***Translation***

CSET CENTER *for* SECURITY *and* EMERGING TECHNOLOGY

*The following white paper by a PRC information security standardization group describes the current state of AI security and AI safety standards in China, with reference to international standards. Appendices list all of China's current (as of October 2019) AI security standards, and describe examples of AI security innovations pioneered by Chinese tech companies (and by IBM).*

**Title**
Artificial Intelligence Security Standardization White Paper (2019 Edition)
人工智能安全标准化白皮书（2019版）

**Author**
The Big Data Security Standards Special Working Group (大数据安全标准特别工作组) of the National Information Security Standardization Technical Committee (SAC / TC 260; 全国信息安全标准化技术委员会; 全国信安标委)

**Source**
China Electronics Standardization Institute (CESI; 中国电子技术标准化研究院; 电子标准院) website, November 1, 2019. The white paper itself is dated October 2019. CESI is a think tank under the PRC Ministry of Industry and Information Technology (MIIT; 工业和信息化部); CESI is also known as MIIT 4th Electronics Research Institute (电子第四研究院; 电子四院).

*The Chinese source text is available online at:*
http://www.cesi.cn/images/editor/20191101/20191101115151443.pdf
*Translator's notes are denoted by superscript letters ([a]) in this translation and do not appear in the Chinese source text. Endnote references appear as bracketed superscript numbers ([1]), and are English translations of the endnote references in the Chinese source text.*
*US $1 ≈ 7 Chinese Yuan Renminbi (RMB), as of May 14, 2020.*

| **Translation Date** | **Translator** | **Editor** |
|---|---|---|
| May 14, 2020 | Etcetera Language Group, Inc. | Ben Murphy, CSET Translation Lead |

# Preface

During the ninth collective study session of the 19th Central Committee Politburo, General Secretary Xi Jinping clearly pointed out that it is necessary to strengthen the determination of the potential risks of the development of artificial intelligence and to strengthen our watchfulness against them, to safeguard the interests of the people and national security, and to ensure the security, reliability, and control of artificial intelligence. After more than 60 years of development, artificial intelligence (AI) has evolved into a discipline for research and development to simulate, extend, and expand human intelligence. In recent years, driven by the three major factors of algorithms, computing power, and data, AI has entered a new stage of accelerated development, becoming a leader of economic development and an accelerator for social development. At present, major countries across the world all regard AI as a national development strategy.

In 2017, China released the New Generation Artificial Intelligence Development Plan[a] to develop a new generation of AI at the national strategic level. With the deep integration of AI in related industries and in people's lives, the risks and challenges to national security, ethics, cybersecurity, personal safety, and privacy protection at multiple levels have also garnered widespread attention from society.

AI security[b] standardization is an important component of the development of the AI industry. It plays a fundamental, normative, and leading role in stimulating healthy and benign AI applications and in promoting the orderly and healthy development of the AI industry. The New Generation Artificial Intelligence Development Plan clearly proposes that "it is necessary to strengthen research into artificial intelligence standard framework systems to gradually establish and improve the technical standards of artificial intelligence basic commonality, interconnection, industry applications, cybersecurity, and privacy protection." Effectively strengthening AI security standardization is the only way to ensure the security of AI.

In order to promote the healthy, rapid, safe, and orderly development and expansion of AI technology applications, the Big Data Security Standards Special Working Group under the National Information Security Standardization Technical Committee (NISSTC) initiated the formulation of the Artificial Intelligence Security Standardization White Paper. This white paper primarily focuses on the security of AI itself, analyzes in detail the current state of AI development, the main security threats, risks, and challenges, and summarizes the progress of standardization in terms of domestic and foreign AI security regulations and standardization organizations. Upon this basis, we have conducted an in-depth analysis of the needs of AI security standardization and have put forward a framework for AI security standards and recommendations for standardization.

[a] Translator's note: For an English translation of the New Generation Artificial Intelligence Development Plan, see: https://www.newamerica.org/cybersecurity-initiative/digichina/blog/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017/

[b] Translator's note: The Chinese word 安全 can be translated as either "safety" or "security." The translator deems that the latter translation, "security," is more appropriate throughout most of this white paper. In cases where the appropriate translation is debatable, this translation defaults to "security."

# AI Security Standardization White Paper (2019)

**Institutional Authors**

China Electronics Standardization Institute
Tsinghua University
Beijing Baidu Network Information Technology Co., Ltd.
Huawei Technologies Co., Ltd.
360 Technology Group Co., Ltd.
Alibaba (China) Co., Ltd.
China Mobile Communications Group Co., Ltd.
Renmin University of China
Ant Financial Services Group
IBM (China) Investment Co., Ltd.
Beijing Topsec Network Security Technology Co., Ltd.
Lenovo (Beijing) Limited
Shanghai Yitu Network Technology Co., Ltd.
Sangfor Technologies Inc.
Shenzhen Tencent Computer System Co., Ltd.
Beijing Sankuai Online Technology Co., Ltd. (Meituan-Dianping)
Qi An Xin Technology Group Inc. (奇安信科技集团股份有限公司)
Shaanxi Network and Information Security Assessment Center
Beijing OrionStar Technology Co., Ltd.
Institute of Automation, Chinese Academy of Sciences
Sichuan University
Big Data Development Administration of Inner Mongolia Autonomous Region
Vivo Communication Technology Co. Ltd.
Peking University
Beijing Shenzhou NSFOCUS Information Security Technology Co., Ltd. (北京神州绿盟信息安全科技股份有限公司)
Alibaba Cloud Computing Co. Ltd.
Information & Data Security Solutions Co., Ltd.
Guangdong OPPO Mobile Telecommunications Corp., Ltd.
Ping An Insurance (Group) Company of China, Ltd.

# AI Security Standardization White Paper (2019)

## Authors

Yang Jianjun Liu Xiangang Wang Jianming Hu Ying Zhang Yuguang Su Hang Liu Yan
Zhang Yi Li Shi Guo Rui Zhu Hongru Wang Xiaopu Cheng Haixu Shangguan Xiaoli
Zhang Feng Luo Hongwei Xu Feiyu Xie Anming Wu Yuesheng Wang Yan Zhao Chunhao
Chen Xingshu Ye Xiaojun Jin Tao Liu Bozhong Wu Yang Luo Zhibing He Ran
Quan Xin Su Yongzi Wu Zijian Wei Yufeng Zheng Xinhua Xie Jiang Jia Ke
Liu Xing Yan Minrui Liu Jianxin He Yuan Yu Le Zhu Jun Li Yi
Bai Xiaoyuan Du Yangzhou Zhou Jun Li Ruxin Wang Haitang Cao Xiaoqi Bao Xuhua
Zhang Dajiang Jiang Weiqiang Chang Ling Peng Juntao Ning Yang Fu Ronghua Wang Jiangsheng
Wang Yanhui Zhao Xiaona Bao Shenfu Zhao Xinqiang Gong Jing Ma Jie Sun Wei
Liu Xiaocen Li Yi Lei Xiaofeng Yu Jingtao Bian Songshan Zhang Hongwei Huang Hanchuan
Yang Fan Li Qingshan Xia Yunming Han Fang Cai Wei

# Contents

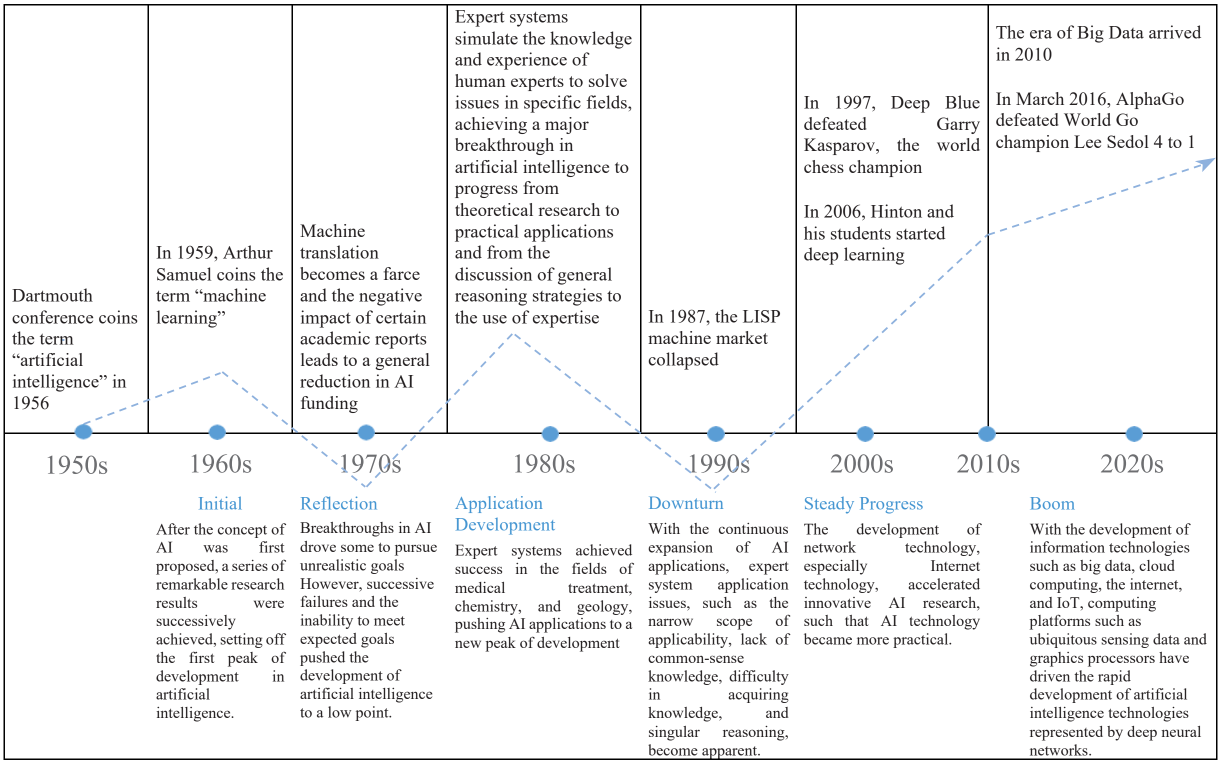
# 1 Outline of artificial intelligence (AI)

Artificial intelligence (AI) is a system of theories, methods, technologies, and applications that uses digital computers and digital computer-controlled machines to simulate, extend, and expand human intelligence, perceive the environment, acquire knowledge, and use knowledge to achieve optimal results[1]. The purpose of research into AI-related technologies is to encourage intelligent machines to listen (e.g., speech recognition and machine translation), to read (e.g., image recognition and text recognition), to speak (e.g., speech synthesis and human-machine dialogue), to act (e.g., robots and autonomous vehicles), to think (e.g., human-machine game-playing and theorem proofs), and to learn (e.g., machine learning, knowledge representation)[2].

## 1.1 AI ushers in third wave of development

As early as 1950, Alan Turing elaborated on thinking about artificial intelligence in “Computing Machines and Intelligence” and proposed to measure machine intelligence with the Turing test. In 1956, the artificial intelligence conference held at Dartmouth College first proposed the concept of "artificial intelligence" - allowing machines to recognize, think, and learn like humans, which marked the beginning of artificial intelligence.

AI entered peak development twice - once in the late 1950s and once in the early 1980s - but because of constraining factors such as technology and cost, development entered a period of slow growth (see Figure 1-1). In recent years, with the development of information technologies such as big data, cloud computing, the internet, and the Internet of Things (IoT), computing platforms such as ubiquitous sensing data and graphics processors have driven the rapid development of artificial intelligence technologies represented by deep neural networks[2]. The third wave of AI development was driven by the three major factors of algorithms, computing power, and data.



Dartmouth conference coins the term “artificial intelligence” in 1956

In 1959, Arthur Samuel coins the term “machine learning”

Machine translation becomes a farce and the negative impact of certain academic reports leads to a general reduction in AI funding

Expert systems simulate the knowledge and experience of human experts to solve issues in specific fields, achieving a major breakthrough in artificial intelligence to progress from theoretical research to practical applications and from the discussion of general reasoning strategies to the use of expertise

In 1987, the LISP machine market collapsed

In 1997, Deep Blue defeated Garry Kasparov, the world chess champion

In 2006, Hinton and his students started deep learning

The era of Big Data arrived in 2010

In March 2016, AlphaGo defeated World Go champion Lee Sedol 4 to 1

Initial: After the concept of AI was first proposed, a series of remarkable research results were successively achieved, setting off the first peak of development in artificial intelligence.

Reflection: Breakthroughs in AI drove some to pursue unrealistic goals However, successive failures and the inability to meet expected goals pushed the development of artificial intelligence to a low point.

Application Development: Expert systems achieved success in the fields of medical treatment, chemistry, and geology, pushing AI applications to a new peak of development

Downturn: With the continuous expansion of AI applications, expert system application issues, such as the narrow scope of applicability, lack of common-sense knowledge, difficulty in acquiring knowledge, and singular reasoning, become apparent.

Steady Progress: The development of network technology, especially Internet technology, accelerated innovative AI research, such that AI technology became more practical.

Boom: With the development of information technologies such as big data, cloud computing, the internet, and IoT, computing platforms such as ubiquitous sensing data and graphics processors have driven the rapid development of artificial intelligence technologies represented by deep neural networks.

Figure 1 -1 AI Development Timeline

## 1.2 Striking progress made in AI technology and applications

AI technology continues to grow, with significant progress made in machine learning algorithms such as deep learning and perceptual intelligence technologies such as speech recognition, natural language processing, and image recognition. Specialized AI refers to AI in specific fields. Because of its singular tasks, clear requirements, clear application boundaries, rich domain knowledge, and relatively simple modeling, it has achieved single-point breakthroughs in computer vision, speech recognition, machine translation, human-computer game-playing, and other aspects that can approach or even exceed human levels.

Meanwhile, key AI technologies, such as machine learning, knowledge graphs, and natural language processing have moved from the laboratory to the marketplace (see Figure 1-2). **Machine learning**, which mainly studies functional units such as computers, is a process of acquiring new knowledge or skills by simulating human learning or reorganizing existing knowledge or skills to improve its performance. As an emerging field in machine learning research, deep learning was proposed by Hinton et al in 2006. **Deep learning**, also known as deep neural networks (neural networks with more than 3 layers), is a method based on representation learning of data in machine learning. In traditional machine learning, hand-designed features are very important for learning effects, but feature engineering is very cumbersome, and deep learning is based on multi-level neural networks, which can automatically learn features from big data. This results in such characteristics as complex model scales, efficient process training, and accurate result training[3]. **The knowledge graph**, which is essentially a structured semantic knowledge base, is a graphic data structure composed of nodes and edges, describing concepts in the physical world and their interrelationships in the form of symbols. **Natural language processing** studies various theories and methods to achieve effective communication between humans and computers in natural language. **Human-computer interaction** mainly studies the exchange of information between humans and computers, including both the exchange of information from humans to computers and from computers to humans. **Computer vision** is the science of using computers to imitate the human visual system, giving computers the ability to extract, process, understand, and analyze images and image sequences in a way that is similar to that of humans. **Biometric feature recognition** refers to technologies that identify and authenticate an individual’s identity through individual physiological or behavioral characteristics. **Intelligent speech** mainly studies the perception, analysis, and synthesis of information represented by human speech through functional units such as computers.

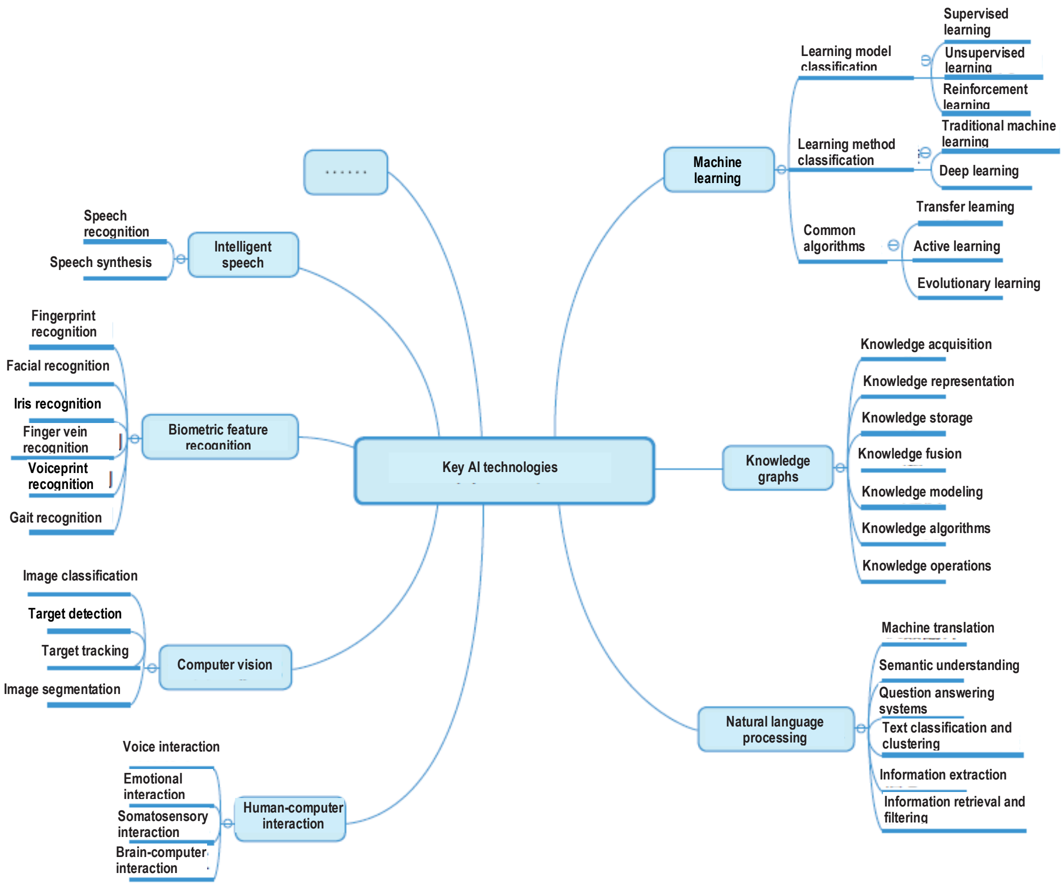
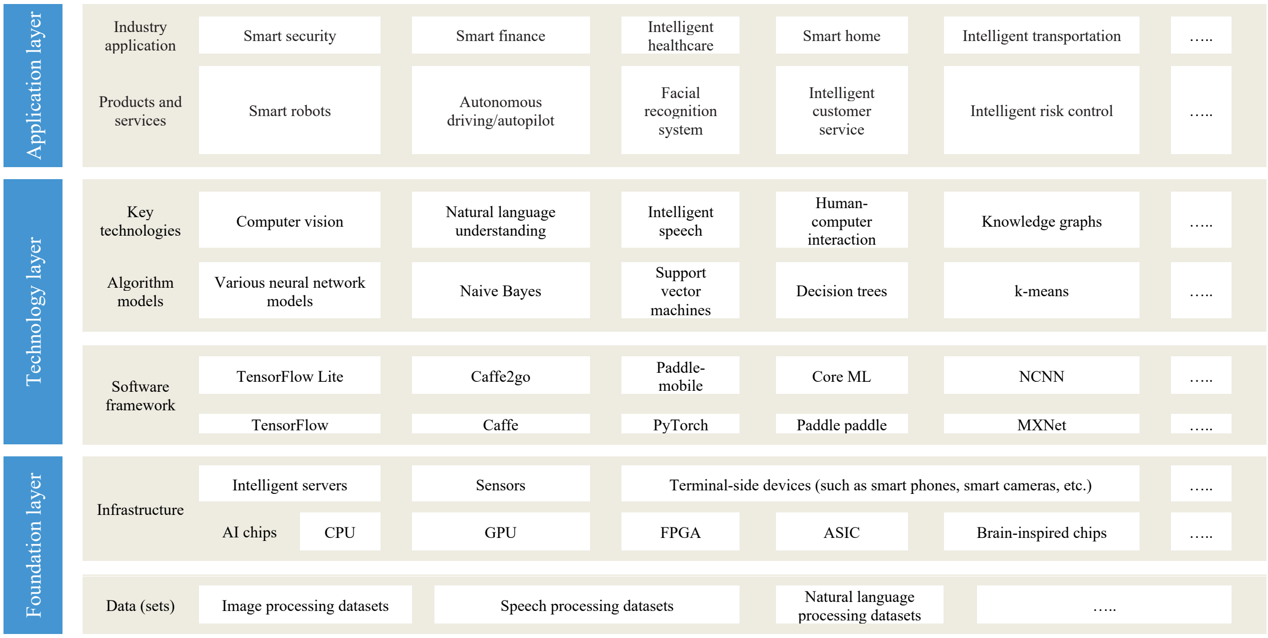


Figure 1-2 Key AI technologies

## 1.3 AI production chain beginning to take shape

From a global perspective, the AI market is developing rapidly and has huge potential. As the world's most important vehicle of AI innovation and applications, China continues to exert efforts to promote the rapid development of the AI industry. In 2018, the scale of China's AI industry was about 34.4 billion yuan Renminbi (RMB), and financing in the AI field reached RMB 79.69 billion[5]. As of September 2019, the total number of AI-related companies in China has exceeded 2,500[5]. Most are engaged in computer vision, speech recognition, language technology processing, and other related businesses[6]. The New Generation Artificial Intelligence Development Plan further proposed that the scale of core domestic AI industries should exceed RMB 1 trillion by 2030 and drive related industries to exceed RMB 10 trillion.

