compliant operations of AI services.

Integration into business: Huawei integrates cyber security and privacy assurance activities into the policies, processes, specifications, and baselines of related businesses, and provides secure and trustworthy products, solutions, and services for customers, so as to guarantee business success.

Director accountability: Directors of business departments at all levels are the primary owners for ensuring cyber security and privacy within their businesses. Process owners at all levels take first-person responsibility for ensuring cyber security and privacy throughout the processes under their charge.

Engagement of all staff: All employees are aware of the importance of cyber security and privacy, have related capabilities, and implement cyber security and privacy requirements in practice. Every one shall be technically and legally accountable for what they do and for the consequences of their actions.

Independent verification: Trust needs to be based on facts, facts must be verifiable, and verification must be based on common standards. On this basis, Huawei performs independent evaluation and verification by layer following the ABC principle — assume nothing, believe no one, check everything.

Openness and cooperation: Being open and transparent, Huawei sincerely and actively communicates and cooperates with stakeholders, such as customers, suppliers, partners, and industry organizations, to implement the concepts of "shared responsibility, capability co-construction, and value sharing" so as to address cyber security and privacy threats and challenges.

Continuous improvement: Cyber security and privacy protection are a continuous risk management and capability building process. There is no absolute security or once-for-all solution. We need to promptly review weaknesses and continuously strengthen the suitability, adequacy, and effectiveness of cyber security and privacy management and technical measures.

The cyber security and privacy system is responsible for:

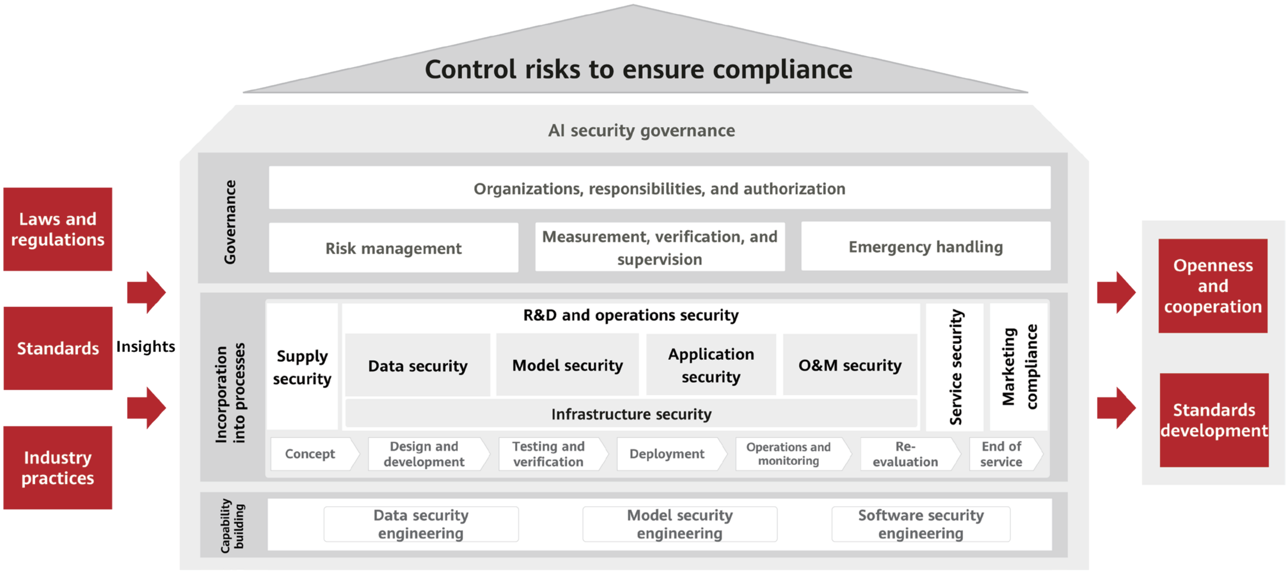
Compliance and risk control: Establish a comprehensive cyber security and privacy compliance system, clarify compliance responsibilities, comply with all applicable cyber security and privacy laws and regulations, implement the "one country, one policy" principle, develop the capability of quickly responding to compliance issues, and prevent risks from spreading.

Product security and privacy assurance: Build product security and trustworthiness along the end-to-end process, clarify product security responsibilities, and ensure fairness and non-discrimination among customers.

