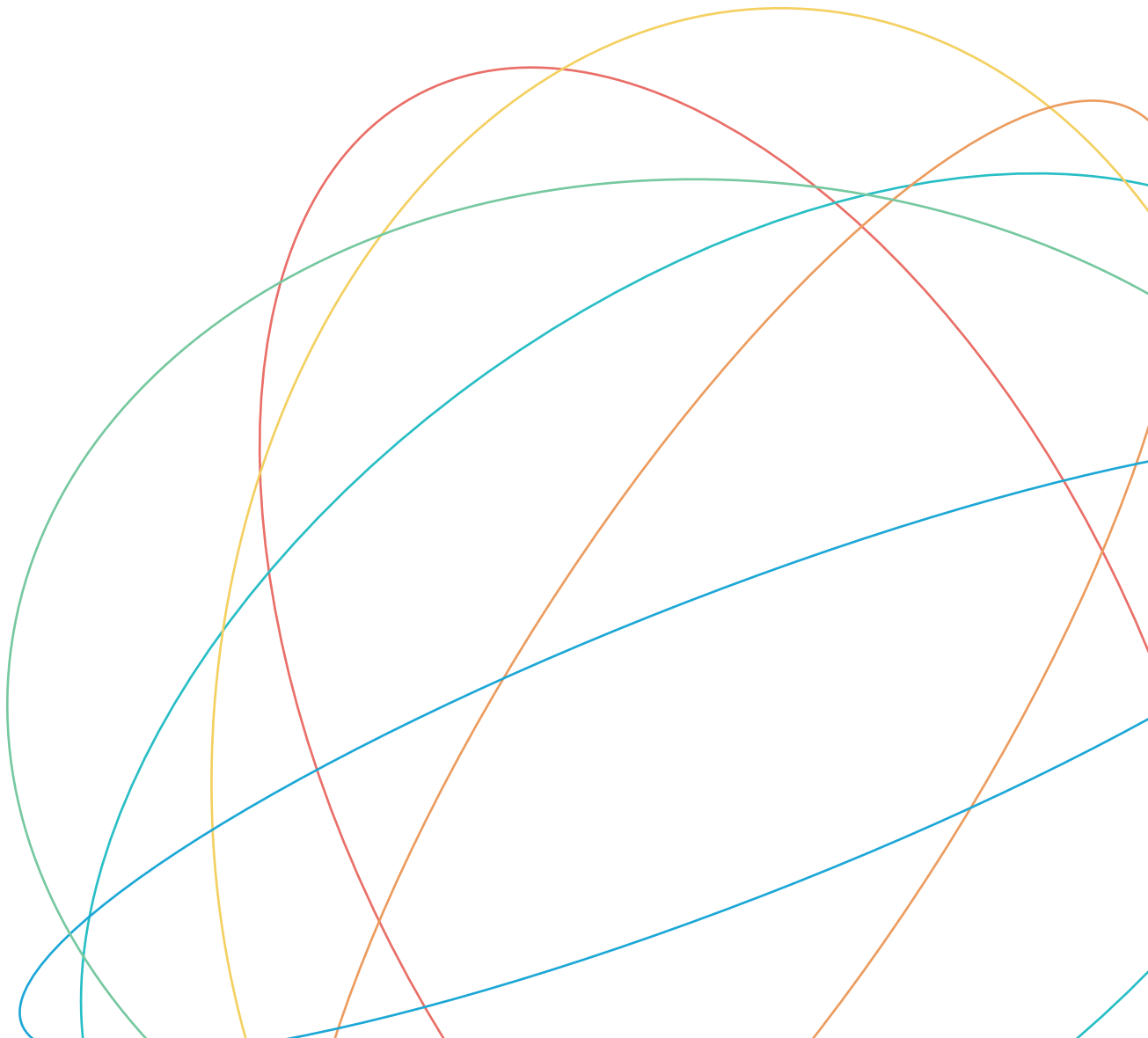




AI Security White Paper



Executive Summary

With the accumulation of big data, dramatic improvements in computing power, and continuous innovation in Machine Learning (ML) methods, Artificial Intelligence (AI) technologies such as image recognition, voice recognition, and natural language processing have become ubiquitous. Meanwhile, AI poses a significant impact on computer security: on the one hand, AI can be used to build defensive systems such as malware and network attack detection; on the other hand, AI might be exploited to launch more effective attacks. In some scenarios, the security of AI systems is a matter of life and death. Thus, building robust AI systems that are immune to external interference is essential. **AI can benefit security and vice versa.**