At present, the global AI industry chain has begun to take shape, forming a multi-layer industrial structure (see Figure 1-3), wherein the basic layer is basic support for AI, providing basic resources such as computing power and data; the technical layer is the technical system of AI, providing software frameworks, algorithm models, and key technologies for algorithm development; and the application layer implements AI applications, providing artificial intelligence products, services, and industry application solutions.



| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Application layer | Industry application | Smart security | Smart finance | Intelligent healthcare | Smart home | Intelligent transportation | ….. |
| | Products and services | Smart robots | Autonomous driving/autopilot | Facial recognition system | Intelligent customer service | Intelligent risk control | ….. |
| Technology layer | Key technologies | Computer vision | Natural language understanding | Intelligent speech | Human-computer interaction | Knowledge graphs | ….. |
| | Algorithm models | Various neural network models | Naive Bayes | Support vector machines | Decision trees | k-means | ….. |
| | Software framework | TensorFlow Lite | Caffe2go | Paddle-mobile | Core ML | NCNN | ….. |
| | | TensorFlow | Caffe | PyTorch | Paddle paddle | MXNet | ….. |
| Foundation layer | Infrastructure | Intelligent servers | Sensors | Terminal-side devices (such as smart phones, smart cameras, etc.) | | | ….. |
| | | AI chips CPU | GPU | FPGA | ASIC | Brain-inspired chips | ….. |
| | Data (sets) | Image processing datasets | Speech processing datasets | | Natural language processing datasets | | ….. |

Figure 1-3 AI supply chain structure

The AI supply chain mainly involves smart infrastructure vendors, smart information and data providers, smart technology service vendors, smart product and service providers, and smart application solution providers. **Smart infrastructure vendors** mainly include vendors of smart chips, smart servers, terminal-side devices (端侧设备), and other hardware that provides basic computing power support for AI and sensors. **Smart information and data providers** mainly include dataset providers and vendors that provide services related to AI data acquisition, labeling, analysis, and processing. **Smart technology service providers** rely on infrastructure and large amounts of data to provide smart technology services, mainly including software frameworks or technology platforms for AI, AI algorithm models and key technical consulting, and online AI services. **Smart product and service providers**, that is, manufacturers of AI products and services, such as intelligent robots, intelligent vehicles, intelligent devices, biometric feature recognition products, natural language understanding products, and computer vision products[7]. These products and services can exist in the form of hardware and software products, cloud services, and API services. **Smart application solution providers** are AI solutions providers in vertical industries and application scenarios such as smart finance, smart manufacturing, smart transportation, smart home, smart healthcare, and smart finance.

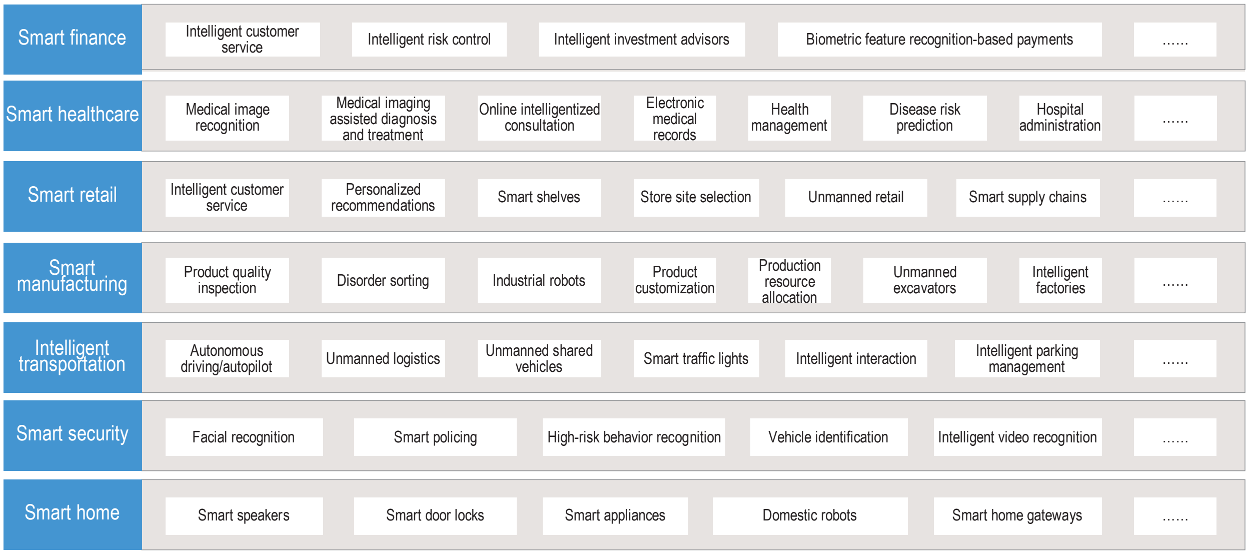
## 1.4 China has a vast range of AI application scenarios

In recent years, China has successively unveiled a number of policies to promote the development of the AI industry and promote the deep integration of AI into the economy and society from various angles. The Ministry of Industry and Information Technology (MIIT) issued the Three-Year Action Plan for Promoting Development of a New Generation Artificial Intelligence Industry (2018–2020),[c] a meeting of the Central Commission for Comprehensively Deepening Reform (中央深改委) reviewed and adopted the Guiding Opinions on Promoting the Deep Integration of Artificial Intelligence and the Real Economy (关于促进人工智能和实体经济深度融合的指导意见), and the Ministry of Science and Technology (MOST) issued the Guidelines for National New Generation Artificial Intelligence Innovation and Development

[c] Translator's note: For an English translation of this action plan, see: https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-chinese-government-outlines-ai-ambitions-through-2020/

Pilot Zone Construction Work.[d] As of 2019, 15 national AI open platforms have been established nationwide, and 21 provincial, municipal, and regional governments have issued policies related to the AI industry[5].

Driven by such factors as national and local policy support and abundant data resources, China's vast AI applications market has become a development advantage. From the perspective of vertical industries (see Figure 1-4), AI is used in such industries as security, finance, retail, healthcare, government, transportation, manufacturing, and the home. Relevant reports show that robotics, security, and healthcare have become popular application scenarios, and the fields of smart military, smart writing, and unmanned surface vessels are, relatively, in their infancy[5]. AI has gone furthest in financial applications, has grown in various aspects of retail, developed rapidly in intelligent medical industry applications, offers broad development prospects in the government and security fields, and has potential for further development in manufacturing[6].



| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Smart finance | Intelligent customer service | Intelligent risk control | Intelligent investment advisors | Biometric feature recognition-based payments | ...... | | | |
| Smart healthcare | Medical image recognition | Medical imaging assisted diagnosis and treatment | Online intelligentized consultation | Electronic medical records | Health management | Disease risk prediction | Hospital administration | ...... |
| Smart retail | Intelligent customer service | Personalized recommendations | Smart shelves | Store site selection | Unmanned retail | Smart supply chains | ...... | |
| Smart manufacturing | Product quality inspection | Disorder sorting | Industrial robots | Product customization | Production resource allocation | Unmanned excavators | Intelligent factories | ...... |
| Intelligent transportation | Autonomous driving/autopilot | Unmanned logistics | Unmanned shared vehicles | Smart traffic lights | Intelligent interaction | Intelligent parking management | ...... | |
| Smart security | Facial recognition | Smart policing | High-risk behavior recognition | Vehicle identification | Intelligent video recognition | ...... | | |
| Smart home | Smart speakers | Smart door locks | Smart appliances | Domestic robots | Smart home gateways | ...... | | |

Figure 1-4 AI industrial applications

## 1.5 Overall AI development level remains in the initial stage

At present, AI has ushered in a new wave of development driven by big data, computing power, and deep learning algorithms. However, considering that the purpose of AI is to "explore the essence of intelligence," humans have yet to develop smart machines with human-like intelligence. Thus, overall, AI is still in its infancy, mainly in three aspects:

**First, much still remains to be accomplished in the research and application of AI**[2]. Although breakthroughs have been made in the field of dedicated AI, dedicated intelligence tends to be weak AI, that is, lacking in autonomy, and cannot truly realize conceptual abstraction, inference decision-making, or problem-solving. Artificial general intelligence (AGI), also called “strong AI” or “humanlike machine intelligence” (类人智能), refers to self-aware AI that can reach the human level and can adaptively respond to external environmental challenges. For instance, the human brain is a general-purpose intelligent system that can learn by analogy and can achieve mastery through comprehensive study. It can see, listen, judge, reason, learn, think, plan, and design for various issues. General intelligence is still in its infancy, and there are still many gaps between general AI and human intelligence (人类智能).

[d] Translator's note: For an English translation of these guidelines, see: https://cset.georgetown.edu/wp-content/uploads/t0110_AI_pilot_zones_EN.pdf

**Second is that there are still many bottlenecks in AI technology that prevent it from being "easy to use."** Although there has been significant progress in information perception and machine learning, conceptual abstraction and planning decisions have only just begun. Deep learning-based AI is more suitable for processing static, single-domain, and single-task application scenarios with rich data or knowledge and complete, deterministic information[9]. As such, it is not a solution for AGI. It can be said that AI is at a technical inflection point between "impractical" and "practical." Along the way, there are still many bottlenecks such as explicability, generalization, energy consumption, reliability, and robustness before it can become "easy to use[10].”

**Third, investment in the AI industry is growing more rational, and the application of complex scenarios will take time.** Since the second quarter of 2018, global investment in the field of AI has gradually declined. In the first quarter of 2019, global AI financing amounted to USD 12.6 billion, a decrease of 7.3% from the previous quarter. Specifically, Chinese financing in the field of AI amounted to USD 3 billion, a year-on-year decrease of 55.8%[11]. In terms of solving practical and complex issues, a large number of tailor-made systems have been able to win in specific field challenges, such as Go and chess, but they cannot meet the requirements of uncertain, dynamic, and multi-domain complex application scenarios.

From another perspective, however, the cooling of industrial investment and the difficulty of implementing technology applications can help people consider the study and application of AI from a more rational perspective. To this end, it is necessary to fully respect the development laws of AI technology and its degree of maturity when strengthening the basic science and key technology research that goes into AI and excavating the pain points and difficulties[18] of specific AI application scenarios to build China's upstream and downstream AI industrial chain. Meanwhile, we should also pay attention to the security risks and challenges that the development and application of AI may bring, reduce the potential threats of AI in complex environments and extreme conditions, and promote the healthy, positive, and orderly development of the AI industry.

# 2 Current state of AI security regulations and policies and standardization

With the continuous development of AI technology and its industry, many countries and regions have formulated laws, regulations, and policies related to AI to promote the healthy, orderly, safe, and controllable development of AI and to explore and implement such development in AI ethics, AI system security, robotics, autonomous driving, and privacy protection.

## 2.1 Current state of AI security regulations and policies

### 2.1.1 International and overseas situation

**(1) United Nations: Focus on personal safety and ethics, gradually deepening the fields of autonomous driving, robots, and AI policing**

At present, United Nations research on AI security mainly focuses on personal safety and ethics and potential threats and challenges, paying special attention to the impact and challenges of AI on personal safety, the safety of society, and economic development as relevant laws and regulations and research results in the field of robotics and other related research are gradually maturing.

In 2016, the United Nations Economic Commission for Europe revised the "Vienna Road Traffic Convention" (hereinafter referred to as the "Convention") through amendments. In September 2017, UNESCO and the World Commission on the Ethics of Scientific Knowledge and Technology jointly released the Report on Robotics Ethics, pointing out that the manufacture and use of robotics have promoted the advancement of AI and discussing the social and economic benefits and ethical issues of these advancements.

In 2017, with the support of the Netherlands and the municipal government of the Hague, the United Nations established an AI and robotics center in the Netherlands to follow up on the latest developments in AI and robotics. This office is also set up to work with the United Nations Interregional Crime and Justice Research Institute (UNICRI) to deal with the security implications and risks of crime-related AI and robotics.

**(2) United States: Focus on AI design security, adopting standard specifications and validation assessments to reduce the risk of malicious attacks**

In February 2019, the U.S. president signed an executive order to launch the “American Artificial Intelligence Initiative.” This initiative proposes to focus on the five areas of AI research and development, data resource sharing, standard and norm formulation, workforce development, and international cooperation. Among these, the goal of standard and norm formulation aims to ensure that technical standards minimize vulnerabilities that allow for malicious attacks and to promote public trust in AI innovation and technologies.

In response to the American Artificial Intelligence Initiative, the United States updated the National Artificial Intelligence Research and Development Strategic Plan in June 2019. Based on the 2016 version, the updated plan proposes eight strategic priorities, including long-term investment in AI research, addressing ethical and legal social impacts, ensuring the security of AI systems, developing shared public datasets and environments, and passing standard assessment technologies.

**Ethical, legal, and social issues with AI:** The plan proposes improvements to fairness, transparency, and accountability through design in order to establish ethical AI and design a framework for such AI. An AI architecture can be designed to include moral reasoning, such as adopting a two-layer monitoring architecture that separates operational AI from assessment monitors, or opting for security engineering to ensure that developed AI behavior is safe and harmless to humans, or a set of theories, principles, and logical constraints that can be used to formulate an ethical architecture. The Defense Advanced Research Projects Agency (DARPA) is carrying out the Explainable Artificial Intelligence (XAI) project which aims to create a set of machine learning technologies that can generate more explainable AI systems.

**Creating robust and reliable AI systems:** Measures should be taken to improve explicability and transparency, build trust, strengthen validation, prevent attacks, and achieve long-term AI security and value adjustments. For instance, the validation and assessment of AI systems should be strengthened to ensure that they meet official norms and meet users' operating requirements. AI system self-monitoring architectures should be studied for use in verifying consistency with designed target behavior. In February 2019, DARPA announced the launch of the Guaranteeing AI Robustness against Deception (GARD) project, which aims to develop a new generation of defense technology to resist adversarial spoofing attacks against machine learning models.

**Building AI public data resources:** Measures such as developing and providing a variety of datasets, developing training and testing resources, and developing open-source software libraries should be adopted. At the same time, the issue of secure data sharing should be considered, and secure data sharing technology and privacy protection technology should be studied. For instance, the VA Data Commons is creating the world's largest linked medical genome dataset.

**Standards and benchmarks:** It has been proposed that a wide range of AI standards should be developed, technical benchmarks should be established, the availability of AI testing platforms should be expanded, and the community should be allowed to participate in standards and benchmark testing. Specifically, in terms of AI standards, it has been proposed that a system of AI standards be established for software engineering, performance, measurement, personal safety, usability, interoperability, security, privacy, and traceability.

### (3) European Union: Focus on the ethics of AI and the challenges that GDPR poses to AI

In 2017, the European Parliament passed a legislative resolution, proposing to develop a "Charter on Robotics" to promote civil law rules on artificial intelligence and robotics. In April 2018, the European Commission released the "Artificial Intelligence for Europe" communication, which establishes EU AI values by improving technical and industrial capabilities, responding to socio-economic changes, and establishing appropriate ethical and legal frameworks.

In May 2018, the EU's General Data Protection Regulation (GDPR) officially entered into force. As for aspects that specifically involve AI: GDPR requires a certain degree of explicability for AI algorithms, which may prove challenging for “black-box” AI systems. At the same time, Article 22 of the GDPR imposes requirements on automated decisions, including profiling: if the legal effect of automated decision-making involves the data subject, or has a similarly significant effect on the data subject, the data subject should have the right not to be the subject of such a decision. If the automated decision is a decision that must be made in order to fulfill the contract between the data subject and the controller, with the express consent of the data subject, the data controller shall implement appropriate measures to protect the data subject’s rights, freedoms, and legal rights and interests, and at least guarantee that the data subject has the right to intervene in automated decision-making, express his or her own opinion, and reject the decision.

On April 8, 2019, the European Commission High-Level Expert Group on Artificial Intelligence released the Ethics Guidelines for Trustworthy AI, which lists the seven principles of trustworthy AI to ensure that the application of AI is ethical and that the technology is robust and reliable so as to give full play to its greatest advantages and minimize its risks. Specifically, trustworthy AI has two components: first, it should respect basic human rights, regulations, and core principles and values; second, it should be technologically safe and reliable in order to avoid unintentional harm caused by insufficient technology.

### (4) Germany: Actively responding to ethical and moral risks of AI, proposing the first ethical standards for autonomous driving

In March 2017, 24 German companies formed the German AI Association to speak on behalf of the industry, including properly responding to negative impacts such as ethical risks. Germany's Data Ethics Commission is responsible for formulating ethics and codes of conduct for the development of AI. All products and services based on algorithms and AI must pass a review to avoid, in particular, illegal phenomena such as discrimination and fraud.

Germany regards autonomous driving ethics as one of the core areas for regulation in the development of AI. On May 12, 2017, Germany passed the first bill on self-driving cars, amending the Road Traffic Regulations, including relevant laws for the first time on the testing of self-driving cars. The purpose of the bill is to protect the personal safety of drivers, which is an important step for Germany to take towards autonomous driving. In May 2018, the German government introduced the first ethical standard for autonomous driving technology, which will require that autonomous vehicles make priority judgments in accident scenarios and add them to the system's self-learning, such as giving human safety priority over that of animals and other property.

In July 2018, Germany's Federal Cabinet adopted the Key Points for a Federal Government Strategy on Artificial Intelligence document, which aims to promote German AI research and development and applications so that it can reach a world-leading level, promote the use of AI in a responsible way, serve as a benefit to society, and unlock new value-added potential. This document established Germany's goals for the development of AI and measures in priority action areas such as research, transformation, talent training, data use, legal protections, standards, and international cooperation. For instance, Germany intends to adopt measures such as opening up government and scientific research data, engaging in data collaborations with national enterprises, building European data areas, expanding interoperability of data systems in the healthcare industry to make data available and usable, and ensuring the transparency, traceability, and verifiability of artificial intelligence systems.

**(5) United Kingdom: Focus on supervision of robotics and autonomous systems, establishing a data ethics and innovation center to provide government advice**

In October 2016, the Science and Technology Select Committee of the House of Commons issued a report on AI and robotics, which studied the supervision of "robotics and autonomous systems" (RAS).

In April 2018, the British government issued the New Artificial Intelligence Sector Deal, which aims to promote the UK as a global AI leader. This document includes investment plans for domestic and foreign technology companies, the expansion of the Alan Turing Institute, the establishment of the Turing Scholars Program, and the launch of the Centre for Data Ethics and Innovation. Of these initiatives, the Centre for Data Ethics and Innovation is an independent consulting agency established by the British government to provide advice to government agencies and industries to support responsible technological innovation and to help build a strong and trustworthy system of governance. The main work of the Centre in 2019 is analyzing the opportunities and risks brought about by data-driven technologies, including algorithmic bias strategy reviews, AI barometer readings, and research into topics such as AI and insurance, smart speakers, and deepfakes.

In April 2018, the Special Committee on Artificial Intelligence under the British Parliament published the report "AI in the UK: ready, willing, and able?" The report believes that there is no need for unified special supervision of AI at present, and regulators in various industries can make adaptive adjustments to supervision depending on actual circumstances. The report calls on the British government to formulate national-level AI guidelines, set basic ethical principles for the development and use of AI, and explore relevant standards and best practices to achieve industry self-regulation. The report's recommendations on some key issues are:

**In terms of maximizing the value of public data,** the report proposes distinguishing between data and personal data, recommending that data access and sharing be promoted through measures such as data trusts, open public data, and open banking data mechanisms.

**In terms of achieving understandable and reliable AI,** it recommends avoiding "black-box" algorithms in specific major fields, encouraging the development of interpretable AI systems, and requiring the use of more technically transparent AI systems in specific security-critical scenarios.

**In terms of addressing algorithmic bias,** it recommends studying training data and algorithm review and testing mechanisms, noting that more measures need to be taken to ensure that data is truly representative and able to represent diverse groups of people and that it will not further aggravate or solidify social injustices.

In addition, in terms of autonomous driving, the United Kingdom introduced the Vehicle Technology and Aviation Bill in February 2017, stipulating that in the event of road testing of self-driving cars, the insurance process can be simplified to help insurers and insurance companies receive compensation. The United Kingdom will also fully allow self-driving cars to legally drive on roads in 2021.

**(6) Japan: Establishing an AI ethics committee to actively carry out research on AI ethics**

In December 2014, the Japanese Society for Artificial Intelligence established an ethics committee to explore the connection between robotics, AI, and social ethics. In June 2016, the ethics committee proposed draft guidelines for AI researchers to follow. The draft emphasized that "independent of the presence or absence of intentionality, AI has the potential to become harmful." The draft stipulates that no direct or indirect use of AI based on intent to harm should be allowed. When the harm is inadvertently imposed, the loss needs to be repaired, and preventive measures should be taken when malicious use of AI is discovered. Researchers must do their utmost to allow people to make equal use of AI and must have the responsibility to explain the limitations and issues of AI to society.

In January 2015, the Ministry of Economy, Trade and Industry compiled the results of the committee's discussion and issued "Japan's Robot Strategy: Vision, Strategy, Action Plan" (also known as the "New Robot Strategy"). Taking the development and promotion of robots as an important growth point for future economic development, the report provides a detailed five-year action plan that focuses on the main application areas of manufacturing, service, agriculture, forestry and fisheries, medical care, infrastructure construction, and disaster prevention. The report also includes specific actions to be carried out, such as robot technology development, standardization, demonstration assessments, personnel training, and regulatory adjustments.

In March 2017, Japan's Artificial Intelligence Technology Strategy Council released the Artificial Intelligence Technology Strategy report, which elaborated on the roadmap formulated by the Japanese government for the industrialized development of AI. The roadmap includes three stages: the development of data-driven AI technology applications in various fields (transition from stage 1 to stage 2 to be completed in 2020); the development of AI technology utilities in various fields (transition from stage 2 to stage 3 to be completed in 2025-2030); and the connection of various fields into an AI ecosystem.

### 2.1.2 Domestic situation

China has issued a series of AI-related policies and regulations and has issued relevant policy documents around promoting industrial technology development, including the New Generation Artificial Intelligence Development Plan ("the Development Plan"), the Three-Year Action Plan for Promoting Development of a New Generation Artificial Intelligence Industry (2018–2020) ("the Action Plan"), the Three-year Action Plan for "Internet+" Artificial Intelligence ("互联网+"人工智能三年行动实施方案), Guiding Opinions on Promoting the Deep Integration of Artificial Intelligence and the Real Economy, and the Guidelines for National New Generation Artificial Intelligence Innovation and Development Pilot Zone Construction Work. These documents all put forward requirements on AI security and ethics **with a primary focus on AI ethics, security supervision, evaluations and assessments, and monitoring and early warning to strengthen the in-depth application of AI technology within cybersecurity**.