Secure operations: Continuously improve the management rules and processes of Huawei-operated business based on the principles of "clear responsibilities and business accountability" and "category-based management and hierarchical protection," enhance cyber security operations and maintenance (O&M) capabilities, strengthen implementation, business self-checks, and internal independent supervision and inspection, and ensure business compliance and secure operations.

Communication and trust building: Conduct "focused, pragmatic, flexible, and effective" communication, build trust through communication, and establish a complete and effective communication mechanism to gain stakeholders' trust and support business operations.

In line with Huawei's overarching cyber security and privacy protection framework, we adhere to the eight governance principles and implement the four responsibilities regarding cyber security and privacy protection. Based on industry standards and Huawei's established practices, we combine governance and embedded management to achieve the objectives of AI cyber security risk control and legal and regulatory compliance, and build an AI cyber security governance architecture.



5.2 Legal and Regulatory Compliance

With the rapid development and widespread application of AI technologies, the regulatory requirements for AI have progressively improved, and AI-related laws and regulations have been gradually released worldwide. Recognizing the complexity and changeability of these laws and regulations, Huawei employs legal experts globally to continuously track, identify, label, and categorize applicable legal requirements on AI. Based on these legal requirements, Huawei has developed its own cyber security strategies and compliance policies, which serve as the strategic framework and baseline to ensure that cyber security compliance requirements are incorporated into the end-to-end business practices and product lifecycle management, ranging from product development to service delivery and support services. This ensures that Huawei's AI systems comply with applicable cyber security laws and regulations of different countries and regions.

5.3 Governance

In terms of governance, we have specified responsibilities and achieved the risk control objectives. Through comprehensive AI risk evaluation and analysis, we prevent and reduce possible AI cyber security and privacy issues, determine responsibilities of each party, and measure, verify, and assess the achievement of results. In addition, crisis exercises are conducted based on major AI scenarios to continuously improve AI cyber security governance.

5.3.1 Organizations, Responsibilities, and Authorization

To implement Huawei's AI business intent and governance principles, Huawei's cyber security and privacy system has built a long-term mechanism for AI cyber security and privacy governance. This effort ensures that Huawei's AI products, solutions, and services as well as Huawei subsidiaries' data processing activities comply with applicable laws and regulations on AI cyber security and privacy. The cyber security and privacy system covers fields such as the R&D, marketing, service, supply, procurement, and manufacturing. It aims to ensure effective implementation of AI cyber security requirements in all regions and processes of Huawei around the world.

To adapt to Huawei's business architecture with multiple business forms, we authorize business units and regions to make decisions and operate independently under the corporate strategic control on the premise of systematic risk prevention.

5.3.2 Risk Management

We identify all AI risks in business scenarios, business processes, and data flows based on applicable laws and regulations, stakeholders' requirements, and technology development trends, as well as objectives and needs of Huawei's business organizations.

We establish AI risk evaluation standards, and organize the business, legal, and risk management experts to determine the risk level of each scenario based on the possibility and impact of risks. We make a business risk map. For medium and high risks, we set measurable and quantifiable objectives and invest resources to preferentially address the risks as planned.

Based on AI risk mitigation objectives, we develop short- and long-term measures and long-standing mechanisms, such as business rules, processes and IT tools, responsibility systems, technical specifications, and training programs, to manage risks. We integrate AI risk management into the corporate business processes, and use IT tools to manage risks in a visible, traceable, and more efficient manner.

We set up a status monitoring, evaluation, and check mechanism for AI risk management, develop metrics, and keep track of AI risk status. Each business organization performs regular self-checks and inspections to make sure related measures are adequate and effective. They verify that AI system security capabilities are implemented through an independent verification system.

5.3.3 Measurement, Verification, and Supervision

Huawei has established an end-to-end cyber security assurance system and incorporated requirements for cyber security, privacy protection, and software engineering capability enhancement into the lifecycle of products, solutions, services, and Huawei-operated business, covering all phases of product development and lifecycle management.

Huawei has fully implemented a multi-layer independent verification mechanism to protect cyber security and privacy. With this mechanism, Huawei can continuously verify products, solutions, services, and Huawei-operated business delivered to customers, and provide evidence-based verification results. This helps promote continuous improvement of the cyber security and privacy system and enhance the trust and confidence of stakeholders.