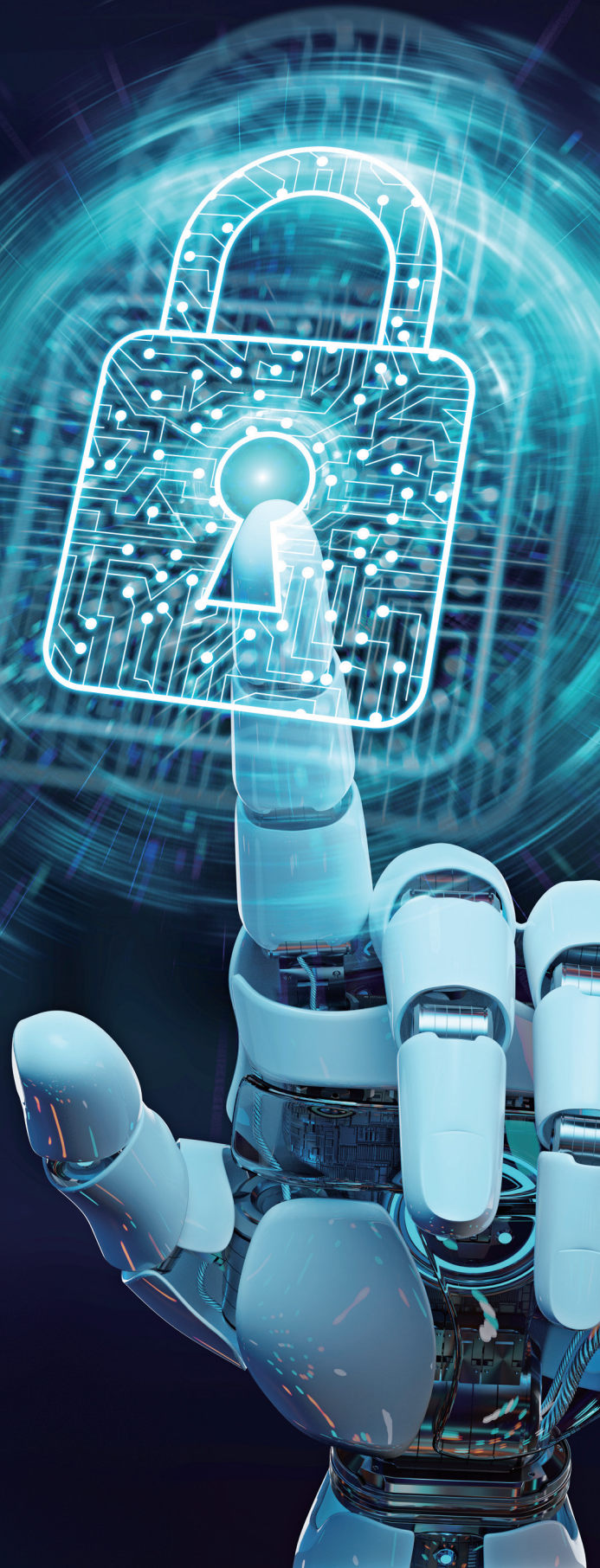
The main purpose of this white paper is to explore the security of AI, in terms of protecting the integrity and confidentiality of AI models and data, and thus preventing attackers from changing the inference results or stealing the data.

Unlike security vulnerabilities in traditional systems, the root cause of security weaknesses in ML systems lies in the lack of explainability in AI systems. This lack of explainability leaves openings that can be exploited by adversarial machine learning methods such as evasion, poisoning, and backdoor attacks. These attacks are very effective and have strong transferability among different ML models, and thus pose serious security threats to Deep Neural Network (DNN)-based AI applications. For instance, attackers can inject malicious data in the training stage to affect the inference of AI models or add a small perturbation to the input samples in the inference stage to alter the inference result. Attackers may also implant backdoors in models and launch targeted attacks or extract model parameters or training data from query results.