The Development Plan proposes that we "formulate laws and regulations and ethical norms related to promoting AI development," and includes such initiatives as accelerating the formulation of relevant security management regulations, with a focus AI sub-fields with a relatively good foundation, such as autonomous driving and service robots; carrying out research into legal issues of AI application-related civil and criminal liability confirmation, privacy, and intellectual property protection and safe use of information, setting up traceability and accountability systems, and identifying AI legal entities and related rights, obligations, and responsibilities; carrying out AI behavior science and ethics research, setting up multilevel assessment structures for ethics and morals and ethics frameworks for human-machine coordination; formulating moral standards and behavior codes for AI product research designers, enhancing assessments for potential AI risks and benefits, and strengthening emergency solutions for complicated AI scenarios.

The Development Plan also proposes "the setting up of an AI security regulation and assessment system" with such initiatives as enhancing research and assessment of AI influence on national security and confidentiality, improving security prevention and protection systems featuring humans, technology, products, and management, and building warning mechanisms for AI security monitoring; strengthening AI cybersecurity technology research and development and enhancing cybersecurity protections for AI products and systems; and building dynamic AI research, development, and application evaluation and assessment mechanisms.

The Action Plan proposes the establishment of a "cybersecurity assurance system," including the development of security technologies such as vulnerability mining, security testing, threat warnings, attack detection, and emergency responses for key AI products or industry applications such as intelligent connected vehicles and smart homes. Such a system can tackle issues, promote the in-depth application of advanced AI technology in the field of cybersecurity, and accelerate the establishment of shared resources such as vulnerability databases, risk databases, and case sets.

Under the guidance of the national artificial intelligence development strategy, relevant national authorities have issued corresponding regulatory documents in sub-fields such as UAVs, autonomous driving, and finance:

**(1) UAVs:** The Civil Aviation Administration of China (CAAC) has issued the Interim Provisions on Issues Related to the Management of Civilian Unmanned Aircraft, Civilian Unmanned Aircraft Air Traffic Management Measures, Regulations on Operation of Light and Small Unmanned Aircraft (for Trial Implementation), Administrative Measures for Civilian Unmanned Aircraft Operators, Administrative Measures for Commercial Flying Operations of Civilian Unmanned Aircraft (for Trial Implementation), Administrative Measures for Civilian Unmanned Aircraft Operators, and other normative documents to clarify regulations on civilian drone flight activities, civilian drone operators, and other safety-related issues.

**(2) Autonomous driving**: The Ministry of Industry and Information Technology (MIIT), the Ministry of Public Security, and the Ministry of Transport jointly formulated and issued the Administrative Rules of Road Testing of Intelligent Connected Vehicles (for Trial Implementation) ("the Administrative Rules”). The Administrative Rules are applicable to the testing of autonomous driving of connected vehicles on public roads in China. Beijing issued the Beijing Implementation Rules for the Management of Autonomous Vehicle Road Testing (for Trial Implementation) and related documents, identifying 33 open roads, totaling 105 kilometers, for testing. Shanghai issued the Shanghai Administrative Measures for the Road Testing of Intelligent Connected Vehicles (for Trial Implementation), delineating the first stretch of 5.6 kilometers of open test roads and issuing the first batch of testing license plates. Chongqing, Baoding, and Shenzhen have also issued corresponding administrative measures for road testing or have solicited opinions to support intelligent connected vehicles in conducting public road testing.

**(3) Finance:** The People’s Bank of China, the China Banking and Insurance Regulatory Commission, the China Securities Regulatory Commission, and the State Administration of Foreign Exchange jointly issued the Guiding Opinions on Regulating the Asset Management Business of Financial Institutions

According to the document, business entities are responsible for disclosing inherent algorithmic flaws in their intelligent investment advisory services. The People's Bank of China issued the Financial Technology Development Plan, emphasizing that it is necessary to accelerate the formulation and improvement of technology and security regulations for the application of AI, big data, and cloud computing in the financial industry, to study and formulate regulations on the supervision of AI financial applications, and to strengthen security certifications for intelligent financial instruments to ensure that AI financial applications are regulated within a safe and controllable range.

## 2.2 AI security work situation of the main standardization organizations

### 2.2.1 ISO/IEC JTC 1

In October 2017, ISO/IEC JTC 1 held a meeting in Russia and decided to establish a new AI sub-committee SC 42 to be responsible for AI standardization. SC 42 has established the five working groups of Foundational Standards (WG 1), Big Data (WG 2), Trustworthiness (WG 3), Use Cases and Applications (WG 4), and Computational Approaches and Computational Characteristics of AI Systems (WG 5). In addition, SC 42 also has established the AI Management Systems Standard Advisory Group (AG1) and the Intelligent System Engineering Advisory Group (AG3).

The SC 42 WG 3 Trustworthiness Working Group focuses on the reliability and ethics of AI. It has carried out standards research and development on topics such as AI credibility, robustness assessments, algorithm bias, and ethics. Its main standards include:

1） **ISO/IEC TR 24027 Information technology — Artificial Intelligence (AI) — Bias in AI systems and AI-aided decision making**, proposed by the United States' NIST, mainly studies algorithmic bias in AI systems and AI-aided decision-making systems.

2） **ISO/IEC PDTR 24028 Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence** mainly studies the connotation of AI trustworthiness, analyzes typical engineering issues and typical related threats and risks of AI systems, and proposes corresponding solutions. The standard defines trustworthiness as the degree of dependability and reliability of AI and proposes a method for establishing the trustworthiness of AI systems from the perspectives of transparency, verifiability, explicability, and controllability.

3） **ISO/IEC TR 24029-1 Artificial Intelligence (AI) — Assessment of the robustness of neural networks — Part 1: Overview**, proposed by France, mainly proposes cross-validation, formal validation, posterior validation, and other methods to evaluate the robustness of neural networks for projects that research AI robustness. **Artificial Intelligence (AI) -- Assessment of the robustness of neural networks -- Part 2: Formal methods methodology** was also applied for as a research item at the meeting in Japan this past October.

4） **ISO/IEC 23894 Information Technology — Artificial Intelligence — Risk Management** reviews the risks of AI and provides processes and methods of AI risk management.

5） **TR Information Technology — Artificial Intelligence — Overview of Ethics and Social Concern** focuses on AI research from the perspectives of ethics and social concern.

In addition to SC 42, the ISO/IEC JTC 1/SC 27 Information security, cybersecurity, and privacy protection subcommittee, in its own WG 5 Identity Management and Privacy Technologies Working Group, established the research project **"AI impact on privacy"** to study the impact of AI on privacy. The ISO/IEC JTC 1/SC 7 Software and Systems Engineering Subcommittee is also developing ISO/IEC/IEEE 29119-11 **Software and Systems Engineering — Software Testing — Testing of AI-Based Systems**, which aims to standardize artificial intelligence system testing.

### 2.2.2 ITU-T

In 2017 and 2018, ITU-T organized the AI for Good Global Summit that focused on strategies to ensure the credible, safe, and inclusive development of AI technology and the right to fair profits. ITU-T is mainly committed to addressing security issues in AI applications such as smart healthcare, intelligent vehicles, spam content management, and biometric feature recognition. In ITU-T, the SG 17 Security Study Group and SG 16 Multimedia Study Group launched the development of security-related standards for AI. Specifically, ITU-T SG 17 has planned to carry out research, discussion, and related standardization projects for AI use in security and the security of AI. The Q9 "Remote Biometrics Study Group" and Q10 "Identity Management Architecture and Mechanisms Study Group" under the ITU-T SG 1 Security Standards Working Group are responsible for the standardization of ITU-T biometric feature recognition. Of the two groups, Q9 focuses on the various challenges of privacy protection, reliability, and security of biometric feature data.

### 2.2.3 IEEE

IEEE has carried out a number of AI ethics studies and has released a number of AI ethics standards and research reports. As early as the end of 2017, IEEE released the report Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Second Edition, which collected the insights and suggestions of over 250 experts from around the world engaged in AI, law, ethics, philosophy, and policy on issues in AI and autonomous systems. The report has been updated to a second edition.

IEEE is developing the IEEE P7000 series of standards to regulate the ethics of artificial intelligence systems. The main standards are introduced as follows:

1） **IEEE P7000 Model Process for Addressing Ethical Concerns During System Design**: The standard establishes a model process where engineers and technicians can deal with ethical issues at all stages of system startup, analysis, and design. The expected process requirements include management and engineering views of new IT product development, computer ethics and IT system design, value-sensitive design, and stakeholder participation in ethical IT system design.

2） **IEEE P7001 Transparency of Autonomous Systems**: Aiming at the transparency of the operation of autonomous systems, this standard provides guidance for the self-assessment of transparency in the development of autonomous systems, helps users understand the reasons for certain decisions made by the system, and proposes mechanisms to improve transparency (such as the need for secure storage of sensors and internal status data).

3） **IEEE P7002 Data Privacy Process**: The standard proposes means of managing the ethical issues of systems and software that acquire personal information, regulates the practice of managing privacy issues during the system/software engineering life cycle, and can also be used to conduct compliance assessments of privacy practices (privacy impact assessments).

4） **IEEE P7003 Algorithmic Bias Considerations**: This standard provides steps to eliminate the issue of negative deviations when creating algorithms and will also include benchmarking procedures and specifications for selecting validation datasets, which is suitable for developers of autonomous or intelligent systems to avoid negative deviations in their code. When using subjective or incorrect data interpretations (such as erroneous causality), negative deviations may occur.

5） **IEEE P7004 Standard for Child and Student Data Governance**: This standard defines how to access, acquire, share, and delete data related to children and students in any educational or institutional environment and provides transparency and accountability processes and certification for educational institutions or organizations that process child and student data.

6） **IEEE P7005 Standard for Transparent Employer Data Governance**: This standard provides guidelines and certifications for storing, protecting, and using employee data in an ethical manner in the hopes of providing clarity and advice for employers to share their information in a secure and reliable environment and how employers can collaborate with employees.

7） **IEEE P7006 Standard for Personal Data Artificial Intelligence (AI) Agent**: This standard involves the issue of automatic decision-making by machines and describes the technical elements required to create and authorize access to personalized AI, including input, learning, ethics, rules, and values controlled by individuals. By allowing individuals to create personal “terms and conditions” for their data, agents will provide people with a way to manage and control their identity in the digital world.

8） **IEEE P7007 Ontological Standard for Ethically Driven Robotics and Automation Systems**: This establishes a set of ontologies with different levels of abstraction, including concepts, definitions, and interrelationships. These definitions and relationships will enable robotics and automation systems to be developed based on worldwide ethics and moral theories.

9） **IEEE P7008 Standard for Ethically Driven Nudging for Robotic, Intelligent and Autonomous Systems**: "Nudges" displayed by robotic, intelligent, and autonomous systems are defined as overt or hidden suggestions or manipulations that are intended to influence the behavior or emotions of a user. This standard establishes a definition of typical nudges (currently in use or that could be created). It contains concepts, functions, and benefits necessary to establish and ensure ethically driven methodologies for the design of the robotic, intelligent, and autonomous systems that incorporate them.

10）**IEEE P7009 Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems**: Autonomous and semi-autonomous systems which remain operational after an intended or unintended malfunction can disadvantage and harm users, society, and the environment. This standard establishes a practical, technical baseline of specific methodologies and tools for the development, implementation, and use of effective fail-safe mechanisms in autonomous and semi-autonomous systems.

11）**IEEE P7010 Wellbeing Metrics Standard for Ethical Artificial Intelligence and Autonomous Systems**: This standard establishes wellbeing metrics relating to human factors directly affected by intelligent and autonomous systems and establishes a baseline for the types of objective and subjective data these systems should analyze and include to increase human wellbeing.

12）**IEEE P7011 Standard for the Process of Identifying and Rating the Trustworthiness of News Sources**: The purpose of this standard is to respond to the negative effects of the uncontrolled spread of fake news by providing an easy-to-understand open rating system for rating online parts of online news providers and multimedia news providers.

13）**IEEE P7012 Standard for Machine Readable Personal Privacy Terms**: This standard provides a means of providing personal privacy terms and a way for machines to read and agree to such terms.

14）**IEEE P7013 Inclusion and Application Standards for Automated Facial Analysis Technology**: This study shows that AI used for automatic facial analysis is susceptible to prejudice. This standard provides phenotypic and demographic definitions that technicians and auditors can use to evaluate the diversity of facial data used for training and benchmark algorithm performance and to establish accuracy reports and data diversity rules for automated facial analysis.

### 2.2.4 NIST

In August 2019, the National Institute of Standards and Technology (NIST) issued guidance on how the government can develop AI technology standards and related tools. This plan outlines a number of initiatives to promote the responsible use of AI and sets out several guiding principles that will provide guidance for future technical standards.

The plan emphasizes the need to develop technical tools to help institutions better research and evaluate the quality of AI systems. These tools include standardized testing mechanisms and powerful performance indicators that can allow governments to better understand various systems and determine how to set effective standards. NIST recommends focusing on the study and understanding of the credibility of AI and incorporating these indicators into future standards and also recommends that AI standards cited in regulation or procurement remain flexible to adapt to the rapid development of AI technology, and that metrics be developed to evaluate the trustworthy attributes of AI systems. AI risk management such as notification risk, monitoring, and mitigation and trust requirements, and methods for the design, development, and use of artificial intelligence should be studied to promote creative problem solving through AI challenge issues and test platforms.

### 2.2.5 TC260

AI security standards are standards and norms related to AI security, ethics, and privacy protection. Broadly speaking, AI security standards involve security standards related to AI algorithm models, data, infrastructure, products, and applications. At present, China's National Information Security Standardization Technical Committee (TC260) **AI security-related standards are mainly focused on biometric feature recognition, smart homes, and other AI application security fields, as well as supporting areas related to data security and personal information protection. There are currently no officially established security standards for AI itself or basic commonality.**

**(1) Common standards for fundamental AI security**

In 2018, TC260 established the standard research project **"Artificial Intelligence Security Standards Research,"** which mainly researched AI security risks, the status of AI security policies and standards, and AI security standard requirements and security standard systems. Much of the material in this white paper comes from such topics.

Under the standard research item **"Information Security Technology and Artificial Intelligence Applications"** established in 2019, the study of the security attributes and principles of AI, security risks, security management, and security engineering practice guidelines is established for the stages of requirement, design, development training, validation and assessment, and operation. Such a study is applicable to organizations such as AI developers, operations managers, users, and third parties as a reference when ensuring the security of AI system engineering.

**(2) Biometric feature recognition security standards**

TC260 has released GB/T 20979-2007 Information Security Technology - Technical Requirements for Iris Recognition System and is developing such standards as Information Security Technology – Biometric Feature Recognition Authentication Protocol Based on Trusted Environment, Information Security Technology - Technical Requirements for Fingerprint Recognition System, Information Security Technology - Security Technology Requirements for Network Facial Recognition Authentication System, Information Security Technology – Biometric Feature Recognition Information Protection Requirements.

**GB/T 20979-2019 Information Security Technology - Technical Requirements for Iris Recognition System** specifies the technical requirements of iris recognition systems that use iris recognition technology to provide support for identity authentication. This standard is applicable to the design and implementation of iris recognition systems based on information security level protection requirements and can serve as a reference in the testing and management of iris recognition systems.

**GB/T 36651-2018 Information Security Technology - Biometric Feature Recognition Authentication Protocol Based on Trusted Environment** specifies a biometric feature recognition authentication protocol based on a trusted environment, including protocol framework, protocol process, protocol requirements, and protocol interfaces. This standard applies to the development, testing, and evaluation of biometric feature recognition service protocols.

**GB/T 37076-2018 Information Security Technology - Technical Requirements for Fingerprint Recognition System**: Security threats and security objectives of fingerprint recognition systems are analyzed, and the technical requirements for the security of fingerprint recognition systems are proposed, regulating the application of fingerprint recognition technology in the field of information security.

**Information Security Technology - Security Technology Requirements for Network Facial Recognition Authentication System** specifies the basic composition, functional requirements, performance requirements, and test methods for facial recognition systems for security surveillance video. This standard applies to the program design, project acceptance, and related product development of video surveillance facial recognition systems for security purposes. Video surveillance facial recognition systems in other fields can be used for reference.

**Information Security Technology – Biometric Feature Information Protection Requirements** studies and formulates security protection requirements for biometric feature identification, including threats and countermeasures for biometric feature recognition systems, security requirements for the secure binding between biometric information and identity subjects, application models, and privacy protection requirements.

### (3) Autonomous driving security standards

TC260 established the Information Security Technology - Cybersecurity Guidebook for Vehicle Electronic Systems standard item in 2017 and established the standard development item Information Security Technology - Vehicle Network Equipment Requirements and the research item Information Security Technology - Vehicle Electronic Chip Security Technology Requirements in 2019. Strictly speaking, these standards belong to the category of vehicle electronics but not to the category of autonomous driving.

### (4) Smart home security standards

The TC260 project standard Information Security Technology - General Technology Requirements for Smart Home Security was launched in 2018 and the project standard Information Security Technology - Technical Requirements and Test Evaluation Methods for Smart Door Locks was launched in 2019.

**Information Security Technology - General Technology Requirements for Smart Home Security** specifies general technology requirements for smart home security, including the overall framework, security models, device security requirements, gateway security requirements, cybersecurity requirements, and application service platform security requirements for smart homes. This standard applies to the safe design and implementation of smart home products and can be used for reference in the security testing and management of smart homes.

**Information Security Technology - Technical Requirements and Test Evaluation Methods for Smart Door Locks** provides security technology requirements and test evaluation methods for smart door locks, where smart door locks refer to door locks that control the door lock actuator to open and close using the identification of fingerprints, finger veins, irises, human faces, and other human biometric features as well as smart cards, wireless remote control codes, static passwords, temporary passwords, and other information.

### (5) Data security and personal information protection standards

Since its establishment in 2016, TC260's Big Data Security Standards Working Group has issued six national standards, 10 research standards, and 18 research items in the direction of data security and personal information protection.

In terms of personal information protection, standards have mainly focused on personal information protection requirements, de-identification technology, app acquisition of personal information, privacy engineering, impact assessment, informed consent, and cloud services. The two standards GB/T 35273 Information Security Technology - Personal Information Security Standards (信息安全技术个人信息安全规范) and GB/T 37964 Information Security Technology - Guide for De-identifying Personal Information (信息安全技术个人信息去标识化指南) have already been published, five standards are being studied, and two standard study items have been initiated.

In terms of data security, standards have mainly focused on data security capabilities, data transaction services, outbound assessment, government data sharing, health and medical data security, and telecommunication data security. The four standards GB/T35274 Information Security Technology - Big Data Service Security Capabilities Security Requirements, GB/T 37932 Information Security Technology - Data Transaction Service Security Standards, GB/T 37973 Information Security Technology - Big Data Security Management Guide, and GB/T 37988 Information Security Technology - Data Security Capability Maturity Model, have already been published, five standards are being studied, and 16 standard study items have been initiated.

In addition, the Standardization Administration of China formally established the National Artificial Intelligence Standardization General Working Group in January 2018, which is responsible for the overall coordination and planning of artificial intelligence standardization, and is responsible for the international and domestic standardization of artificial intelligence. At present, the National Artificial Intelligence Standardization General Working Group has published such achievements as the Artificial Intelligence Standardization White Paper 2018[e] and the Artificial Intelligence Ethical Risk Analysis Report (人工智能伦理风险分析报告) and is researching AI terminology, AI ethical risk analysis assessments, and other standards.

### 2.2.6 Other standardization organizations

The **China Communications Standards Association** has carried out standard research work on vehicle electronics, smart homes, and other aspects and has currently issued such standards as the YDB 201-2018 Technical Requirements for the Security Capabilities of Smart Home Terminal Equipment and the T/CSHIA 001-2018 Technical Requirements for the Security of Smart Home Network Systems. However, relevant research work is still mainly focused on AI in specific application scenarios. Its research standards include Guidelines for the Evaluation of Artificial Intelligence Products, Applications, and Services Security; Artificial Intelligence Service Platform Security; Research on Artificial Intelligence Terminal Product Standards Systems; Personal Information Protection Technical Requirements and Assessment Methods for Mobile Intelligent Terminal Artificial Intelligence Capabilities and Applications; and Security Technical Requirements and Test Assessment Methods for Mobile Intelligent Terminal Facial Recognition.

**The China Artificial Intelligence Open Source Software Development League** is a community organization engaged in work related to artificial intelligence open-source software. The league has developed evaluation standards for products or services such as machine translation and intelligent assistance as well as reliability evaluation standards for deep learning algorithms, primarily including T/CESA 1039-2019 Information Technology - Artificial Intelligence - Machine Translation Capability Level Assessment, T/CESA 1038-2019 Information Technology - Artificial Intelligence - Intelligent Assistant Capability Level Assessment, and T/CESA 1026-2018 Artificial Intelligence - Deep Learning Algorithm Assessment Specifications. T/CESA 1026-2018 Artificial Intelligence - Deep Learning Algorithm Evaluation Specifications puts forward an evaluation index system and evaluation process for AI deep learning algorithms, as well as assessments for the requirement stage, design stage, implementation

[e] Translator's note: For an English translation of this white paper, see: https://cset.georgetown.edu/wp-content/uploads/t0120_AI_standardization_white_paper_EN.pdf

stage, and operations stage to guide deep learning algorithm developers, users, and third parties in assessing the reliability of deep learning algorithms.

## 2.3 AI ethical and moral work situation

Given the complexity of AI, it inevitably determines such fields as technology, ethics, law, and morals. To ensure the healthy development of AI, a corresponding ethical and moral framework must be established. In terms of the ethics of AI, there are rich research results at home and abroad. Among them, Asilomar's AI Principles and the AI ethics standards initiated by the IEEE organization have become the world's most influential research achievements of AI ethics. In addition to the widely reached consensus, many countries and institutions have also issued their own relevant guidelines[15].

**(1) Asilomar's AI principles**

Asilomar's AI Principals were proposed at the Asilomar Conference on Beneficial AI in January 2017, and the ethics and value principles that it advocates for include security, failure transparency, judicial transparency, responsibility, value alignment, human values, personal privacy, liberty and privacy, shared benefit, shared prosperity, human control, non-subversion, and prohibition of an AI arms race.