Based on the principles of "independence and objectivity; many eyes and many hands; professionalism and dutifulness; continuous improvement; openness and transparency," Huawei continuously builds AI testing and evaluation standards, cases, and platforms through organizations in and outside Huawei to make AI systems compliant with cyber security standards and best practices throughout the lifecycle. Identifying and assessing cyber security risks of AI systems is the basis of AI testing and evaluation. AI cyber security testing and evaluation should cover multiple aspects, such as data security, model robustness, backdoors in algorithms, adversarial example attacks, and framework defects.

Huawei has established an AI cyber security testing and evaluation system. With the continuous evolution and wider application of AI technologies, this system is continuously optimized for effective measurement, verification, and supervision, so as to ensure the secure, reliable, and responsible use of AI systems.

5.3.4 Emergency Handling

Huawei has established a mature emergency handling system and organizes crisis exercises to improve business departments' ability of handling crises. The exercise plan must clearly define the responsibilities of each role and allow the team to identify and shore up weak links. Through AI cyber security emergency exercises, the exercise plans are checked and optimized to ensure their effectiveness in case of real AI cyber security incidents.

5.4 Incorporation into Processes

Based on international laws, regulations, standards, and best practices, Huawei gradually incorporates AI cyber security governance requirements into all business processes, such as R&D, marketing, service, supply, procurement, and manufacturing, and continuously optimizes the requirements to control AI risks, thus achieving the quality objectives of AI systems.

Taking the IPD process as an example, in the management phase of a product, we develop AI governance requirements, including cyber security and privacy requirements, analyze AI scenarios and application risks, identify AI risks, and make mitigation plans. After that, we take risk mitigation measures in subsequent phases, such as system design, documentation development, integration verification, and marketing, to mitigate risks. The AI model development process provides data lifecycle specifications and traceability requirements from data collection, import, storage, training, to application, and defines execution standards for AI data authenticity and traceability, guiding through AI development and making the process controllable. We establish AI datasets and model metadata standards, adapt the standards to industry characteristics, and unify tool chains for AI development and build. This enables online AI operations and end-to-end traceability that covers raw data, datasets, models, and software, so that datasets can be managed and models can be traced. In this way, guidance is provided for the entire development process of AI products, cyber security risks are controlled, and AI products can be sold, delivered, and maintained after they are released.

5.5 Engineering Capability

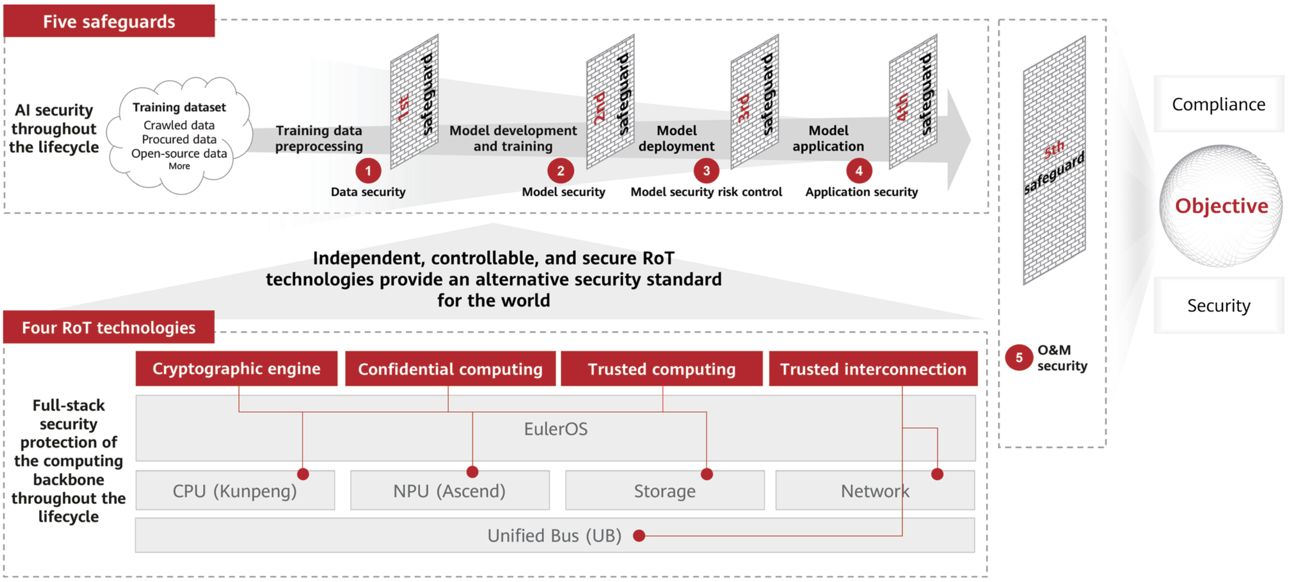
Huawei has developed an engineering capability building framework covering AI security. It continuously builds cyber security engineering capabilities, covering all needed security protection techniques. Based on software security engineering capabilities, Huawei focuses on improving engineering capabilities of data security and model security. Data security engineering capabilities include security protection measures for data collection, storage, processing, and transmission. For example, advanced data encryption techniques, access control, and data anonymization are used to prevent data leakage, tampering, and abuse and protect data integrity and confidentiality. Model security engineering capabilities include model security evaluation, defense against adversarial attacks, and model hardening, aiming to improve the robustness and protection of AI models.