**(2) IEEE**

In March 2017, IEEE published the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems in *IEEE Transactions on Robotics and Automation* magazine, proposing the establishment of AI ethical design principles and standards to help people avoid both fear and blind worship of AI, thereby promoting AI innovation. The following five principles were proposed: 1) human rights: ensure that they do not violate internationally recognized human rights; 2) well-being: give priority to indicators of human well-being in their design and use; 3) accountability: ensure that their designers and operators are responsible and accountable; 4) transparency: ensure that they operate in a transparent manner; 5) use with caution: minimize the risk of abuse.

**(3) United States**

The United States Public Policy Council issued the Statement on Algorithmic Transparency and Accountability on January 12, 2017, which proposed the following seven guidelines: 1) awareness; 2) access and redress; 3) accountability; 4) explanation; 5) data provenance; 6) auditability; 7) validation and testing.

**(4) European Union**

On April 8, 2019, the European Commission released the Ethics Guidelines for Trustworthy AI compiled by the High-Level Expert Group on Artificial Intelligence, listing the seven principles of AI trustworthiness, including: 1) human agency and oversight; 2) safety; 3) privacy and data governance; 4) transparency; 5) diversity, non-discrimination, and fairness; 6) social and environmental well-being; 7) accountability.

**(4) Japan**

The Japanese Society of Artificial Intelligence (JSAI) has issued the Japanese Society of Artificial Intelligence Ethical Guidelines, requiring members of JSAI to follow and practice the following guidelines: 1) contribution to humanity; 2) obedience to laws and regulations; 3) respect for the privacy of others; 4) fairness; 5) security; 6) acting with integrity; 7) accountability and social responsibility; 8) communication with society and self-development; 9) AI obedience to ethical guidelines.

**(5) United Kingdom**

In April 2018, the Special Committee on Artificial Intelligence under the British Parliament published the report "AI in the UK: ready, willing, and able?" The guidelines that it put forward include five aspects: 1) Artificial intelligence should be developed for the common good and benefit of humanity; 2) artificial intelligence should operate on principles of intelligibility and fairness; 3) artificial intelligence should not be used to diminish the data rights or privacy of individuals, families, or communities; 4) all citizens have the right to be educated to enable them to flourish mentally, emotionally, and economically alongside artificial intelligence; 5) the autonomous power to hurt, destroy, or deceive human beings should never be vested in artificial intelligence.

**(6) Canada**

The Montreal Declaration on Responsible AI published in Canada proposed seven values and emphasized that they are all moral principles that should be observed in the development of artificial intelligence: well-being, autonomy, justice, privacy, knowledge, democracy, and accountability.

**(7) China**

On February 25, 2019, the Ministry of Science and Technology announced the establishment of the National New Generation Artificial Intelligence Governance Specialist Committee (国家新一代人工智能治理专业委员会) to further strengthen research on AI-related laws, ethics, standards, and social issues and to participate in international exchanges and cooperation on AI-related governance. On June 19, 2019, the committee published the Governance Principles for a New Generation of Artificial Intelligence: Develop Responsible Artificial Intelligence,[f] proposing that the development of artificial intelligence should follow eight principles: 1) harmony and friendliness; 2) fairness and justice; 3) inclusivity and sharing; 4) respect for privacy; 5) secure and controllable; 6) shared responsibility; 7) open collaboration; 8) agile governance.

In April 2019, the National Artificial Intelligence Standardization General Working Group published the Artificial Intelligence Ethical Risk Analysis Report. The report proposes two ethical guidelines for AI. The first is the principle of human fundamental interests (人类根本利益), which means that AI should adopt the ultimate goal of achieving the fundamental interests of humans; the second is the principle of responsibility, which refers to the establishment of a clear responsibility system in both the development and application of AI-related technologies. Under the principle of responsibility, the principle of transparency should be adhered to in the development of AI technology, and the principle of parity of authority and responsibility should be adhered to in the application of AI technology.

In addition, given that robots are a representative AI product, a number of ethical principles, norms, guidelines, and standards have been launched in Japan, South Korea, the United Kingdom, Europe, and UNESCO in terms of robotics principles and ethical standards.

[f] Translator's note: An English translation of these governance principles is available online at: https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-chinese-expert-group-offers-governance-principles-responsible-ai/

# 3 Analysis and connotations of AI security risks

The application of AI is changing the development trajectory of the human economy and society and is bringing tremendous changes to people's production and lives. However, AI also poses risks and challenges for all of society that cannot be ignored. AI security risk refers to the possibility that security threats use the vulnerability of AI assets to trigger AI security incidents or affect related parties. To make good use of AI technology, it is necessary to fully and clearly understand new attack threats, AI security risks, and the impact on relevant parties.

## 3.1 New attack threats

As an information system that uses AI technology, AI systems, in addition to being threatened by traditional cyberattacks such as denial-of-service (DoS), will also face attacks that specifically target AI systems. These attacks particularly affect systems that use machine learning[13].

**(1) Attack methods**

**The first is an adversarial example attack,** which refer to adding subtle, typically unrecognizable interference to inputs, causing the model to give a wrong output with high confidence. Studies have shown that deep learning systems are susceptible to well-designed adversarial examples, which may lead to erroneous or missed judgments. Adversarial example attacks can also come from the physical world, attacking autonomous driving through carefully constructed traffic signs. For example, research by Eykholt et al.[12] shows that a slightly modified physical parking sign can cause a real-time target detection system to mistake it as a speed limit sign, which may cause traffic accidents.

Attackers use elaborately constructed adversarial examples and can also initiate spoofing attacks such as imitation attacks (模仿攻击) and evasion attacks (逃避攻击)[14]. Imitation attacks achieve the purpose of obtaining the victim's authority by imitating the victim's example. Currently, they mainly appear in image recognition systems and speech recognition systems based on machine learning. Evasion attacks are early forms of attacks against machine learning, such as spam detection systems and malicious program detection systems in PDF files. By generating several adversarial examples that can successfully evade detection by the security system, a malicious attack on the system is achieved.

**The second is data poisoning,** which mainly adds carefully constructed abnormal data to the training data, destroying the probability distribution of the original training data and causing the model to produce classification or clustering errors under certain conditions. Because data poisoning attacks require attackers to access training data, this type of attack is typically more effective for online learning scenarios (that is, the model uses online data to continuously learn and update the model) or systems that need to be periodically retrained to update the model. Typical scenarios include recommendation systems, adaptive biometric systems, and spam detection systems. Correctly filtering training data can help detect and filter abnormal data, thereby minimizing possible data poisoning attacks.

**The third is model stealing**, which either sends a large number of prediction queries to the target model and uses the received responses to train another model with the same or similar function or uses reverse attack technology to obtain the model parameters and training data. For models deployed in the cloud, attackers typically use certain application programming interfaces (APIs) provided by the machine learning system to obtain the preliminary information on the system model and then use the preliminary information to reverse-analyze the model to obtain the training data inside the model and data acquired at runtime[4]. For models deployed privately on users' mobile devices or servers in data centers, attackers can use traditional security techniques such as reverse engineering to restore the model files directly.

**The fourth is AI system attacks.** Typical attacks on machine learning systems are attacks that affect the confidentiality of data and the integrity of data and calculations, as well as other forms of attacks that result in denial of service (DoS), information disclosure, or invalid calculations. For example, a control-flow hijacking attack (控制流攻击) on a machine learning system may disrupt or circumvent machine learning model inference or lead to invalid training. The complex device models (such as hardware accelerators) used by machine learning systems are mostly paravirtualized or simulated, and may be subject to such attacks as device spoofing (设备欺骗), memory remapping attacks at runtime (运行时内存重新映射攻击), and man-in-the-middle device attacks.

**(2) Attack impact**

Deep learning is susceptible to data poisoning, evasion attacks, imitation attacks, model stealing, and adversarial example attacks.

**First, the model may be attacked during training, testing, and inference.** Data poisoning attacks are mainly targeted at the training process, while adversarial example attacks can be targeted at the training, testing, and inference processes. Model stealing attacks are targeted at the inference process. Evasion attacks and imitation attacks are targeted at the testing and inference process.

**Second, attacks compromise the confidentiality, integrity, and availability of data and models.** The above attacks typically lead to consequences such as model decision errors and data leakage. Data poisoning attacks destroy training datasets, mainly affecting data integrity and model availability. Model stealing attacks mainly affect data confidentiality, model confidentiality, and privacy. Adversarial examples, evasion attacks, and imitation attacks do not destroy training datasets but rather mainly affect model integrity and usability.

## 3.2 Hidden dangers of AI

Generally, AI assets are mainly composed of data, algorithm models, infrastructure, product services, and industry applications, as shown in Figure 3-1. A trained model can be understood as a function of y=f(x). The model for deep learning training typically includes two parts. One is the description of the network structure or the definition of the network structure, and the other is the specific parameter value of each layer of the network[4].

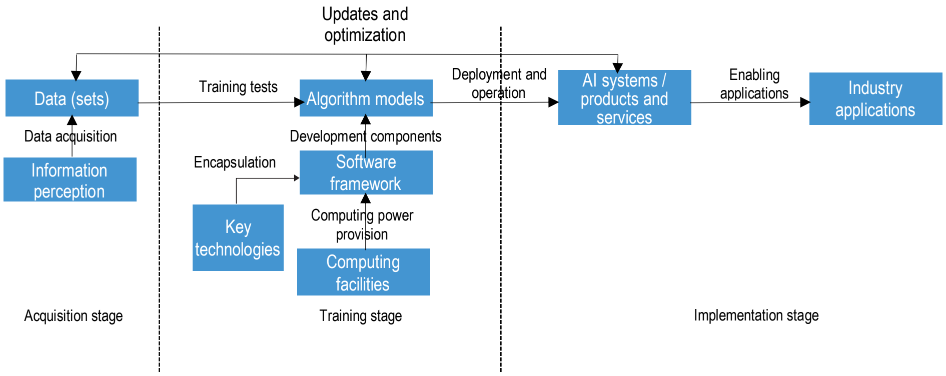


Figure 3-1 AI assets

### 3.2.1 Hidden dangers of algorithm models

An **algorithm model** is the core of an AI system, and security risks in the algorithm model may bring fatal security consequences to the AI system.

**(1) Algorithm models have hidden defects in such aspects as robust balance and data dependence**

**First, it is difficult to balance accuracy with model robustness.** AI algorithm models generally rely on the construction of probability and statistical models, and there is a trade-off between accuracy and robustness. Research by Eykholt, et al.[12] showed that for robustness against adversarial example attacks, models with higher accuracy generally have worse robustness, and the logarithm of classification error rate has a linear relationship with the model robustness.

**Second, datasets have a great impact on model accuracy.** At present, AI is still in a stage of knowledge learning driven by massive data, and the quantity and quality of datasets are among the key factors that determine the quality of a model. The application of a model may result in unexpected situations, and training data may have difficulty covering such situations, such that results are inconsistent with expectations or even harmful. Normal environmental changes may also produce dataset noise, which threatens the reliability of a model. For example, affine transformation, light intensity, angle, and contrast changes can have unpredictable effects on visual model predictions.

**Third, reliability challenges must be faced.** Application scenarios with high real-time performance (such as autonomous driving) require that the algorithm model is available at all times. If the data is subject to directional interference before entering the core AI module, it will lead to immediate erroneous judgment.

**(2) Algorithms may have hidden prejudices or biases, resulting in biased results or improper handling.**

**Bias and discrimination** refer to when the designer or developer of an algorithm has a certain subjective bias in his or her perception of things or inadvertently uses a training dataset with deviations, resulting in a decrease in model accuracy or classification errors or even discriminatory results when the model is used. Training set deviations are typically due to incorrect application or failure to consider statistical methods and rules. For example, when data is selected subjectively rather than objectively or when non-random data is selected, a selection bias will occur. When assumptions can be explained by relevant information, a confirmation of deviation will occur. If a bias already exists in an algorithm, the bias may be further strengthened in the algorithm after deep learning. If the algorithm is applied to crime evaluation, credit lending, employment assessments, or other occasions where personal interests are concerned, the resulting discrimination may seriously harm personal rights and interests.

**(3) There are issues with the explicability and transparency of results in AI algorithm decision-making characterized by the "black box" feature**

Deep learning has achieved unprecedented progress in many complex tasks, but deep learning systems typically require millions or even billions of parameters. As such, it is difficult for developers to annotate complex neural networks in an interpretable manner, thus leading to a veritable "black box."

**First, AI algorithms based on neural networks have "emergence" (涌现性) and "autonomy," which can easily result in algorithm black boxes**[16]. Emergence, i.e. intelligence, is a complex behavior generated by simple rules underlying an algorithm. Algorithmic behavior is not a single behavior with clear boundaries but rather an evolution of collective behavior. In other words, the effect of a behavior is neither determined by a "certain" behavior nor completely determined by its antecedent. Autonomy: the autonomy of today's weak AI mainly refers to the self-organization of intelligent behavior. Deep learning algorithms can learn and evolve from vast quantities of unlabeled big data without the intervention of programmers.

**Second, the application of AI in important industries faces explicability challenges.** When AI is applied to key industries such as finance, healthcare, and transportation, which are critical to personal and property safety, humans' sense of security, trust, and approval of an algorithm may depend on the transparency and intelligibility of the algorithm. In addition to the technical opacity of the AI model itself, there may also be opacity in data and data processing activities.

### 3.2.2 Hidden dangers for data security and privacy protection

Data is a basic resource of AI, and machine learning requires large amounts of diverse and high-quality data for training. In terms of different engineering stages, vast quantities of raw data, training and test datasets, and actual application system input data (field data) are acquired. AI system data life cycles are as shown in Figure 3-2.

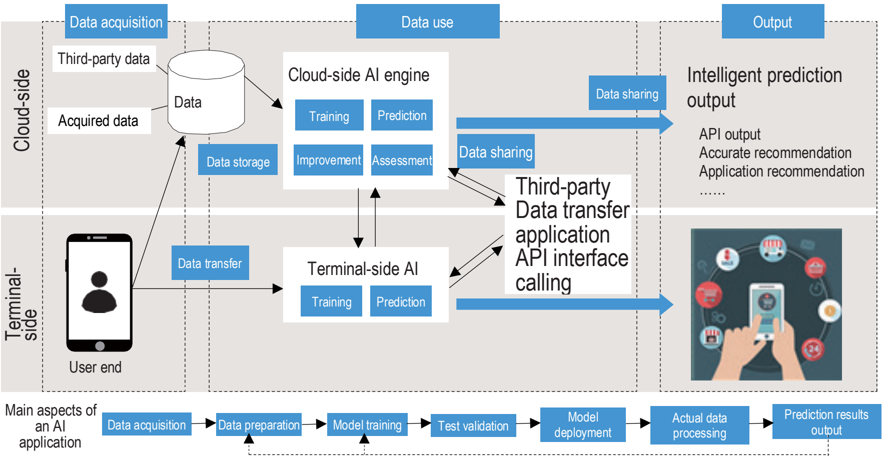


Figure 3-2 AI system data life cycle

### (1) Hidden dangers of data acquisition

In the data acquisition stage, the AI system acquires a large amount of training data and application data through user provision, automatic acquisition, and indirect acquisition methods for training models or inference predictions. Common data acquisition stage issues mainly include:

**First is excessive data acquisition.** Because model training requires the acquisition of a large amount of diverse data, it is difficult to ensure that data acquisition follows the necessary principles and to clarify the range of data acquisition for the purpose of use. In AI application scenarios, to optimize products and identify specific targets, the acquisition terminal may acquire a large amount of environmental data or user behavior logs. For instance, in order to avoid hitting pedestrians, autonomous vehicles will collect enough environment information to judge whether there are pedestrians.

**Second is data acquisition that is inconsistent with user authorization.** In addition to acquiring data directly from users, there are often situations where training data is acquired from such channels as online data sources and business procurement. Certain challenges must be faced to ensure user authorization in such scenarios. For instance, public data sources are typically limited to scientific research. If such data is used for commercial purposes, acquisition thereof may face the risk of inconsistency between data usage and user authorization.

**Third is compliance issues with the acquisition of personal sensitive information.** Currently well-developed applications such as computer vision and speech recognition typically must acquire sensitive personal biometrics such as faces and voiceprints, and the acquisition of personal sensitive information such as biometrics poses legal compliance risks. For instance, in August 2019, Swedish data regulators issued a 200,000 krona [$20,000] GDPR fine to a local high school on the grounds that the school used a facial recognition system to record student attendance.

**Fourth is data quality issues.** The accuracy of AI models is limited by the quality of training data and application data, and insufficiencies in the size, diversity, and balance of training datasets, low quality of dataset labeling, data poisoning attacks, and data noise will all affect the quality of training data[8].

**Fifth is the difficulty in guaranteeing a user's right to opt out.** Since AI systems are typically deployed from trained models, it is difficult for users to opt out of the data used in model training.

### (2) Hidden dangers of data use

The data use stage, that is, data analysis and processing, involves the model training process and operation deployment process. These processes include data preparation, model training, test validation, model deployment, actual data processing, and predicted result output. Of these processes, data preparation primarily comprises preprocessing and labeling operations on acquired original data to generate datasets for training.

**First is the problem of re-identification of anonymous data.** In the data pre-processing stage, data pre-processing techniques are typically used to improve data quality. This process merges externally acquired data with the user's own data, which may cause the data that has been anonymized to be re-identified.

**Second is the hidden danger of data labeling and compliance issues**. Limited by the costs of data labeling, most companies resort to outsourcing data labeling together with independent labeling. Because data labeling personnel can directly access original data and datasets, if data security management is not standardized, there may be risks of internal personnel stealing data, accessing data without authorization, contaminating training datasets, or leaking data. In addition, at present, most data labeling relies on manual labeling. When manually processing personal sensitive information, you may encounter privacy compliance risks that exceed user authorization. For instance, intelligent voice assistants are exposed to employee monitoring and analysis events.

**Third is privacy compliance issues with automated decision-making.** AI systems trained with personal information often need to consider compliance issues for automated decision-making. The EU GDPR and Article 29 Working Party documents have relevant requirements for automated decision-making and require that automatic decision-making does not discriminate against vulnerable groups. Meanwhile, AI systems must properly explain the significance of data processing and possible results. However, the black-box feature of machine learning algorithms makes the data processing process less interpretable.

**(3) Hidden data dangers at other stages**

**Hidden dangers of data storage:** The data of the AI system is usually stored in a storage system such as a database or data warehouse in the cloud or stored in file form on the terminal-side device. The hidden dangers of data storage are mainly reflected in the security of the storage media of data and models. If there is a security breach in the storage system or the model storage file is damaged, it may cause data leakage. For instance, in February 2019, SenseNets (深网视界) was exposed to a leak of facial recognition information of more than 2.5 million people, mainly due to a failure to password-protect their internal MongoDB database.

**Hidden dangers of data sharing:** In the process of data acquisition and data labeling, many AI companies will rely on third-party companies or use crowdsourcing to achieve the acquisition and labeling of vast quantities of data. The data protection capabilities of multiple parties involved in the data link are uneven, which may bring about potential risks of data leakage and misuse. In the operation stage of AI systems, there are many situations in which data is shared or disclosed to third parties. For instance, machine learning algorithms or model training are completed by third parties, and data interaction with third-party artificial intelligence APIs is required.

**Hidden dangers of data transfers:** AI systems are usually deployed on the cloud side and terminal side. Large quantities of data are transmitted between the cloud side and the terminal side, and there are hidden dangers in traditional data transmission.

### 3.2.3 Hidden dangers for infrastructure

Infrastructure is software and hardware that AI products and applications generally rely on, such as software frameworks, computing facilities, and smart sensors. Among them, the software framework is the engineering implementation of general algorithm model components, which provide integrated software packages and algorithm call interfaces for AI application development. Infrastructure provides computing power resources for AI, typically derived from the edge computing power of smart chips, smart servers, or terminal-side devices. Infrastructure security risks are the fundamental risks faced by AI, mainly including:

**First, open source security risks**: Today's rapid development of AI technology and the industry is largely due to the open sourcing of mainstream AI software, frameworks, dependent libraries, and other necessary experiments and production tools, and more and more entrepreneurs can rely on open source achievements for AI research. The open source community has played a key role in the development of AI in terms of function optimization and framework design, but it often overlooks the security risks of its achievements.

**Second, software framework security risks:** In recent years, domestic cybersecurity companies have repeatedly discovered security vulnerabilities in machine learning-related software frameworks, tools, and dependent libraries such as TensorFlow and Caffe. These vulnerabilities may be used in cyberattacks, bringing new threats and challenges to AI applications.

**Third, traditional software and hardware security risks:** AI infrastructure is composed of software and hardware, and as such, it also faces the security risks of traditional software and hardware. For this reason, attention must be paid to AI service interface security, software and hardware security, and service availability.

**Fourth, system complexity and uncertainty risks:** In many cases, AI systems are designed to operate in complex environments, with a large number of potential states that cannot be examined or tested in detail, and the system may face situations that were never considered in the design process.

**Fifth, system behavior unpredictability:** For AI systems that learn after deployment, the behavior of the system may be mainly determined by learning under unsupervised conditions. In this case, it may be difficult to predict the behavior of the system.

**Sixth, human-computer interaction security risks:** In many cases, the performance of AI systems is greatly affected by human interaction, and different groups of people may have different reactions such that changes in human responses may affect the safety of the system.

### 3.2.4 Hidden dangers for application security

Product application refers to hardware, software, systems, and services that collect, store, transmit, exchange, and process information in accordance with AI technology, such as intelligent robots and autonomous driving. Industry applications are the applications of AI products and services in industry, such as intelligent manufacturing, smart healthcare, and intelligent transportation. The hidden dangers of product services and industry applications are mainly manifested in:

AI applications are built on data, algorithm models, and infrastructure, and the hidden dangers of algorithm models, data security, privacy protection, and infrastructure will persist, **such that AI applications have a greater attack surface and privacy protection risks become more prominent.**

**Autonomous driving:** Due to the addition of connection control functions and new interfaces between IT back-end systems and other external information sources, network attack surfaces have grown significantly, such that autonomous driving now faces the vulnerability risks of the physical debugging interfaces, internal microprocessors, carrying terminals (运载终端), operating systems, communication protocols, and cloud platforms.

**Biometric features:** At the data acquisition stage, AI may face attack threats such as presentation attacks, replay attacks, and illegal tampering. In the biometrics storage stage, AI mainly faces threats to the biometrics database. In the biometrics comparison and decision-making stage, AI faces security threats such as comparison result tampering (比对结果篡改), decision threshold tampering (决策阈值篡改), and hill-climbing attacks (爬山攻击). There are threats such as illegal eavesdropping, replay attacks, and man-in-the-middle attacks on biometric data transmitted between biometric feature recognition modules (生物特征识别模块).

**Smart speakers:** There are vulnerabilities in the six aspects of hardware security, operating systems, application layer security, network communication security, AI security, and personal information protection. For example, for an open physical port or interface, an attacker can take advantage of the insecurity of the interface and memory chip. For instance, the hardware chip of the speaker can be directly disassembled and a back door can be embedded in the Flash chip to monitor and obtain the control of the smart speaker, tamper with the operating system, or steal personal data.

### 3.2.5 Abuse of AI

The abuse of AI has two layers of meaning: one is the improper or malicious use of AI technology to cause security threats and challenges; the second is the use of AI technology to cause uncontrollable security risks.

**On the one hand, the application of AI in attack methods such as fraud, dissemination of bad information, and password cracking has brought new challenges to traditional security detection.** Specifically, **the first is that cyberattack automation has become an obvious trend**; for instance, in the network field, attack activities that require a large amount of highly skilled labor (such as advanced persistent threat [APT] attacks) have gradually been highly automated. **Second, the spread of bad information (不良信息) has become more concealed**. Criminals can use AI technology to make the spread of all kinds of bad information more targeted and hidden, bringing tremendous hidden dangers and challenges to maintaining cybersecurity. **Third, more and more AI is used in fraud and other illegal crimes.** In 2017, there were many cases of criminals using speech synthesis technology to disguise the relatives of their victims in Zhejiang, Hubei, and other places, causing serious consequences and a negative impact on society. **Fourth, the probability of password cracking has increased.** The success of using AI technology to crack login verification codes is improving and becoming difficult to prevent. In 2018, a team at Northwest University established a set of verification code solvers (验证码求解器) based on AI technology. Using only 500 target verification codes to optimize the solver, it was able to obtain a verification code within 0.05 seconds.

**On the other hand, with the cross-integration of AI innovation technology into various fields, though it has been beneficial in promoting these fields, the issues of AI abuse have gradually grown prominent.** ISO/IEC PDTR 24028 divides the abuse of AI into three levels: misuse, which means the excessive dependence on AI will lead to unpredictable negative results; disuse, which means insufficient dependence on AI will lead to negative results; and abuse, which means that there is inadequate respect for the interests of end users when building AI systems. Because the application boundary of innovative technology is difficult to control, it may trigger risk of abuse, such as using AI technology to imitate humans, such as face swapping, handwriting forgery, voice mimicry, and chat robots. In addition to triggering ethical and moral risks, it may accelerate the application of technology in black and gray areas, blur the reasonable boundaries of technology application, and increase the risk of AI abuse.

## 3.3 Security implications

As an emerging industry that is developing rapidly, AI has both benefits and risks. In particular, with the in-depth application of AI in important industries such as defense, healthcare, transportation, and finance, if a serious security incident occurs or if AI is used improperly, it may affect national security, social ethics, cybersecurity, personal security, and personal privacy. Specifically, cybersecurity mainly affects the security of cyberspace (such as information security and system security), and personal privacy affects the rights and interests of personal information protection. This section mainly describes the impact on national security, social ethics, and personal safety.

**First, national security:** AI can be used to build new types of military strike forces and poses a threat to national defense and national security. The application of autonomous systems and robotics in the military, such as the production of AI weapons with automatic target recognition and precision strike capabilities, the production of military-related camouflage and decoys by generating adversarial networks, and AI systems that help improve radio spectrum allocation through electromagnetic countermeasures and machine learning will accelerate combat speed and enhance combat capabilities. **The use of AI to customize and disseminate information to target users can achieve the purpose of mobilizing public opinion.** By collecting user behavior data, using machine learning to analyze the user's political orientation and other portraits, pushing the desired content for users with different tendencies, and also influencing people's judgment on things by learning and simulating the speech of real people, once maliciously used, AI may cause a wide range of effects. **The massive application of AI in intelligence analysis (情报分析) has increased the risk of leakage of important state data.** AI technology has many uses in intelligence collection and analysis. For instance, intelligence workers can use it to obtain more data from monitoring, social media, and other channels, and by using AI technology to mine and analyze massive data, many important sensitive data can be obtained.

**Second, social ethical challenges: "Machine replacement of humans" will impact the employment of low- and medium-tech labor, which will exacerbate social differentiation and inequality in the long run.** The large-scale use of industrial robots and various intelligent technologies has exposed workers in labor-intensive, repetitive, and highly process-oriented industries to the threat of unemployment. Although some research reports have questioned excessive unemployment concerns, in the long run, AI will exacerbate social differentiation and inequality. This is particularly so for people with low levels of education, as the growing popularity of AI will make them significantly less competitive. **Reliance on AI technology will have an impact on existing social ethics, affecting existing interpersonal concepts and even human communication.** For instance, relying on personal data analysis, intelligent companion robots may be able to better understand the psychology of individuals and thus better meet user needs with extreme consideration and obedience. This may in turn reduce people's social needs. **The rapid development of AI technology has impacted the relatively stable written legal system. A lag and lack of legislation is unavoidable and may trigger legal difficulties that are difficult to pursue liability for.** For instance, given that AI is currently being applied to the healthcare and transportation industries, if AI results in medical misdiagnoses or autonomous driving results in traffic accidents, who should bear responsibility? How should the guardianship responsibilities of the designer and user of the AI model be distinguished from the responsibility of the AI system itself? Since AI has the ability to replace humans in decision-making and action, should AI be given certain subject rights under law? Relevant authorities, academia, and industries must carefully consider and discuss these issues.

**Third, personal safety risks: AI may impair personal safety due to flaws or malicious attacks in applications that are critical to personal safety.** With the in-depth integration of AI and IoT, smart products are increasingly being used in people's homes and in such fields as healthcare and transportation that are related to personal safety. Once these smart products (such as smart medical devices and autonomous vehicles) are subjected to network attacks or have loopholes and defects, they may endanger personal safety. **The application of AI in the development of weapons and other attack fields, if unconstrained, will pose a grave threat to personal safety.** AI technology may be used to develop weapons, and AI weapons developed with the help of facial recognition, automatic controls, and other technologies, such as "killer bees" (杀人蜂), can achieve fully automatic attack targeting. If AI weapons are allowed to choose and kill human beings, it will pose a grave threat to our personal safety and freedom.

## 3.4 Attributes and implications of AI security

In response to new threats such as adversarial examples, data poisoning, and model theft faced by AI, AI algorithm models, data, infrastructure, and product applications mainly face hidden dangers in such aspects as algorithmic bias, algorithm black boxes, algorithm defects, data security, privacy protection, software and hardware security, abuse, and ethics.

In order to prevent new AI attack threats and security risks, traditional cybersecurity attributes such as confidentiality, integrity, availability, controllability, and non-deniability must be expanded. For example, addressing algorithmic bias requires insisting on fairness, algorithm black boxes require enhanced explicability and transparency, algorithm flaws require improved robustness, issues of abuse require a focus on controllability, ethical and moral issues require being people-centered, and the protection principles of data security, privacy protection, and software and hardware security are similar to traditional principles.

In summary, this white paper gives the principles, attributes, and implications of AI security as follows:

**(1) AI security principles**

１） **Human orientation**: This refers to the concept that the research, development, and application of AI should aim at human goodness and human well-being and guarantee human dignity, and basic rights and freedoms.

２） **Parity of authority and responsibility:** Mechanisms are established to ensure that designers and operators of AI can be responsible for their results, such as accurate records, auditability, minimization of negative effects, trade-offs, and remediation.

３） **Classification**: Considering that the overall development of AI is still in its infancy, different standards for the classification and categorization of AI capabilities and specific functions can be established according to the maturity of the development of different AI technologies and the security needs of different application areas.

**(2) AI security attributes**

As an immature innovative technology, AI not only must ensure traditional security attributes such as confidentiality, integrity, and availability of AI assets, but also must consider robustness, transparency, and fairness in order to ensure its security in the in-depth application of important industries.

１） **Confidentiality**: This ensures that algorithm models and data in an AI system will not be leaked to unauthorized persons at any link in the life cycle (such as acquisition, training, or inference), such as preventing model theft attacks.

２） **Integrity**: This ensures that AI systems are not implanted, tampered with, replaced, or forged in any part of the life cycle (such as acquisition, training, or inference), algorithm models, data, infrastructure, and product applications, such as guarding against sampling attacks and data poisoning attacks.

３） **Availability**: This ensures that the use of AI algorithm models, data, infrastructure, and product applications will not be unreasonably rejected. Availability includes recoverability, that is, the ability of the system to quickly recover its operating state after an incident.

４） **Controllability**: This refers to the ability to control AI assets to prevent AI from being abused intentionally or unintentionally. Controllability includes verifiability and predictability. Verifiability means that AI systems should keep records and be able to test and verify the effectiveness of algorithm models or systems.

５） **Robustness**: This refers to the robustness of AI in the face of abnormal interference or input. For AI systems, robustness is mainly used to describe the ability of artificial intelligence systems to maintain their performance levels under conditions such as external interference or harsh environmental conditions. Robustness requires that AI systems take reliable preventive measures to prevent risks, that is, to minimize unintentional and accidental injuries and prevent unacceptable injuries.

６） **Transparency**: This provides visibility into the functions, components, and processes of the AI system. Transparency does not necessarily require the disclosure of AI algorithm source code or data, but depending on the different security levels of AI applications, transparency may have different levels of implementation and performance. Transparency typically includes explicability and traceability such that users can understand the decision-making process and causality of AI. Explicability refers to the mapping relationship between the feature space and the semantic space of the algorithm in an AI scenario, which enables the algorithm to understand the machine from a human perspective.

７） **Fairness**: This refers to the establishment of diverse design teams in the development process of AI systems, taking various measures to ensure that data is truly representative and representative of a diverse group of people to avoid biased or discriminatory results emerging from AI.

８） **Privacy**: Citizens' personal information is protected in accordance with personal information protection principles such as clear purpose, choice of consent, minimum use, openness and transparency, and subject participation.

**(3) AI security implications**

This white paper holds that AI security is still a part of cybersecurity, and according to the definition of cybersecurity under the Cybersecurity Law, AI security is defined thus: AI security refers to taking necessary measures to prevent attacks, intrusions, interference, destruction, and illegal use of AI systems and accidents to keep the AI system in a stable and reliable state of operation, so that it obeys the security principles of AI in being people-centered with parity of authority and responsibility to guarantee the integrity, confidentiality, availability, robustness, transparency, fairness, and privacy of AI algorithm models, data, systems and product applications.

# 4 AI security standards systems

## 4.1 Analysis of requirements for AI security standardization

At present, Chinese AI security-related standards mainly focus on application security standards in some fields such as biometric feature recognition and autonomous driving, as well as security standards for such supporting fields as big data security and personal information protection. However, there are few basic security standards directly related to the security or basic commonality of AI. Based on AI security attributes and connotations, incorporated with the current security risks faced by AI, with reference to the development direction of AI security policies and standards at home and abroad, this paper analyzes the standardization needs of AI from the dimensions of AI algorithm model security, data security and privacy protection, infrastructure security, product and application security, testing, and evaluation.

Incorporating the results of AI security risk analysis with the current state of standardization and the division of AI modules under the AI Standardization White Paper (2018), AI itself has standardization needs in the following aspects:

**(1) Standardization needs of AI data, algorithms, and model security**

Considering that AI algorithm models face security challenges such as issues of robustness and adversarial example attacks, SC42 has also developed standards for AI credibility and neural network robustness assessments. We recommend that the security requirements of the AI algorithm model be first addressed with full consideration of the robustness and credibility needs of the AI algorithm model when applied in China. Furthermore, AI algorithm model security indicators should be studied to develop algorithm model security assessment requirements and algorithm model standards for trustworthiness.

**(2) Standardization needs of AI data security and personal information protection**

The integrity, security, and personal information protection capabilities of AI data are important prerequisites for ensuring AI security, and research has already been carried out at home and abroad into standards and technologies related to AI data security and privacy protection. We recommend that standardization research be carried out to highlight data security and privacy protection risks. **First, research should address outstanding issues such as data poisoning, reverse attacks, and model theft faced by AI datasets**. Given the AI data life cycle, AI security standardization should cover such aspects as dataset protection, algorithm model protection, and resilience against reverse attacks. **Second, the effects of privacy protection should be balanced against AI analysis.** Security risks such as reverse engineering and privacy abuse should be prevented, and research into AI privacy protection requirements and technical standards should be carried out.

**(3) Standardization needs of AI infrastructure security**

The AI system includes the cloud-side, edge-side, terminal-side, and network transmission, and AI infrastructure faces risks in terms of software framework vulnerabilities as well as conventional software and hardware security risks. In addition to traditional cybersecurity requirements such as service interface security, software and hardware security, and service availability, we recommend that AI-specific security requirements be incorporated with special system security requirements. AI information system security standards should be developed with a focus on basic components and AI systems and platforms, such as open source algorithm frameworks, code security, and system security engineering.

**(4) Standardization needs of AI product and application security**

AI involves multiple dimensions such as data, algorithms, and infrastructure with a wide range of products and applications. These products and applications have the characteristics of high complexity, wide attack surfaces, and different security capabilities, and they still must also meet common security requirements. We recommend that priority be given to the standardization needs of smart products and applications with mature industrial development, urgent security needs, and imperfect standards. Analyze and review security risks in scenarios such as national and personal security, ethics, and intelligentized attacks, and study the basic common and specific security needs of different products and applications. Priority can be given to the selection of AI products such as smart door locks, smart speakers, and smart customer service for standardization and for the development of product and application guides and evaluation standards.

In addition, given that AI relies on data, algorithm models, and its underlying infrastructure and given its complex system of components, diverse risk dimensions, complex supply chain, and high-security operating requirements, we recommend that different types of standards for AI security risk management, supply chain security management, and safety operations be developed for AI research and application institutions and AI products and applications.

**(5) Standardization needs of AI testing and evaluation security**

The complexity of AI makes it face various risks and challenges, including algorithm model security risks, data security and privacy risks, and infrastructure security risks. We recommend that AI security test and evaluation indicators be designed so that they are fully compatible with existing supporting security standards and that priority be given to the development of testing and evaluation standards for such topics as algorithm robustness, AI system application trustworthiness, privacy protection, dataset security, and application.

## 4.2 Relationships between AI security standards and standards in other fields

The rapid development of AI in recent years is inseparable from the strong support of infrastructure such as big data and cloud computing. The formation of intelligentized services is also inseparable from the support of infrastructure. Therefore, in order to guarantee the security of AI, we should not only consider the security of AI itself, but we must also comprehensively consider fundamental security threats and challenges such as those facing data security, algorithm model security, infrastructure security, and cybersecurity. Standard development work in areas with AI characteristics must be fully compatible with current big data and cloud computing standards and standard systems. In particular, we must develop both fundamental AI standards and also security standards for the specific security requirements of AI scenarios. In addition, in more mature technologies and application scenarios such as biometric feature recognition and situational awareness, AI plays an increasingly important role. Existing security standards for such technologies and scenarios should be taken into account in standardization work. The security risks facing such AI scenarios should be comprehensively considered when proposing standard development plans.

## 4.3 AI security standards systems

As shown in Figure 4-1, the AI security standard architecture includes six parts: one, foundation; two, data, algorithms, and models; three, technology and systems; four, management and services; five, testing and evaluation; and six, products and applications, which mainly reflect the composition of each part of the standard system relationship.

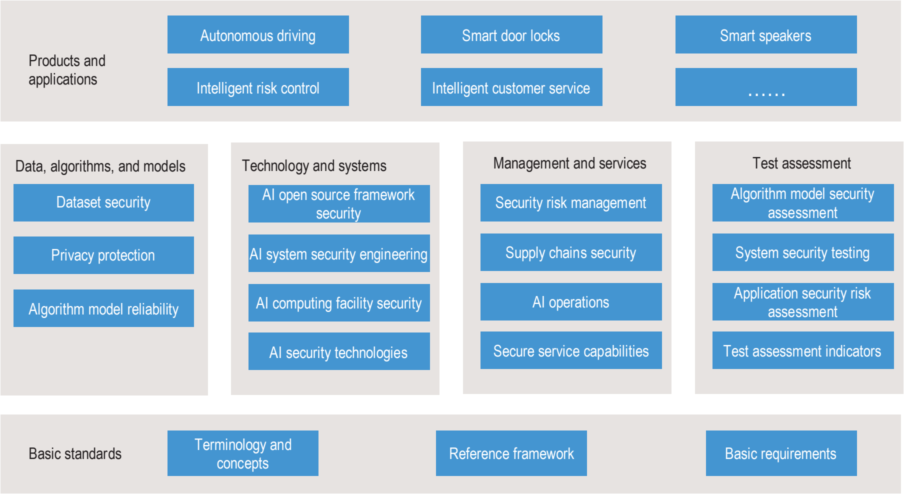


Figure 4-1 AI security standards systems

### 4.3.1 Fundamental security standards for AI

Fundamental AI security standards include AI concepts and terminology, security reference architecture, and fundamental security requirements.

– AI security concepts and terminology are the basic language for technical communication in AI security. Standardizing the definition of terms and the relationship between terms helps to accurately understand and express technical content and facilitate technical communication and research. Such standards must fully consider the normative definitions of AI concepts and terms issued by ISO, ITU-T, China's National Artificial Intelligence Standardization General Working Group, and other national and international standards organizations.

– The AI security reference architecture is the basis for understanding and further researching AI security. Through the security analysis of AI roles, an AI security model can be proposed to standardize AI security architecture, thereby helping to form an accurate understanding of structural levels, functional elements, and the relationships between them.

– AI fundamental security requirements standards are fundamental security principles and requirements for AI that are proposed in response to AI security risks, regulations, and policies to provide basic support for the system of AI security standards. These standards can guide relevant parties when building AI security and proposing requirements for data protection, algorithm security, and the design, development, and implementation of internal information systems. These standards provide technical requirements for the implementation of AI security practices and effectively protect the security of AI.

### 4.3.2 Security standards for AI data, algorithms, and models

Data, algorithm, and model security standards are proposed for prominent AI data, algorithm, and model security risks, including dataset security risks, privacy protection, and algorithm model trustworthiness.

– Dataset security standards mainly focus on the life cycle of AI data, guarantee the security of data labeling processes and data quality, guide the security management and protection of AI datasets, and reduce AI dataset security risks.

– Privacy protection standards are based on the privacy risks faced in the stages of AI development, operation, and maintenance, and AI privacy protection safety standards are formulated for privacy acquisition, utilization, storage, and sharing with a focus on preventing privacy data security risks from excessive privacy data acquisition, reverse engineering, and the abuse of privacy data. Such standards should be fully compatible with TC260's existing personal information protection standards and should focus on addressing privacy protection issues that are common to AI scenarios.

– Algorithm model trustworthiness standards mainly focus on security requirements such as algorithm model robustness, security protections, explicability, and algorithm bias. They address the issues of algorithm robustness and stability in natural operations, propose recoverability requirements and practical guidelines for extreme situations, and ensure the security of AI by implementing trustworthy AI algorithm models.

### 4.3.3 Security standards for AI technologies and systems

Technology and system standards are used to guarantee the security of AI open source frameworks and AI system security engineering.

– AI open source framework security standards put forward security requirements for AI server-side, client terminal-side, edge-side, and other computing and operating frameworks. In addition to open source framework software security, interface security, and traditional software platform security requirements, specific security requirements should be proposed for AI open source frameworks to ensure the security of the underlying support system for AI applications in training, operations, and other processes.

– AI system security engineering standards aim to address the security needs of analysis, design, development, testing, evaluation, operations, and maintenance. AI application development requirements and guidelines for outstanding risks, such as privacy protection and model security risks, are proposed with respect to aspects of data protection, model security, and code security, and security engineering implementation guidelines have been developed.

– AI computing infrastructure security standards aim to address the security needs of smart chips, smart servers, and other computing infrastructure and propose AI computing infrastructure security requirements and guideline standards.

– AI security technology standards are formulated in response to AI security protection and detection technologies, such as privacy-based machine learning, data bias detection, face swapping detection, adversarial example resistance, and federated learning.

### 4.3.4 Security standards for AI management and services

AI security management and service standards are mainly meant to guarantee AI management and service security and mainly cover security risk management, supply chain security, and AI safety operations.

– AI security risk management standards mainly respond to the multi-dimensional security risks of AI data, algorithm models, technologies and systems, management and services, and products and applications from the perspective of risk management. These standards propose security requirements and practical guidelines for technology, personnel, and management to provide guidance on and reduce overall AI security risks.

– AI supply chain security standards mainly review the supply chain security management needs of typical products, services, and roles from the perspective of supply chain security management. They refer to existing ICT supply chain security management standard development concepts and propose practical guidelines for AI supply chain security management to ensure the supply security of AI production factors.

– AI safety operation standards mainly aim to address the safe operation of AI services after they are launched, submitted, or officially in operation based on typical industry practice cases. These standards provide practical guidance from the perspectives of personnel safety, operational safety, and emergency response to reduce the safety risks of AI business continuity.

– AI security service capability standards mainly regulate the technical and management capabilities required for AI service providers to provide AI services externally.

### 4.3.5 Security standards for AI testing and evaluation

Testing and evaluation standards mainly analyze key security testing and evaluation points of AI algorithms, AI data, AI technology and systems, and AI applications and refine them into evaluation indicators for AI security testing. These standards analyze the key security testing points of products and applications with mature applications and urgent security needs and propose standards for such fundamental aspects of evaluation as AI algorithm models, system security, application risk, and testing and evaluation indicators. This includes but is not limited to:

– AI algorithm model security evaluation standards, which mainly focus on whether AI algorithms meet security requirements.

– AI system security evaluation standards, which mainly focus on whether the operation of AI systems meet security requirements.

– AI application security risk evaluation standards, which mainly focus on whether the application of AI meets security requirements.

– AI security testing and evaluation standards mainly refine AI security testing and evaluation indicators based on AI security requirements and specific object security requirements and lay the foundation for carrying out AI security testing and evaluation.