5.6 Openness and Cooperation

Huawei maintains open communication with users, partners, regulators, and other stakeholders to promote AI security governance. We establish open dialog platforms to share information and discuss with stakeholders, so as to address AI security challenges together. Through cross-disciplinary and cross-industry cooperation, Huawei develops more comprehensive and forward-looking AI security standards and specifications to drive progress of the industry. In this way, we can better understand the security requirements and challenges of AI in different business domains and develop more secure and reliable AI products and services.

06 Huawei's Practices of Protecting AI System Cyber Security

It is impossible to address all AI security issues, which involve data, models, and applications, using a single method. Therefore, Huawei has built a multi-layer safeguard system to ensure compliance and security.



6.1 Four Root of Trust (RoT) Technologies that Secure the Computing Backbone

Foundation models are more and more complex as the training data increases. Model training and inference are conducted in AI computing centers, where different users share the computing power. This inevitably causes cyber security threats. In this case, model vendors must consider:

1. How to securely transmit training data and AI models from the company to the AI computing center

2. How to securely store AI models in the computing center

3. How to protect AI models during model training and inference

4. How to protect computing platform software from malicious implantation or tampering

6.1.1 Cryptographic Engine

Based on the AI computing center solution, we propose the AI Guard solution for model and data protection. It builds on the technique of secure storage and authorization of keys provided by Huawei Ascend AI computing center's AI-VAULT, a secure, trusted public facility. AI Guard uses cryptographic algorithms to protect the core asset ownership of data and model owners. The solution is as follows:

1. As the owner, the AI model vendor encrypts the training data and AI model locally and protects the encryption key.

2. The AI model vendor establishes a trust relationship with AI-VAULT in the AI computing center, and securely registers the key with AI-VAULT through a secure channel.

3. The AI model vendor sends the encrypted training data or AI model to the AI computing center.

4. The AI computing center provides integrity protection for the training or inference container image.

5. When the AI model runs in a container, it needs to pass identity authentication to obtain the decryption key stored in AI-VAULT.

6. When the AI model runs in a training and inference container, only necessary permissions are granted to protect the container.

7. In the preceding steps, the cryptographic engine built in the chip can be used for acceleration to reduce the usage of computing resources.

6.1.2 Confidential Computing

The encrypted model and data are decrypted before running on the CPU, leaving them at security risks. To address the risks, currently, there are two technical approaches:

Cryptographic computing: In this approach, cryptography is used as the RoT, and user data is encrypted before being computed. The computing does not require decryption. Cryptographic computing provides high security and avoids hardware dependency. However, the performance loss is huge in most cases.

Confidential computing: In this approach, user-encrypted models and data are decrypted and then computed in the trusted execution environment (TEE), safe from unauthorized access or tampering. Hardware-level security assurance prevents high-privilege operating systems, even hypervisors, from accessing or tampering with data and code in the TEE. This approach has little performance loss but depends on the hardware.