### 4.3.6 Security standards for AI products and applications

Product and application standards mainly ensure the security of AI technologies, services, and products in specific application scenarios. Standards can be developed in areas with mature applications, extensive use, or urgent security requirements such as autonomous driving, smart door locks, smart audio, smart risk control, and smart customer service. When developing standards, they must be fully compatible with general AI security requirements and take into account the specificity, urgency, and representative AI security risks of products and applications.

# 5 Recommendations for AI security standardization work

**(1) Attach importance to improving a system of AI security standards**

We recommend launching AI standardization work, coordinating the planning of a system of AI security standards, strengthening research into fundamental AI security standards, and deepening the work of AI application security standards. **The first is the overall planning of a system of AI security standards.** In order to ensure the orderly progress of the development of AI security standards, we recommend investigating and analyzing the need for AI security standardization in China and giving priority to research on a system of AI security standards. The system of standards should cover the security needs of multiple objects such as the foundation, platform, technology, products, and applications of AI and should be able to clearly define relationships with related standards such as big data security, personal information protection, cloud computing security, and IoT security. **Second, pay close attention to research and to implement the AI ethical principles.** Focus on the outstanding issues of AI algorithm discrimination and algorithmic bias. Analyze and review the ethical needs of AI in various scenarios, develop and refine AI ethical principles, and guide the implementation of principles and requirements related to AI standards.

**(2) Accelerate the development of standards in key areas**

AI has the characteristics of wide coverage, complex application scenarios, and many types of security requirements. We recommend that AI security standardization promotion plans be established to develop AI security standards in accordance with the concept of “emergency use first, driven by safety incidents.” The development of standards should be accelerated in key areas, and AI security standardization should be advanced in an orderly fashion. First is fundamental AI security standards research. China’s AI security standards mainly focus on the field of application security, lacking security standards for AI itself or basic commonality. We recommend that fundamental AI security standards be strengthened with requirements for AI security aspects such as monitoring and early warning, risk assessment, security accountability, and research personnel security guidelines based on the New Generation Artificial Intelligence Development Plan. Standards research should be conducted for such aspects as AI security reference architecture, security risks, ethical design, and security assessment to grasp AI algorithm security threats and protection needs to clarify general algorithm security principles and requirements and to strengthen AI algorithm model security and robustness. AI algorithm models and smart product security requirements and evaluation methods should be standardized to address the data quality and dataset security issues faced by AI. **Second is a deepening of AI application security standards.** Given that AI is becoming ever more integrated with various application fields, we should begin to study AI products and application security standards. We recommend that smart product security standards be ahead of AI application standards. AI security standardization characteristics and requirements should be extracted based on the development of standards for smart door locks, smart homes, and other fields that have been carried out by the National Information Security Standardization Technical Committee. Afterwards, we should give priority to areas where there is an urgent need for standardization, more mature applications, urgent security needs, wide applications, or areas that are relatively sensitive. We should develop AI products and application security standards to improve the security requirements of existing AI standards. **Third is to develop standards according to the concept of “fulsome research, with priority to urgent needs, driven by security incidents.”** We recommend that priority be given to security standards with pressing security needs and mature applications. For instance, standards such as Artificial Intelligence Security Reference Framework, Artificial Intelligence Dataset Security, Artificial Intelligence Data Labeling Security, Machine Learning Algorithm Model Trustworthiness, Artificial Intelligence Open Source Framework Security, Artificial Intelligence Application Security Guide, and Artificial Intelligence Security Service Capabilities Requirements should be prioritized. Simultaneously, research work on fundamental AI standards should be conducted with the study and application of security risk assessment standards and key AI product and service security standards such as smart manufacturing and intelligent networked vehicles. We should gradually promote research on AI security standards in other fields.

**(3) Diligently promote the application of AI security standards**

In order to improve the effectiveness and operability of AI security standards, address prominent AI security risks, and explore a path to standardization of the most difficult and pressing issues of AI security, we recommend that in-depth application and practical work be carried out for AI security standards. **First, improve the pilot mechanisms for AI security standards.** Select a number of pilot enterprises to carry out an evaluation of the applicability of the standards and the effectiveness of implementation, establish the concept of work of “tracking practice, discovering issues, summarizing experiences, improving standards, and providing feedback for the next step of standardization,” and promote the rapid and high-quality development of AI security standardization. **Second, improve the research, promotion, and application of promotion mechanisms for AI security standards.** Organize universities, scientific research institutes, and enterprises in jointly breaking past AI security standardization difficulties, give play to the advantages of various institutions, establish AI security standards research, promotion, and application mechanisms that integrate "industry, university, and research," and promote the benign development of the AI industry.

**(4) Effectively strengthen the training of AI security standardization talent**

Talent is the cornerstone of AI security standardization work, and we recommend that a multi-level and multi-type AI security talent training mechanism be established and improved upon. **First, train AI security professionals.** Establish training programs for professional technology, standard setting, publicity and habituation training (宣贯培训), and testing and evaluation. **Second, encourage universities, research institutes, and enterprises to establish collaboration.** Explore training paths that integrate talent into AI security. **Third, strengthen support for artificial intelligence security and standardization projects.** Optimize the ratio of scientific research resources and management and evaluation mechanisms to ensure that relevant talent fields concentrate on solving key AI security issues.

**(5) Actively participate in international AI security standardization**

International organizations such as ISO and IEEE have organized and undertaken a number of AI security standardization research tasks and have achieved certain standardization results in some areas. We recommend that the achievements of international and overseas standardization work be fully digested, absorbed, and incorporated with China's AI security needs in order to explore AI security standardization work paths with Chinese characteristics. **First, closely track and study the dynamics and development trends of AI security standardization work at home and abroad.** Consolidate the research achievements of international AI security standardization, absorb the experiences of standards development overseas, and promote the better development of AI security standardization in China. **Second, continuously enhance the influence of China's international standards in the field of AI security.** Diligently support the participation of Chinese institutions and experts in international standardization work, strengthen research into AI security standards, and encourage Chinese experts to serve in international standardization organizations and act as editors of international standard projects. **Third, give full play to China's international standardization exchange and cooperation mechanisms.** Given the rich application scenarios of China's AI industry, standards cooperation and exchange mechanisms in the field of AI security should be established to enrich the achievements of Chinese AI security standardization work with the help of international and overseas efforts.

**(6) Establish an AI high security risk early warning mechanism as soon as possible**

In view of AI technologies, products, and applications with high security risks, we recommend that an early warning mechanism be studied and proposed for highly dangerous AI security risks. **First, establish a catalog of highly dangerous AI security risks.** Review AI technologies, products, and applications with outstanding security risks that may create high-impact security issues and classify and categorize risk items within the catalog. **Second, establish an AI high security risk early warning mechanism.** Incorporate technology, application, and product characteristics and propose an early warning plan for risks. **Third,**

**study and formulate highly dangerous AI safety risk management standards.** Starting with standards, propose a security risk management plan for highly dangerous AI that incorporates techniques, management, evaluation, and other means and covers such aspects as risk identification, analysis, and processing.

### (7) Effectively improve AI security supervision support capabilities

Standardization can strongly support the implementation of AI security supervision, and we recommend the development of an index evaluation system for AI security standards. **First, establish a sound AI supervision system and formulate supporting standards.** The government should use standards as a powerful starting point to establish a monitoring system that runs through the entire cycle of AI development, design, data acquisition, and market application to prevent AI from being used illegally or used in areas that deviate from its intended purpose. **Second, we recommend accelerating research into AI supply chain security management mechanisms.** Develop supporting standards for AI supply chain security, propose AI supply chain procurement requirements for telecommunications, energy, transportation, power, finance, and other industries, and promote the piloting of relevant standards in key areas to provide a useful reference for Chinese party and government departments, and for key industries, to manage AI supply chain security risks and offer practical guidance for enterprises as they strengthen AI supply chain management.

**Appendix A**

# AI-related security standards

## A.1 TC260 AI Security Standards Research Projects

Table A-1 Domestic AI Security Standards Research Projects

| No. | Standard Content | Standard Type |
|---|---|---|
| 1 | **"AI Security Standards Research"**<br>Prepared by the National Information Security Standardization Technical Committee. This was China's first national AI security standard research project. This project investigated the policies, standards, and industry status of AI security at home and abroad, analyzed security threats and risk challenges faced by AI, reviewed security cases in various AI application areas, refined the need for AI security standardization, and studied the system of AI security standards. | Research |
| 2 | **"Security Guide for AI Applications"**<br>Prepared by the National Information Security Standardization Technical Committee. This project aims to use AI applications as a starting point for analyzing AI application security and to lay a foundation for proposing AI security application standards. | Research |

## A.2 TC260 AI Security-Related Standards

Table A-2 Domestic AI Security Standards

| No. | Standard Content | Standard Type |
|---|---|---|
| 1 | **GB/T 20979-2019 "Information Security Technology - Technical Requirements for Iris Recognition System"**<br>Proposed by the National Information Security Standardization Technical Committee. This standard specifies the technical requirements of iris recognition systems that use iris recognition technology to provide support for identity authentication. | Revised |
| 2 | **GB/T 36651-2018 "Information Security Technology - Biometric Feature Recognition Authentication Protocol Based on Trusted Environment"**<br>Proposed by the National Information Security Standardization Technical Committee. This standard specifies a Biometric Feature Recognition authentication protocol based on a trusted environment, including protocol framework, protocol process, protocol requirements, and protocol interfaces. | Formulated |
| 3 | **GB/T 37076-2018 "Information Security Technology - Technical Requirements for Fingerprint Recognition System"**<br>Proposed by the National Information Security Standardization Technical Committee. This standard analyzes security threats and security objectives of fingerprint recognition systems, regulates the potential security risks of fingerprint recognition systems, proposes technical requirements for the security of fingerprint recognition systems, and regulates the application of fingerprint recognition technology in the field of information security. | Formulated |

| | | |
|---|---|---|
| 4 | **"Information Security Technology - Cybersecurity Guidebook for Vehicle Electronic Systems"**<br>Prepared by the National Information Security Standardization Technical Committee. By absorbing and adopting practical experience in industry and academia, this provides practical guidance for the cybersecurity activities of vehicle electronic systems. | Formulated |
| 5 | **"Information Security Technology - Vehicle Network Equipment Requirements"**<br>Prepared by the National Information Security Standardization Technical Committee. This aims to propose a solution to the standards issues of information security technology requirements for in-vehicle network equipment in the intelligent connected automotive industry. This establishes scientific and unified standards for in-vehicle network equipment information security technology requirements. | Formulated |
| 6 | **"Information Security Technology - General Technology Requirements for Smart Home Security"**<br>Prepared by the National Information Security Standardization Technical Committee. This standard specifies general technology requirements for smart home security, including the overall framework, security models, device security requirements, gateway security requirements, cybersecurity requirements, and application service platform security requirements for smart homes. This standard applies to the safe design and implementation of smart home products and can be used for reference in the security testing and management of smart homes. | Formulated |
| 7 | **"Information Security Technology - Technical Requirements and Test Evaluation Methods for Smart Door Locks"**<br>Prepared by the National Information Security Standardization Technical Committee. The goal of this standard is to stipulate technical requirements for information security and testing and evaluation methods for smart door locks and to address the new security issues for smart door locks, such as Tesla coil attacks (特斯拉线圈攻击), biometric information counterfeiting, and remote control risks. In this way, this standard can allow research institutions to standardize information security design and development of products at the beginning of the product and application design stage. This will comprehensively improve the security of products, promote the healthy and orderly development of the industry, safeguard cyberspace, including smart door lock systems, and guarantee the security of people's lives and property. | Formulated |

## A.3 ISO/IEC JTC1/SC42 AI Security-Related Standards

Table A-3 State of SC 42 AI Security-Related Standards Development Work

| Name of Working Group | Convener | State of Work |
|---|---|---|
| WG 1 Foundational Standards Working Group | Canada | ISO/IEC 22989 "Artificial Intelligence Concepts and Terminology"<br>ISO/IEC 23053 "Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)" |
| WG 2 Big Data Working Group | United States | ISO/IEC 20546 "Information technology – Big data – Overview and vocabulary" (Published)<br>ISO/IEC 20547-2 "Information technology – Big data reference architecture – Part 2: Use cases and derived requirements" (Published)<br>ISO/IEC 20547-5 "Information technology – Big data reference architecture – Part 5: Standards roadmap" (Published)<br>ISO/IEC 20547-1 "Information technology – Big data reference architecture – Part 1: Framework and application process"<br>ISO/IEC 20547-3 "Information technology – Big data reference architecture – Part 3: Reference architecture" |

| | | |
|---|---|---|
| WG 3 Trustworthiness Working Group | Ireland | ISO/IEC TR 24027 “Information technology – Artificial Intelligence (AI) – Bias in AI systems and AI aided decision making”<br>ISO/IEC TR 24028 “Information technology – Artificial intelligence – Overview of trustworthiness in artificial intelligence”<br>ISO/IEC TR 24029-1 “Artificial Intelligence (AI) – Assessment of the robustness of neural networks – Part 1: Overview”<br>ISO/ TR 23894 “Information Technology – Artificial Intelligence – Risk Management”<br>TR “Information Technology – Artificial Intelligence – Overview of Ethics and Social Concern” |
| WG 4 Use Cases and Applications Working Group | Japan | TR “Information Technology – Artificial Intelligence – Use Cases” |
| WG 5 Computational Approaches and Computational Characteristics of AI Systems Working Group | China | TR Artificial Intelligence — Study on Computational Approaches and System Standardization |

**Appendix B**

# Cases of AI application security in practice

(listed in no particular order)

## B.1 AI security in practice at Baidu

Adversarial example attacks have truly migrated from the laboratory environment into actual cyber confrontations, increasing the threats to personal privacy, property security, traffic safety, and public safety. This has hampered the trustworthy and healthy development of AI in various industries. For the generation of adversarial examples involving fundamental issues such as the explicability of deep neural networks, the industry has yet to come up with a comprehensive solution. In the process of AI implementation, Baidu has gradually developed an overall solution, AdvBox, that covers security verification, model reinforcement, adversarial example detection, and formal verification of model robustness.

1） **Security verification:** This aims to verify model security, typically from the two dimensions of sensitivity to environmental changes and adversarial examples. Forms of verification include white box verification and black box verification. For models with deep neural network structures and parameters that are completely public, AdvBox constructs targeted adversarial example datasets to test a model's adversarial performance under different disturbances. When only limited model information can be acquired, AdvBox verifies the prediction results of different inputs through a trial model. Taking facial recognition APIs provided by various public cloud service providers as an example, different adversarial examples can be uploaded, and prediction results are recorded to gradually enhance the effect of the simulated attack. In addition, because adversarial examples are often migratory, that is, the adversarial examples constructed by a model can often be used to deceive other black box models, AdvBox uses adversarial example migration when testing model security. Users can use AdvBox to simulate attacks against models, using generated adversarial examples to deceive their target model without accessing the target model the generated adversarial samples to deceive the target model, but instead only locally attacking a similarly constructed model.

2） **Model reinforcement:** After discovering the security flaws of the AI model, AdvBox will adopt technical means such as adversarial training and input data preprocessing and a generative adversarial network (GAN) to strengthen and protect the model. Of these means, adversarial training refers to adding adversarial examples to the training set so that the model has better resistance to adversarial examples. In addition, AdvBox also reduces the impact of disturbances on model prediction accuracy through data preprocessing, modification of the model's activation function or loss function, and network feature compression. Finally, AdvBox trains a neural network model based on GAN to enhance the generalization ability of the model. The reinforced model can effectively reduce the success rate of black box and white box attacks.

3） **Adversarial example detection:** As more and more AI services are now provided externally as APIs, to defend against adversarial example attack risks for AI APIs exposed to public networks, based on its model reinforcement, AdvBox detects the compliance of input data with its adversarial example detection technology. Comprehensively used detection methods include local intrinsic dimension (局部本征维数)-based methods, model explicability methods, and consistency comparisons of continuous frame prediction results. In deep learning-based detection tasks, once AdvBox detects an adversarial example, it will immediately prevent the example from entering by reporting the anomaly. This is of great significance for application scenarios such as malware detection and internet content compliance review. Black and gray product practitioners (黑灰产从业者) always attempt various methods to bypass malware detection and content review, and adversarial examples are bound to become one of their tools. Therefore, it is very important to have a model in place for detecting adversarial examples. Specifically, the adversarial example detection process is shown in Figure B-1, where x is the input, function T is the transformation performed on the input, and function D is a measure of the inconsistency of the prediction results. If the inconsistency of the prediction result exceeds the threshold after different changes are made, it is determined to be an adversarial example.

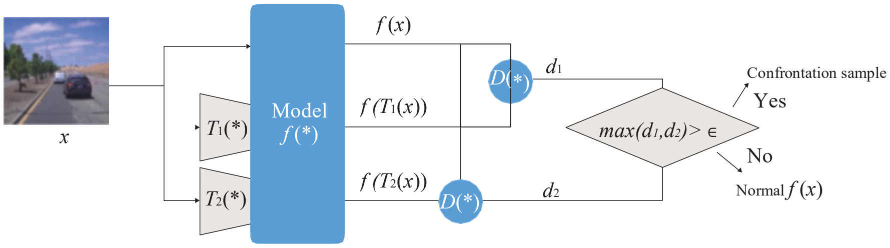


Figure B-1 Adversarial example detection process

4） **Formal verification of model robustness:** Given the ultra-high dimensional characteristics and complex structure of deep learning models, if a sample testing method is used for security verification, there is no way to determine whether the selected limited samples can fully cover all representative scenarios, and thus model security cannot be effectively proven. In addition, traversal authentication (遍历验证) likewise cannot effectively prove model security. AdvBox uses formal verification methods for model security protection. Specifically, from the construction of adversarial examples, AdvBox will find the range and lower bound for required disturbance security. In extreme scenarios, the model is automatically invalidated, and alternative solutions are enabled to avoid AI misjudgment causing harm to personal property or the public.

## B.2 AI security in practice at OrionStar

AI training data is the core data asset of all AI technology companies, and the security of training data is the core foundation to ensure that the effectiveness of an AI model meets expectations. Speech recognition, speech synthesis, visual recognition, and natural speech processing all require a great deal of training data support. AI data includes voice data files in common formats, dialogue corpus text files, facial image data, or item image data. The amount of data tends to be quite large. If AI data is not properly managed, it may easily lead to issues such as chaotic AI data management, low AI dataset security, privacy leaks, and poor AI model training results.

When strengthening AI core data asset security management while also ensuring the efficiency and stability of the AI training process, it becomes difficult to apply traditional data encryption and decryption methods to AI data security protection plans. OrionStar developed its own AI automated training platform. Through this unified training platform, training tasks are separated from the training data to achieve automation of the AI training process and visualization of the training results. Training data is hidden in an isolated storage medium to ensure the security of big data assets. This can also optimize the scheduling management of GPU computing resources to maximize their utilization.

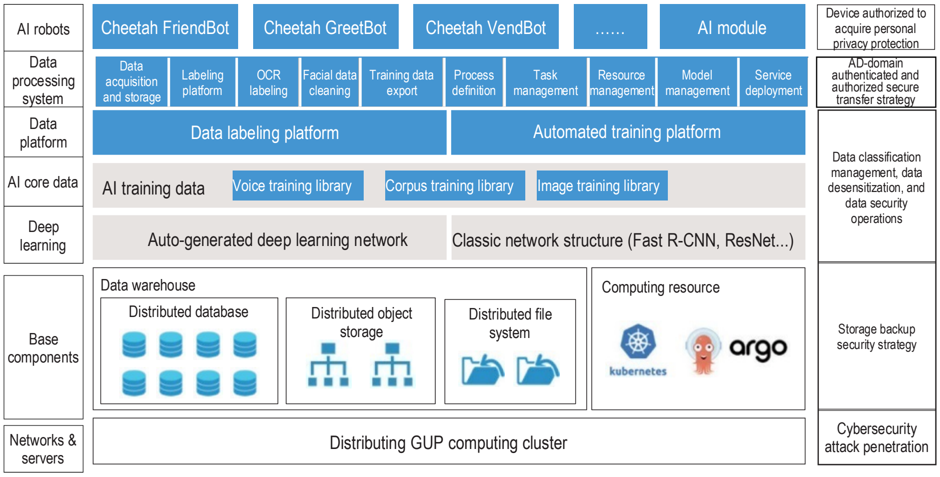


Figure B-2 OrionStar AI automated training platform architecture

As shown in Figure B-2, based on the business characteristics and data ownership of OrionStar's AI, OrionStar designed a system architecture for its AI automation training platform with reference to relevant big data security standards. First, in terms of business model design, the GPU service cluster used by the AI training service platform adopts unified domain account authorization management.

Using the distributed file storage method, group management is carried out according to different training needs of speech, vision, and natural language understanding. When research personnel initiate training tasks, they only need to select the training target, set the data range, and select the type of training. The system can then allocate computing resources and mount relevant training data based on preset information. The training platform ensures that 1) the business as a whole is reasonable and clear about data authorization boundaries, 2) data processing logic is based on available and invisible security principles, and 3) the application output of data is based on the value of the data rather than the output of bare data.

Secondly, the automated training platform builds a comprehensive data management and control system based on data service stages, including data security management and control, before, during, and after data processing. Specifically, at the data compliance layer, the platform refers to GB/T 35273-2017 Information Security Technology - Personal Information Security Specifications, GB/T 35274-2017 Information Security Technology - Big Data Service Security Capability Requirements, GB/T 31168-2014 Information Security Technology - Cloud Computing Service Security Capability Requirements, and the ISO 27001 series of standards and achieves personal privacy information protection and cloud service security and guarantees the security of big data services.

Then, based on the standard big data security classification requirements, the automated training platform uses the active directory (AD) domain authentication and authorization method to ensure that employees can only log in to the system after authorization. The data range and data scale of all training tasks are then managed through data classification, with authorization based on the role and responsibilities of an employee's position. A jump server is then used to keep track of all employee operations.

Finally, the automated training platform provides comprehensive protection for infrastructure security, storage security, system security, application security, and platform network service security. Based on the Big Data Security Standardization White Paper, OrionStar has established AI data security systems, principles, strategies, management plans, and implementation rules to ensure the security of AI data throughout the processes of resource integration, sharing, release, and exchange.

## B.3 AI security in practice at Tsinghua University

Although AI is rapidly promoting a technological revolution and industrial progress, its security risks are often overlooked. Research has found that many algorithms that perform well on datasets are easily deceived by adversarial examples that are invisible to the human eye, resulting in AI system misjudgment. The question of how to build a security platform that can support protection against different types of adversarial examples has attracted the attention of researchers in this field. Using these platforms can make it easier for researchers to implement counterattacks and defenses and thus develop more robust and more secure deep learning models while providing model performance benchmarks for AI adversarial research. However, most of the platforms currently available only support adversarial attack algorithm research and lack support for defense algorithms. Also, most platforms do not support a flexible definition of functions such as loss functions or adversarial detection. Furthermore, there is a lack of uniform standards and comparison indicators for robustness evaluations and model comparison.

Tsinghua University has developed the RealSafe algorithm platform to respond to typical AI adversarial attack and defense issues for different application scenarios and different threat scenarios. This platform incorporates a standard programming library with the three layers of system, algorithm, and application. RealSafe is an open source standard programming library, available to industry for non-commercial use for free. This programming library provides platform support for China's development of theoretical algorithms and standard formulation for AI security. As for the three layers, the system layer can implement common universal modules with different adversarial attack and defense algorithms and includes different kinds of module support, such as model structure and loss function settings. The algorithm layer supports the efficient implementation of mainstream adversarial attack and defense methods from the two aspects of adversarial attack and adversarial defense. The application layer supports adversarial attack and defense verification of related applications from different levels such as image, video, voice, and network data. The development of the adversarial attack and defense platform will greatly reduce the threshold for the development and use of related models. Furthermore, by developing general algorithm modules, new model development costs can be reduced. This platform can provide a unified evaluation standard for various AI model security and for various AI model attack and defense algorithms.

The RealSafe platform meets the current standardization needs of AI methods and the deficiencies of existing platforms. It can extract common modules of different algorithms, support flexible settings for different security applications, improve the development and research efficiency of AI attack and defense algorithms, and evaluate the security of mainstream AI methods, and largely meet the needs of future AI security standardization testing. As in Figure B-3, the platform includes:

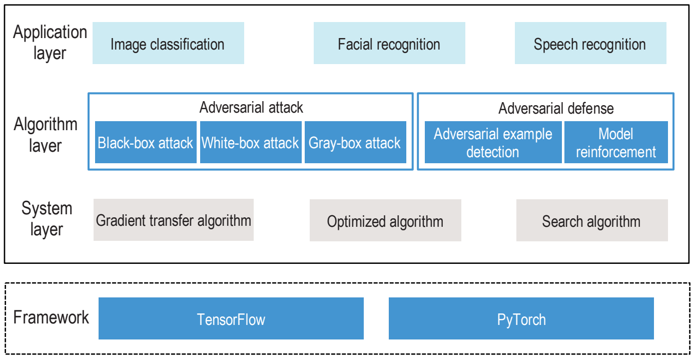


Figure B-3 AI security platform architecture

1) **Attack and defense algorithm performance benchmarks:** Based on the defined model interfaces for different application scenarios, attack and defense algorithms are implemented for different threat scenarios. Benchmarks are then provided for evaluating the robustness of specific AI models, reducing the development cost to evaluate the robustness of new models and to attack new models. Current AI security platforms do not provide full support for typical attack and defense algorithms and thus cannot provide an objective and comprehensive evaluation benchmark. This platform establishes a complete index system for the security issues of AI models to achieve an evaluation of AI algorithm security.

2） **The algorithm platform is divided into the three layers of the application layer, system layer, and algorithm layer**. Specifically:

**System layer:** This layer provides a unified basic algorithm interface for the system layer with the help of reasonable abstraction, including: gradient transfer algorithms that support different machine learning frameworks, such as TensorFlow and PyTorch, various optimization algorithms, including algorithms based on stochastic gradient descent and quasi-Newton methods, and various search algorithms, such as an adversarial example search based on random walk that is used in some black box attack methods.

**Algorithm layer:** This layer provides support for typical adversarial attack algorithms and adversarial defense algorithms for multiple threat models. Adversarial attacks include: (1) black-box and gray-box attacks, including methods based on decision boundary attacks and migration-based attacks, and (2) white-box attacks, including multiple attack methods based on gradient information. Adversarial defense includes defense based on detection of adversarial examples and defense based on model enhancement, for example.

**Application layer:** This layer provides support for typical models of typical application scenarios, such as image classification and facial recognition, provides model interfaces to access the system layer for adversarial attack and defense models, thereby reducing the development cost of adding new application models to the framework, and provides model security and an attack and defense algorithm performance evaluation interface.

## B.4 AI security in practice at YITU Technology

In response to typical AI security issues, YITU has proposed relevant requirements and solutions for AI system security technologies. YITU divides AI systems into two categories: equipment elements and network elements. Among them, equipment elements refer to computer products that store, process, or apply AI-related information, including chips, operating systems, application software, and other related hardware and software. Network elements refer to computer products or applications, including network transmission, that transmit, exchange, or share AI-related information. Table B-1 shows each component element and corresponding security requirements.

Table B-1 Security requirements of the components of the AI system

| Element | Components | Privacy protection | Algorithm Calibration | Transparent Supervision |
|---|---|---|---|---|
| AI device | Chips | √ | √ | √ |
| | Operating system | √ | — | — |
| | Application | √ | √ | √ |
| | Other | √ | √ | √ |
| Network | Network transmission | √ | — | — |

From the perspective of actual application scenarios, as shown in Figure B-4, YITU divides its AI system into four parts, namely, the cloud-side, edge-side, terminal-side, and network transmission, and has established a security management center based on these four parts along with an AI system security system architecture.

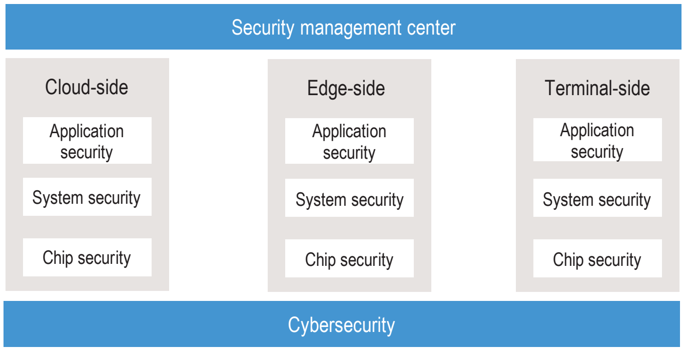


Figure B-4 AI system security architecture

Taking facial recognition as an example, the facial recognition information system includes private cloud storage on the cloud-side and model calculation using vast quantities of data, edge-side recognition algorithms, terminal-side data acquisition, and the components of a network link that transmit data between the cloud, the edge, and the terminal. Specifically,

**1）A "security committee" is established to ensure complete AI application security from "production" to "application"**

YITU strictly controls and executes all aspects of its security strategy through the "security committee" to ensure that the information security strategy and measures of AI applications are properly implemented.

**2）The YITU QuestCore (“求索”) chip adopts a complete security design to ensure the secure and reliable use of the chip at all levels**

The YITU QuestCore neural network chip acts as a complete security technical solution with management specifications from algorithm development to firmware release and firing to ensure the security and reliability of the chip. At a technical level,

**Encrypted area within chip:** Encrypted areas come prefabricated within the QuestCore chip, and unique identification ensures that the system is not stolen, avoiding the risk of server cloning.

TrustZone technology: Chip hardware and software resources are divided into a secure domain and a non-secure domain. This technology ensures that key computing processes and core key exchanges are performed in the secure domain.

**Data model encryption:** This provides the encryption function for the algorithm's core data model. This ensures that model data can only be accessed and called by trusted applications. This also ensures that data model files are not lost.

**Self-deployed transmission encryption algorithm:** Secure and reliable data is ensured with the use of a self-deployed encryption algorithm. In addition, this chip also adopts security design solutions such as firmware read protect technology and physical attack protection.

**3）Security designs have been added to AI algorithms to guarantee that “algorithmic bias,” “model disclosure,” “sample attack,” “data corruption,” and other issues are avoided**

In terms of algorithm design, algorithm explicability is ensured through technical and management measures to avoid algorithmic bias. In the application of AI, a multi-dimensional direction is used to increase complexity in combination with noise training to prevent feature reverse attacks, thereby ensuring that algorithm model security does not leak. In addition, in the algorithm model training process, related processes and tools are designed. By adding appropriate labels in the algorithm model and incorporating these labels with input samples in practical applications, when an attacker changes the input sample, the contaminated parameters can be located and the contaminated sample can be corrected or rejected, thereby preventing malicious attacks against the sample to a certain extent. In addition, data and data transmission paths can be encrypted with a variety of symmetric and asymmetric encryption algorithms and multiple digital signature technologies which cannot be cracked by current theoretical computing power. By ensuring data integrity and untamperability, data-centric security can be achieved. YITU has strengthened the audit design for AI applications to ensure comprehensive, standardized, independent, and secure audits.

**4）For data labeling specifications, labeled data verification, input and output verification, and other processes, a data governance mechanism with technology and system protection was established.**

In terms of training data security, a focus was placed on preventing intruders from contaminating data. On the one hand, automated tools are used to perform basic labeled data verification. During the input and output process of the training model, deviation verification is again performed to prevent contaminated data from further contaminating the AI model. On the other hand, security awareness training is carried out during personnel management, and secure operation regulations have been formulated and implemented.

**5）Basic environment security components, encrypted communications, and self-developed encryption algorithms for functions such as "password-free secure login" under the LAN environment are deployed.**

Using this basic environment security component, when cluster login is used, users can only log in through the master node, rather than logging in directly to child nodes in the cluster. In port management, this basic environment component can perform "whitelist" management and permission control on application ports. In system security, YITU has ensured system security by strengthening operation audits, implementing audit modules, and embedding self-deployed operating systems. In terms of cybersecurity, "cloud," "edge," and "end" terminal identity authentication mechanisms were designed, and a self-deployed encryption algorithm was deployed in the transmission channel.

## B.5 AI security in practice at IBM

AI is of great significance and increasingly impacts human decision-making. To this end, IBM has proposed three AI principles, including:

**1）Focus on purpose:** The purpose of AI developed and applied by IBM is to enhance human intelligence and capabilities. AI should continue to be controlled by humans.

**2）Emphasize transparency:** People's trust in the use of AI is crucial. The foundation of trust is transparency. People need to know how AI is reliable, fair, and explicable.

**3）Popular skills:** AI will be widely used in all aspects of human life. Only by popularizing AI skills can the public fully understand, trust, and utilize AI.

To achieve AI's reliability, fairness, accountability, and implementation throughout the AI life cycle, IBM has developed and open-sourced a series of key technologies for trusted AI. Specifically, the AI Fairness 360 Toolbox (AIF360) can be used to detect and mitigate bias in machine learning models; the Adversarial Robustness 360 Toolbox (ART) can be used to quickly make and analyze the attack and defense methods of machine learning models; and the AI Explainability 360 (AIX360) can be used to support the explainability of machine learning models and algorithms. This series of open source technologies helps to promote the innovation and application of trusted AI. In order to meet the needs of enterprises for technology integration support, assurance, and services, IBM has integrated this series of technologies and IBM's years of AI practice into the enterprise version of IBM Watson OpenScale to meet the needs of enterprises. Taking AIF360 as an example, AIF360 embodies the transparency of IBM's trusted AI.

AIF360 can help detect, mitigate, or eliminate bias in machine learning models. AIF360 contains a comprehensive set of datasets, models, indicators, and algorithms to detect, mitigate, or eliminate bias in machine learning models.

### 1 ）Bias and machine learning

AIF360 has designed fairness indicators and bias mitigators to check whether datasets, machine learning models, and machine learning algorithms are biased. Fairness indicators can be used to check for bias in machine learning workflows. Bias mitigators can be used to overcome bias in the workflow to produce fairer results.

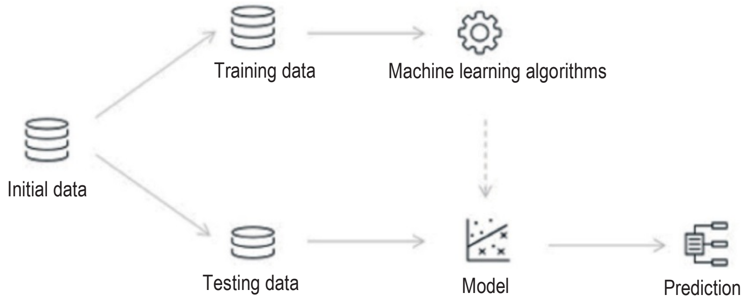


Figure B-5 AIF360 Workflow

As shown in Figure B-5, bias can enter the system at any link in the workflow. Training datasets, algorithms, and test datasets may all cause bias. Specifically, (1) in training datasets, training data may be biased towards specific types of instances. (2) In algorithm modeling, the algorithm may generate a model weighted for specific features in the input. (3) In test datasets, test datasets may have biased expectations for correct answers.

### 2 ）Detecting and mitigating bias

The bias detection process starts with an initial dataset, and uses a random segmentation algorithm to split the initial dataset into a training dataset and a test dataset.

First, when the initial dataset is loaded, a protected attribute is established with a privileged group and non-privileged group. During testing, the percentage of favorable results between the privileged and non-privileged group are compared, and the former is subtracted from the latter. If the result is negative, it means that the attribute is not good for the vulnerable group, and the result is the detected deviation.

To reduce this type of bias in training datasets, AIF360 provides a variety of indicators and bias mitigation algorithms for detecting bias. For example, the "reweighting algorithm" (重新加权算法) can apply appropriate weights to different groups in the training dataset so that the training dataset does distinguish in terms of sensitive attributes. The new training dataset generated by this conversion will then obtain fairer results in terms of age attributes for the privileged and non-privileged groups.

Then, the same method used to test the original training dataset is used to check the results of the new converted training dataset in eliminating deviations. For the existing dataset, if the deviation result is 0, it proves the effectiveness of the bias mitigation method, indicating that the bias of the biased dataset has been eliminated.

## B.6 AI security in practice at Sangfor Technologies

AI technology empowers the cybersecurity field, which helps manufacturers, enterprises, and individuals effectively improve their ability to deal with cybersecurity issues such as cyber fraud and malicious attacks. Compared with traditional technologies, AI has the advantages of a strong capacity for generalization, slow degradation, and low broadband and memory use. As such, AI can offer breakthrough cybersecurity protection. Sangfor Secure Cloud Brain (深信服安全云脑) has threat intelligence and detection capabilities for unknown threats. Specifically,

1） The Threat Intelligence Center gathers vast quantities of intelligence data from channels such as Sangfor's online security devices, third-party security vendors, and security communities. The Threat Intelligence Center was established based on big data and cloud computing technology, using various data analysis and artificial intelligence algorithms.

2） The Unknown Threat Detection Center can use threat detection engines such as cloud sandboxes, AI antivirus, AI detection engines, honeypot technology, and URL identification engines, and incorporates manual analysis by senior security analysts to appraise and identify unknown and suspicious threats.

As shown in Figure B-6, in terms of the physical architecture, the Secure Cloud Brain is divided into the access layer, platform layer, engine layer, data layer, and application layer.

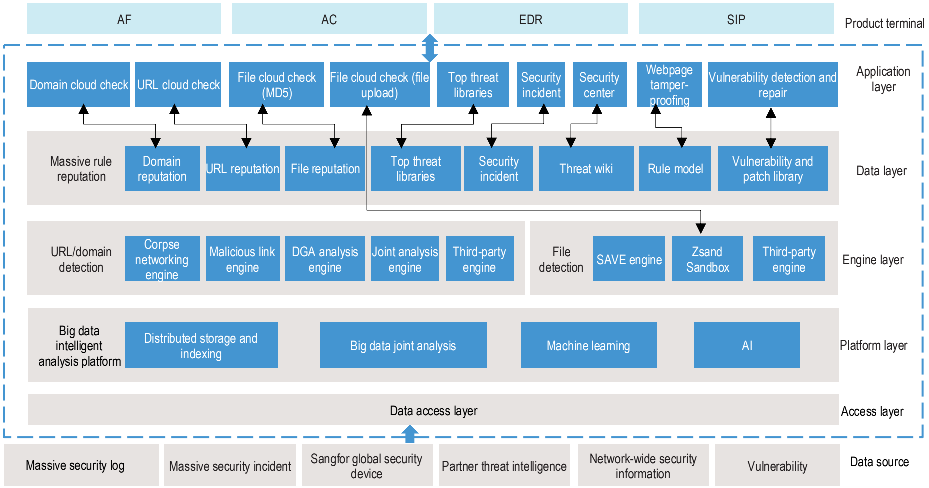


Figure B-6 Secure Cloud Brain architecture

Attackers' attacks on AI are becoming more diversified and ever-evolving. This may cause the AI model of the existing Secure Cloud Brain to fail or misjudge information. In order to ensure the reliability of the Secure Cloud Brain, the model should be updated in time to respond to attackers' attacks on the model. However, several security risks may be faced when updating the model: 1) if updates are only based upon user-side data, the unbiasedness and completeness of the data cannot be guaranteed, which may reduce the reliability of the user-side model; 2) if updates are based on a small amount of local data, the brain may be easily deceived by an attacker's carefully forged data, destroying the distribution of real data such that the model makes wrong judgments; 3) if updates are based on all real data but the sample distribution is not balanced, the model may result in an insufficient prediction accuracy for a small number of types.

To this end, Sangfor has established a unified data acquisition and processing mechanism for the Secure Cloud Brain to ensure the diversity and reliability of data sources and to update model training in the cloud and distribute the new model to all security devices so as to improve the anti-attack capabilities and accuracy of the AI model. In addition, in the application process, Sangfor has improved the explainability, anti-attack capabilities, and accuracy of the Secure Cloud Brain AI model. Specifically:

**Explainability:** By incorporating AI and heuristic rules, the capabilities of security experts and data experts are utilized such that data experts can continuously update models to improve AI generalization capabilities. Then, security experts can extract and update the detected variants to enhance the intuitiveness of the rules. The closed loop iteration of AI + rules can enhance explicability.

**Anti-attack capabilities:** The model updates online based on a massive reputation database (信誉库) and multi-dimensional intelligence sources. By incorporating an in-depth analysis system based on association analysis and intelligent algorithms, various data layers are screened and comprehensively analyzed to ensure the diversity and reliability of data sources and to ultimately output high-accuracy judgment results.

**Robustness:** a. The model will continue to evolve and periodically self-test. Once accuracy is below a certain threshold or new samples that are not covered by the model appear, the model will trigger training in the cloud to ensure the model's robustness. b. Association analysis is as shown in Figure A-7. The engine is mainly constructed based on algorithms such as deep learning and graph computing (图计算) and assists the analysis of security experts to form human-computer knowledge (人机共智). The core association module uses deep learning algorithms to analyze DNS traffic characteristics and co-occurrence patterns. The family clustering module then calculates community association analysis-related technologies based on graphs to discover unknown families through clustering. The proof module combines multi-dimensional information to prove the results of the previous two modules, thereby enhancing the robustness of the results.

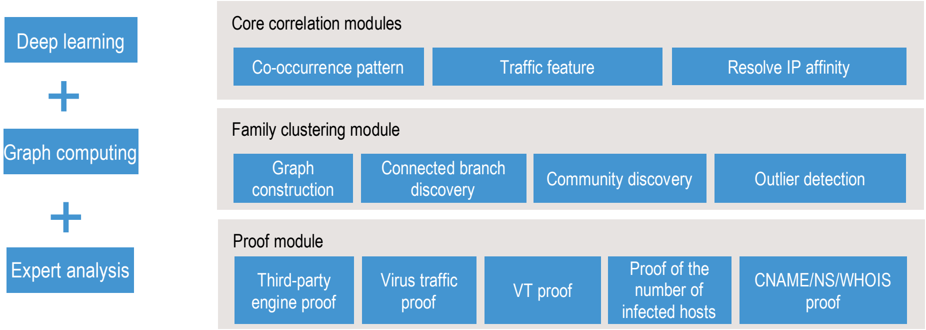


Figure B-7 Correlation analysis engine architecture diagram

## B.7 AI security in practice at 360 Total Security

Intelligent vehicles are an important asset of intelligent transportation. With the development of intelligentization and network connections, they bring security risks to intelligent transportation. Facing increasingly severe safety challenges, 360 has created a dynamic defense system for intelligent transportation security through 360's AI analysis-based “security brain” that comprehensively guarantees intelligent traffic safety.

The overall framework of the intelligent vehicle security brain includes the telematics (车联网) security analysis engine and the telematics secure operations service platform. At the same time, through security protection and security resource services, it provides secure operations support such as security warnings, analysis decisions, and system management, as shown in Figure B-8:

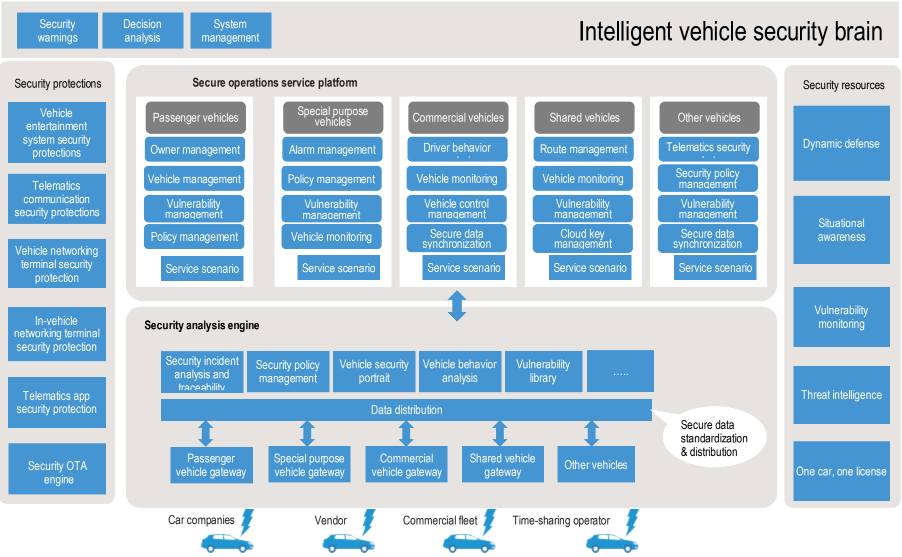


Figure B-8 Intelligent vehicle security brain framework

### 1） Security analysis engine

After standardizing vehicle data, the security analysis engine performs security portraits, behavior analysis, and strategy management for the vehicle.