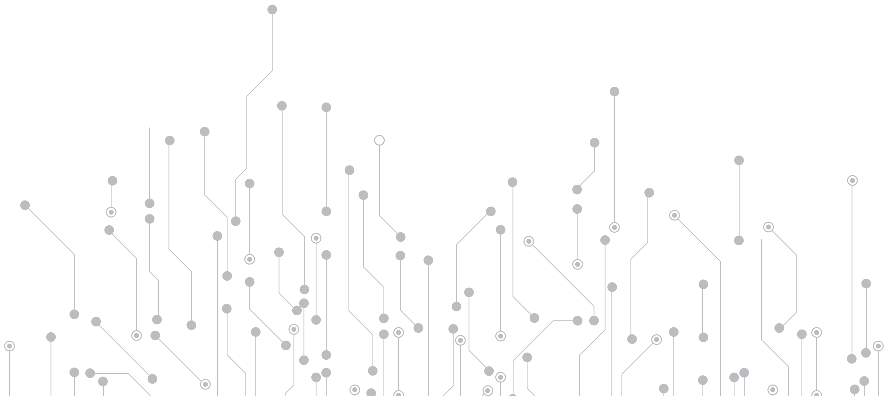
By combining the advantages of both approaches, Huawei not only implements the Virtualized Arm Confidential Compute Architecture (virtCCA) for the CPU, but also uses the innovative security technology of PCIe to securely connect NPUs to virtCCA for AI confidential computing.

6.1.3 Trusted Interconnection

Based on the PCIe, NPUs can be securely connected to the virtCCA on the CPU to implement board-level AI confidential computing. However, foundation model training requires NPU clusters. Therefore, device-to-device (D2D) high-speed interconnection is also required between NPUs. The traditional solution encrypts data transmission to ensure D2D interconnection security. Due to the latency-sensitive foundation model training based on Transformer, the traditional solution has a high performance overhead. Huawei's innovative LingQu UB interconnection protocol defines a secure channel at the bottom layer. In this way, it establishes a secure transmission channel between NPUs during D2D interconnection, providing secure D2D transmission with zero overhead.

6.1.4 Trusted Computing

To protect the computing platform software from malicious implantation or tampering, Huawei develops the trusted computing 3.0 solution based on its in-house BMC chips and China's national standards on trusted computing. Guided by the principles of "overall security and active immunity," trusted computing 3.0 uses collaborative software-hardware design to build a secure and trusted computing and communication environment for the computing platform. It enhances the proactive, dynamic, overall, and accurate protection capability of the system, and builds trusted, manageable, and controllable security protection for the computing platform. Independent trusted software and hardware for detection are used together with the service systems. In this way, security detection is performed during computing. The trusted mechanism effectively copes with advanced persistent threats (APTs) and supply chain attacks, meeting the requirements of GB/T 22239-2019 Information Security Technology — Baseline for Classified Protection of Cybersecurity for trusted verification and proactive defense.



6.2 First Safeguard: Data Security

To ensure data compliance, data security check measures are developed, and a solution is formed for training data compliance governance.

1. Dataset source check: High-risk datasets are excluded from the data production line.

2. Dataset content check: Datasets are checked for non-compliance, privacy violations, and copyright risks based on data features.

3. Model compliance check: In addition to model file integrity, the check also includes whether the corresponding dataset has non-compliance handling or tracing records.

4. Backtracking of live network issues: Backtracking is conducted for abnormal responses caused by problematic data to resolve issues that are not addressed by existing means, such as data cleansing, alignment, and risk control measures.

6.3 Second Safeguard: Model Security

Technical capabilities are systematically built for model security from many dimensions, such as security data, algorithms, testing, and evaluation. Alignment is implemented to improve the intrinsic security of models.

1. Data: Secure datasets that contain texts, images, and the combinations are continuously built. Based on these diversified datasets, Huawei uses supervised fine-tuning (SFT) to enhance the security of its models.

2. Algorithm: SFT and reinforcement learning (RL) algorithms are used to analyze generalization issues of different semantic types during alignment training in multimodal understanding. Dedicated security algorithms for multimodal understanding are designed to prevent known AI security vulnerabilities and malicious prompt attacks. This ensures model alignment at the security level.

3. Testing and evaluation: Security tests are performed both manually and automatically. This not only improves the test efficiency, but also makes the tests more comprehensive and accurate. We continuously evaluate model security and enrich security datasets to adapt to changing security threats and challenges. We have built an automated security testing and evaluation system to continuously evaluate foundation models that support Chinese and English languages, multimodal interaction, and agents.

6.4 Third Safeguard: Model Security Risk Control

This safeguard aims to mitigate risks. It safeguards security and privacy of foundation models by controlling risks at the input and output ends, and provides risk detection and defense. Foundation model applications are protected from many known security and privacy risks, such as content non-compliance, privacy breach, intellectual property right and portrait right infringement, information breach, source tracing failure, and malicious code execution.

1. Security defense: A Q&A library that delineates which questions should be answered and which should not is built and continuously optimized.

2. NLU: The risk control model and security dialog model are trained to improve the context/multi-turn dialog detection capability of the risk control model.

3. Text moderation: Moderation is provided for PDF, Word, PPT, and Excel files.

4. Image check: The capability of detecting non-compliant images is built and enhanced based on accumulated non-compliant image identifiers and elements.

6.5 Fourth Safeguard: Application Security

This safeguard aims to mitigate the risks of agents and application frameworks. To this end, Huawei has built an application safeguard solution.

First, to ensure normal operations of the agent, the task planning and execution safeguard solution is used to ensure that the core functions of task planning and execution run properly. The task planning safeguard uses RL technologies to automatically analyze tasks orchestrated by foundation models to prevent malicious tasks. The task execution safeguard checks the generated code to prevent malicious code generation and execution by foundation models, vulnerable code, calls of malicious external plug-ins, and access to malicious websites.

Second, to ensure the normal running of the application framework, technologies such as unauthorized access prevention for APIs, access control, container isolation, and operator security protection are used to mitigate risks such as unauthorized access to the application framework.

Third, to ensure application security, the full-stack security engineering capability of the traditional software system is enhanced.

6.6 Fifth Safeguard: O&M Security

In the AI system O&M phase, Huawei protects the lifecycle security of training data, models, and generated content involved in the foundation model system. We build a system to detect malicious behaviors of foundation models, improve the interception capability, and perform unified operations of the live network system. We propose to build a system that intercepts malicious behaviors through online and offline detection, so as to continuously enhance our detection capabilities.

1. Online detection capability library: To develop a solution for detecting high-risk malicious behavior patterns, it is advised to build capabilities for real-time detection of prompt attack patterns, malicious intents, and malicious behavior rules. This includes establishing a library of prompt attack patterns, such as target hijacking and reverse exposure based on adversarial tokens. It is also required to create a library of malicious intent patterns, such as non-compliant content and portraits that infringe personal privacy.

2. Offline detection capability library: An offline detection system is needed, which integrates multiple detection solutions (such as rules, classification models, and large language models' detection capabilities).

3. Malicious behavior detection (securing AI with AI): An attack detection engine is built on the Ascend card to detect malicious behaviors in both online and offline modes. The engine uses the feature detection model based on the deep neural network (DNN) to extract features of inference requests and identify the matching attack patterns. In addition, it detects AI attacks, such as physical adversarial examples and foundation model prompt injection.

4. Trustworthy audit: In the computing power environment, the Merkle tree is used to store integrity proofs to the secure hardware, building a basis for trustworthy audit of AI system logs. With this function, if a malicious actor or O&M employee tampers with a log file, the log integrity damage will be detected and a warning will be generated very fast.

Huawei Technologies Co., Ltd.
Huawei Base in Bantian, Longgang District, Shenzhen
TEL: +86 755 28780808
ZIP CODE: 518129
www.huawei.com

Trademark Notice

HUAWEI, HUAWEI, and are trademarks or registered trademarks of Huawei Technologies Co., Ltd.
Other trademarks, product, service and company names mentioned are the property of their respective owners.

General Disclaimer

This document is provided "as-is" without warranty of any kind, express or implied. All warranties are expressly disclaimed. Without limitation, there is no warranty of non-infringement, no warranty of merchantability, and no warranty of fitness for a particular purpose. Huawei assumes no responsibility for the accuracy of the information presented. Any information provided in this document is subject to correction, revision, and change without notice. Your use of, or reliance on, the information provided in this document is at your sole risk. All information provided in this document on third parties is provided from public sources or through their published reports and accounts.