### 2） Security resources

This product contains basic security resources such as password certificates and secure storage, supports SM2, SM3, SM4, and other forms of state secret algorithm (国密算法) communication, and provides resources such as vulnerability monitoring and threat intelligence.

### 3） Security protections

This product includes vehicle entertainment system security protection, vehicle network terminal security protection, telematics communication security protection, and telematics app security protection. Security patches are issued promptly to protect other vehicles in advance and to ensure the overall security of intelligent transportation.

### 4） Operations service platform

Targeted services are provided based on scenarios, such as passenger vehicles, special vehicles, commercial vehicles, and shared vehicles, with a focus on intelligent transportation security in different scenarios. By integrating resources and optimizing AI algorithms, functions such as vulnerability management, security data synchronization, and security policy management are provided to enhance the security operations capabilities of intelligent transportation.

The core capabilities of the intelligent vehicle security brain include defense, monitoring, upgrades, and services, as shown in Figure B-9.

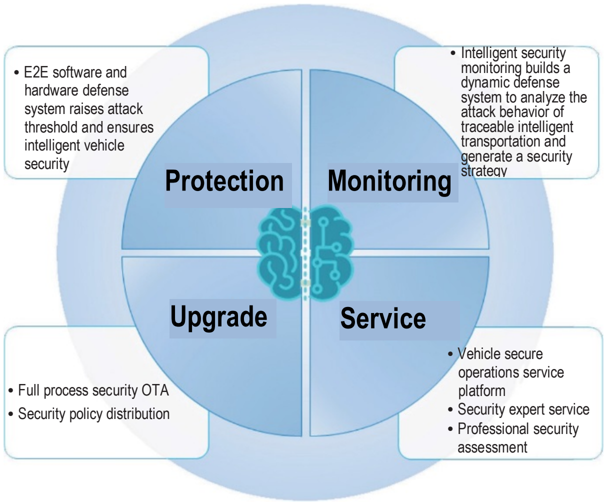

• E2E software and hardware defense system raises attack threshold and ensures intelligent vehicle security
Protection
Monitoring
• Intelligent security monitoring builds a dynamic defense system to analyze the attack behavior of traceable intelligent transportation and generate a security strategy
Upgrade
Service
• Full process security OTA
• Security policy distribution
• Vehicle secure operations service platform
• Security expert service
• Professional security assessment

Figure B-9 Diagram of intelligent vehicle security brain core capabilities

The intelligent vehicle security brain includes three aspects: a unified data access layer, a secure data computing platform layer, and a security analysis application layer, as shown in Figure B-10.

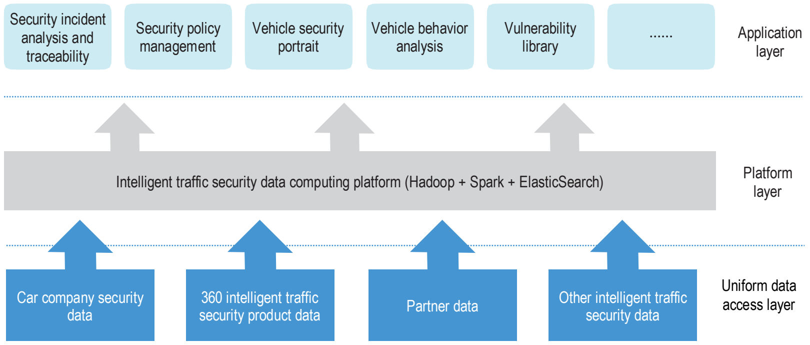

Security incident analysis and traceability
Security policy management
Vehicle security portrait
Vehicle behavior analysis
Vulnerability library
......
Application layer
Intelligent traffic security data computing platform (Hadoop + Spark + ElasticSearch)
Platform layer
Car company security data
360 intelligent traffic security product data
Partner data
Other intelligent traffic security data
Uniform data access layer

Figure B-10 Intelligent vehicle security brain system framework

### 1） Uniform data access layer

This layer accesses the vehicle's telematics security data, 360 telematics security product data, and partner data through a set of standard data access interfaces and forwards this security data to the security data computing platform layer to achieve an efficient and real-time data channel with the automaker's platform and to meet the data access requirements of intelligent connected vehicles across different automakers, different models, and different regions.

### 2） Security data computing platform layer

The Telematics Security Storage Center and Computing Center can compute and analyze vast quantities of telematics data. In terms of data processing capabilities, on the one hand, large-scale Hadoop clusters and Spark clusters provide computing power, and on the other hand, by building GPU clusters, they support the analysis and mining of vehicle portraits and vehicle behavior data based on AI technology.

### 3） Security analysis application layer

Security portraits and vehicle behavior analysis are conducted by analyzing vehicle behavior data, providing vehicle threat intelligent support for secure operations. This is primarily used to establish vehicle information security incident analysis, traceability, and response mechanisms, and by applying AI technology, this can achieve vehicle information security awareness, early warning, dynamic defense, and security strategy updates. In addition, the application layer will also establish a vehicle information security vulnerability database through continuous monitoring to discover networked vehicle security vulnerabilities and monitor for potential vehicle decoding (破解) behaviors.

## B.8 AI security in practice at Alibaba

### (1) Real person authentication

To address the inability of traditional network identity authentication methods to resolve many security threats and risks brought about by identity fraud and identity legitimacy issues, to ensure that the identity of users in e-commerce, social networking, new retail, and other businesses is true and effective, and to accurately identify false identities, Alibaba has used AI technology to build a real person authentication system. The technical framework of the real person authentication system is shown in Figure B-11.

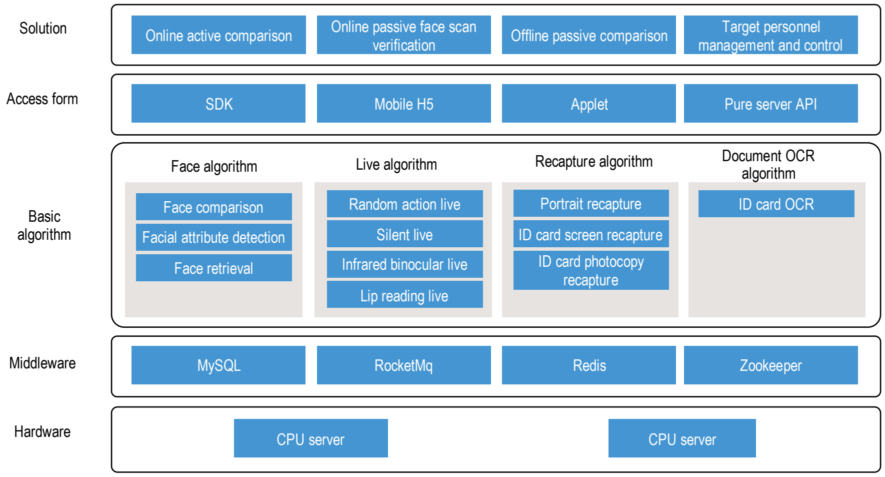


Figure B-11 Technical framework of real person authentication system

The business architecture of the real person authentication system is shown in Figure B-12. The real person authentication system can use living body (活体) detection, face comparison, and other biometric technologies, document OCR identification technologies, and other methods to provide real-name verification and biometric services, including:

1） **Real-name verification:** OCR identification technology is adopted for documents to automatically identify and read information such as name, ID number, and expiration date and to verify and compare this information with an authoritative database. The comprehensive recognition rate of this technology is over 99%.

2） **Biometrics:** Photos are retrieved through video to verify whether the object of identification is the same person. In order to prevent spoofing attacks, the goal of identifying real people can be achieved through interactive operations. In some scenarios with limited use of facial recognition, human-based identity and motion recognition can provide an effective supplement.

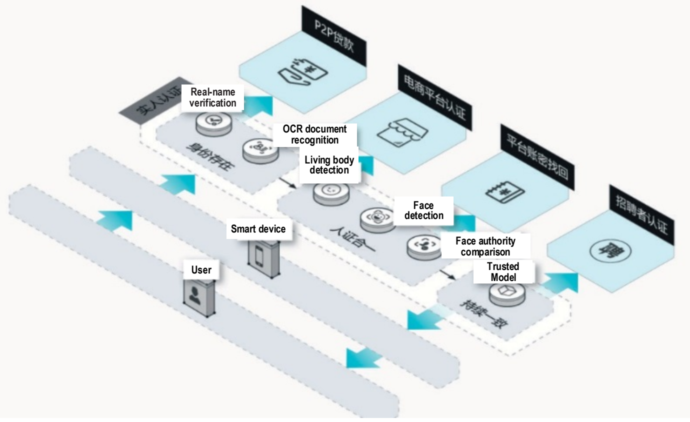


Figure B-12 Business architecture of real person authentication system

## (2) Adversarial example attack and defense — Question answering verification codes

Adversarial attack technologies for AI models themselves present many AI security issues and risks. Alibaba summarizes adversarial example attack defense technology based on business requirements as shown in Figure B-13. This can be divided into three stages: adversarial attack, security assessment, and adversarial defense.

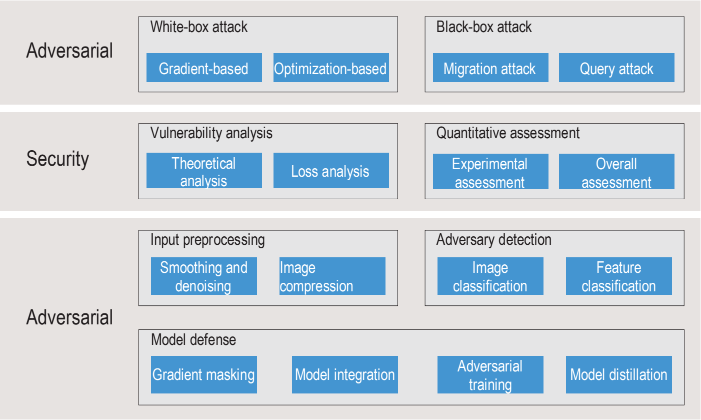


Figure B-13 Summary of adversarial example attack defense technology

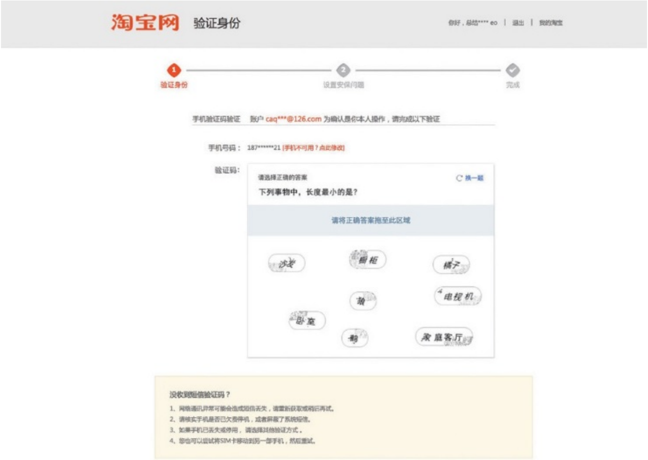
Alibaba uses the above technology system to conduct many application practices such as question answering verification codes, security assessments, and AI firewalls in its business.

In recent years, for the most common character verification codes, criminals have collected a large number of images through various means and have used machine learning technology to train optical character recognition (OCR) models to achieve automatic recognition of verification codes, with an accuracy of over 80%. This has allowed them to commit criminal acts such as stealing user accounts and maliciously registering coupon codes (薅羊毛).

Alibaba Security officially launched a new generation of AI verification code products at the end of 2018 for use on Taobao and Tmall as well as in other business scenarios. This product combines knowledge graph and adversarial example attack technology to lower machine decoding by black and gray products (黑灰产) without affecting the user experience.

### 1) Knowledge graphs

A rich common sense question and answer library based on a structured knowledge graph can be used to effectively avoid attacks, as shown in Figure B-14. The so-called knowledge graph is composed of some entities, entity attributes, and relationships between entities, such as the comparison of the weights of elephants and tigers. This knowledge graph can generate a billion-level question bank, and the response time for a user is about 9 seconds with an answer pass rate of 90%. Most ordinary users have no difficulty in answering these questions, but it is extremely difficult for a machine to answer automatically.



手机验证码验证　账户 caq***@126.com 为确认是你本人操作，请完成以下验证

手机号码：187******21 [手机不可用？点此修改]

验证码：

请选择正确的答案

下列事物中，长度最小的是？

请将正确答案拖至此区域

没收到短信验证码？

1. 网络通讯异常可能会造成短信丢失，请重新获取或稍后再试。
2. 请核实手机是否已欠费停机，或者屏蔽了系统短信。
3. 如果手机已丢失或停用，请选择其他验证方式。
4. 您也可以尝试将SIM卡移动到另一部手机，然后重试。

B-14 Common knowledge question and answer database case based on knowledge graph [The question lists a number of household objects and animals and asks: "of the following things, which is the shortest in length?"]

**2) Adversarial examples**

As answers to the human-machine verification quiz are displayed to users in the form of pictures, if black and gray products wish to crack this kind of verification code, they must rely on AI technology to automatically recognize text within the picture. Alibaba applies the latest adversarial example technology in the field of AI research to specifically add interference to the original image without affecting naked eye recognition. However, it will significantly reduce the recognition rate of the AI model, thereby preventing the cracking of the coding platform while maintaining the user experience.

## B.9 AI security in practice at Huawei

Biometric feature recognition is a technology used for identification based on biometric feature information, and AI technology is widely used in biometric feature recognition. At present, biometric feature recognition is mainly used in fingerprint/face unlocking and fingerprint/face payment scenarios to solve security issues such as identity authentication in physical scenarios.

Biometric feature information generally has strong privacy attributes. To ensure the security of biometric feature information, Huawei mobile phones complete processes such as acquisition of fingerprint/face images, feature extraction, feature comparison, and feature storage in a chip-level isolated Trusted Execution Environment (TEE). Fingerprint/face feature data is then encrypted and stored through TEE's secured storage or the Replay Protected Memory Block (RPMB), and the built-in security chip is used to encrypt/decrypt fingerprint/face feature data to ensure that sensitive personal data does not leave the TEE.

Biometric feature information data such as fingerprints/faces are collected and transmitted to a secure isolated area through a secure acquisition device and a secure channel. This data then only exists in the TEE together with other sensitive data such as keys. At the same time, fingerprint/face verification (such as live detection, feature extraction, and feature comparison), password verification, and other operations are all completed in TEE. Fingerprint/face feature information and encryption keys in the built-in security chip cannot be obtained externally to ensure that face data will not be leaked. This then protects the security of sensitive user data and related services. In Android's facial recognition framework, the framework is only responsible for processing data such as authentication initiation and results for fingerprints/faces and does not involve fingerprint/face data itself. Android's third-party applications also cannot obtain the user's fingerprint/face data, nor can they pass the user's fingerprint/face data out of the terminal device. This mechanism further protects the user's personal privacy and the security of sensitive data. The specific framework is shown in Figure B-15.

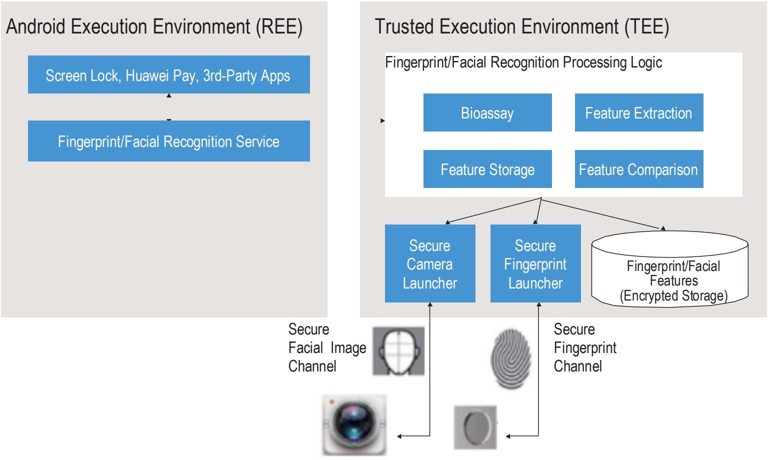

Android Execution Environment (REE)
Screen Lock, Huawei Pay, 3rd-Party Apps
Fingerprint/Facial Recognition Service
Trusted Execution Environment (TEE)
Fingerprint/Facial Recognition Processing Logic
Bioassay
Feature Extraction
Feature Storage
Feature Comparison
Secure Camera Launcher
Secure Fingerprint Launcher
Fingerprint/Facial Features (Encrypted Storage)
Secure Facial Image Channel
Secure Fingerprint Channel

Figure B-15 Biometric feature data secure storage/processing framework

For mobile phone users, apart from the leakage of biometric information, snooping content displayed on the user's mobile phone screen may also lead to user privacy and personal data leakage risks. Huawei has proposed solutions that incorporate AI technology for snooping prevention and other user privacy protections. The mobile phone can use AI technology to detect the presence of the owner/non-owner or multiple people viewing the screen. When the mobile phone recognizes that there are strangers or multiple people viewing the screen, private information displayed on the mobile phone screen will be automatically hidden to protect the user's private information. The specific implementation logic flow is shown in Figure B-16:

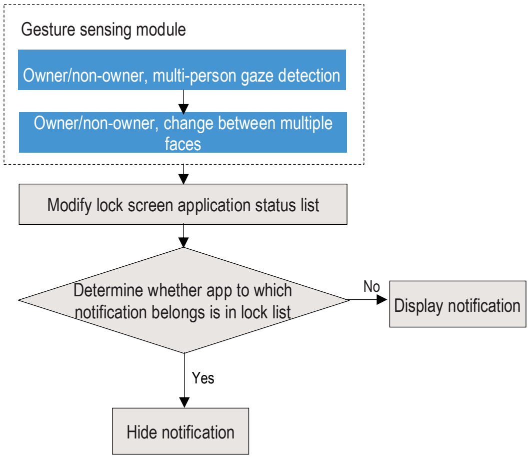
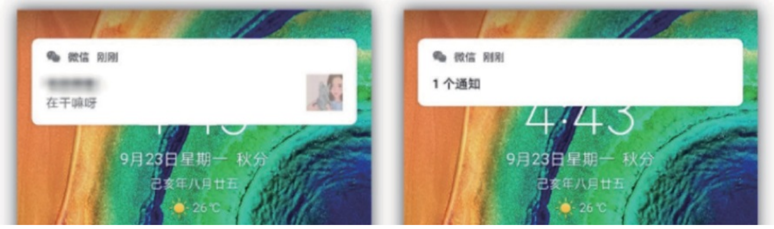

Gesture sensing module
Owner/non-owner, multi-person gaze detection
Owner/non-owner, change between multiple faces
Modify lock screen application status list
Determine whether app to which notification belongs is in lock list
No
Display notification
Yes
Hide notification

Figure B-16 Logic flow chart of message hiding based on AI technology

Through the posture sensing module, the mobile phone can recognize viewership by the owner/non-owner or multiple people. When a change is identified in the face of the owner/non-owner or multiple people, it will indicate that others are viewing. At this time, by modifying the locked application status list accordingly, the locked screen status of applications will be changed. The mobile phone will then determine whether the notifications displayed on the screen belong to applications in the locked list. If a notification is for an application on the list, it will hide the application's notification. Otherwise, the notification will be displayed. Simply put, as shown in Figure B-17, when a user uses his or her phone alone, information will be expanded. When the phone detects that a stranger/multiple people are viewing the screen, the user's privacy will be protected by “collapsing” (折叠) the information to ensure that application information is only visible to the user himself or herself.



微信 刚刚

[redacted]

在干嘛呀

微信 刚刚

1 个通知

Figure B-17 Message hiding function based on AI technology [the left screen features a username and the text of an SMS message; the right screen reads "1 notification"]

## References

[1] National Artificial Intelligence Standardization General Working Group, Artificial Intelligence Standardization White Paper (2018), January 2018.

[2] Tan Tieniu: "History, Status, and Future of Artificial Intelligence." *Seeking Truth*, April 2019.

[3] Alibaba Cloud: Chinese Enterprises 2020: Artificial Intelligence Application Practices and Trends. August 2019.

[4] Liu Yan: AI Security: Introduction to Adversarial Examples. Machinery Industry Press, June 2019.

[5] National Industrial Information Security Development Research Center: 2019 China Artificial Intelligence Industry Development Index. September 2019.

[6] Deloitte: White Paper on China's Artificial Intelligence Industry. November 2018.

[7] Tsinghua University: Artificial Intelligence Chip Technology White Paper (2018). December 2018.

[8] Information & Data Security Solutions Co., Ltd.: AI Data Security Risk and Governance. September 2019.

[9] Zhang Bo: A new journey towards the third generation of artificial intelligence. September 2019.

[10] Tan Tieniu: Artificial Intelligence - Angel or Demon. June 2018.

[11] China Academy of Information and Communications Technology (CAICt). White Paper on the Development of Artificial Intelligence: Industrial Applications (2018), December 2018.

[12] Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., Song, D.: Robust physical-world attacks on deep learning visual classification. CVPR 2018.

[13] ISO/IEC PDTR 24028 Information technology – Artificial intelligence (AI) – Overview of trustworthiness in artificial intelligence.

[14] Li Pan, Zhao Wentao, Liu Qiang, et al.: Review of machine learning security issues and their defense technologies [J]. Computer Science and Exploration, 2018(2): 171-184.

[15] National Artificial Intelligence Standardization General Working Group: Artificial Intelligence Ethical Risk Analysis Report. April 2019.

[16] Liu Jinyang: "Complexity and Ethical Challenges of Artificial Intelligence Algorithms." *Guangming Daily*, September 2017.

[17] European Union: Ethics guidelines for trustworthy AI. April 2019.

[18] Cong Mo.: Laughing about the development of AI in China, Zhang Bo, Li Deyi, Zhang Zhengyou and Xiao Jing discuss "Five Questions of AI" on the same stage. AI Technology Review, September 2019.