Huawei is dedicated to AI security research with the goal of providing a confident and secure AI application environment, and contributing to the AI-enabled intelligent world. In order to tackle the new AI security challenges, this white paper proposes three layers of defense for deploying AI systems:

- **Attack mitigation:** Design defense mechanisms for known attacks.
- **Model security:** Enhance model robustness by various mechanisms such as model verification.
- **Architecture security:** Build a secure architecture with multiple security mechanisms to ensure business security.

There is a long way to go before the industry achieves secure and robust AI systems. On the technology side, we need to continuously research AI explainability to understand how to implement a sound AI foundation and build systematic defense mechanisms for secure AI platforms. In terms of business, we need to analyze the business models in detail, and deploy tested and verified AI security technologies. In the intelligent world, where all things can sense, be connected, and be intelligent, Huawei will cooperate with our customers and partners to provide secure and trustworthy AI services that evolve alongside their needs.



Contents

1. Striving Toward an Intelligent World 02

2. Five Challenges to AI Security 03

3. Typical AI Security Attacks 04

3.1 Evasion 04

3.2 Poisoning 05

3.3 Backdoor 05

3.4 Model Extraction 05

4. AI Security Layered Defense 06

4.1 Defense Against AI Security Attacks 06

4.2 AI Model Security 08

4.3 Security Architecture of AI Services 10

5. Collaborating for a Safe and Smart Future 12

References 13

01 Striving Toward an Intelligent World

With the accumulation of big data, dramatic improvements in computing power and continuous innovation in Machine Learning (ML) methods, Artificial Intelligence (AI) technologies such as image recognition, voice recognition, and natural language processing have become ubiquitous. More and more companies will increase their investment in AI research and deploy AI in their products. According to Huawei's Global Industry Vision (GIV), by 2025, 100 billion connections will be achieved globally, covering 77 percent of the population; 85 percent of enterprise applications will be deployed on the cloud; smart home robots will enter 12 percent of all families, forming a billion-dollar market.

The development and extensive commercial use of AI technologies have heralded the coming of an intelligent world. McCarthy, Minsky, Shannon, et al., proposed the concept of "Artificial Intelligence" in 1956. 60 years later, AI technologies are blossoming with AlphaGo developed by Google DeepMind defeating the human Go champion. Today, with development of chips and sensors, the intention to "Activate Intelligence" is affecting a broad range of sectors, including:

- Intelligent transportation — choose the best route
- Intelligent healthcare — understand your health
- Intelligent manufacturing — adapt to your needs

Researchers at the University of California Berkeley believe that the rapid rise of AI in the past two decades can be attributed to the following reasons [1]: (1) Massive amounts of data: With the rise of the Internet, data grows rapidly in the form of voice, video, and text. This data provides sufficient input for ML algorithms, prompting the rapid development of AI technologies. (2) Scalable computer and software systems: The success of deep learning in recent years is mainly driven by new dedicated hardware, such as CPU clusters, GPUs, and Tensor Processing Units (TPUs), as well as software platforms. (3) The broad accessibility of these technologies: A large amount of open-source software now bolsters data processing and AI-related work, reducing development time and costs. In addition, many cloud services provide developers with readily available computing and storage resources.

In several types of deployments, such as robots, virtual assistants, autonomous driving, intelligent transportation, smart manufacturing, and Smart Cities, AI is moving toward historic accomplishments. Big companies such as Google, Microsoft, and Amazon have taken AI as their core strategy for future development. In 2017, Google DeepMind unveiled the AlphaGo Zero, which honed its Go-playing prowess simply by playing games against itself and defeated the champion-defeating version of AlphaGo after only three days of self-training. AlphaGo Zero is able to discover new knowledge and develop rule-breaking policies, revealing the tremendous potential of using AI technologies to change human life.

What we see now is only a start. In the future, we will have a fully connected smart world. AI will bring superior experience to people everywhere, positively affect our work and life, and boost economic prosperity.

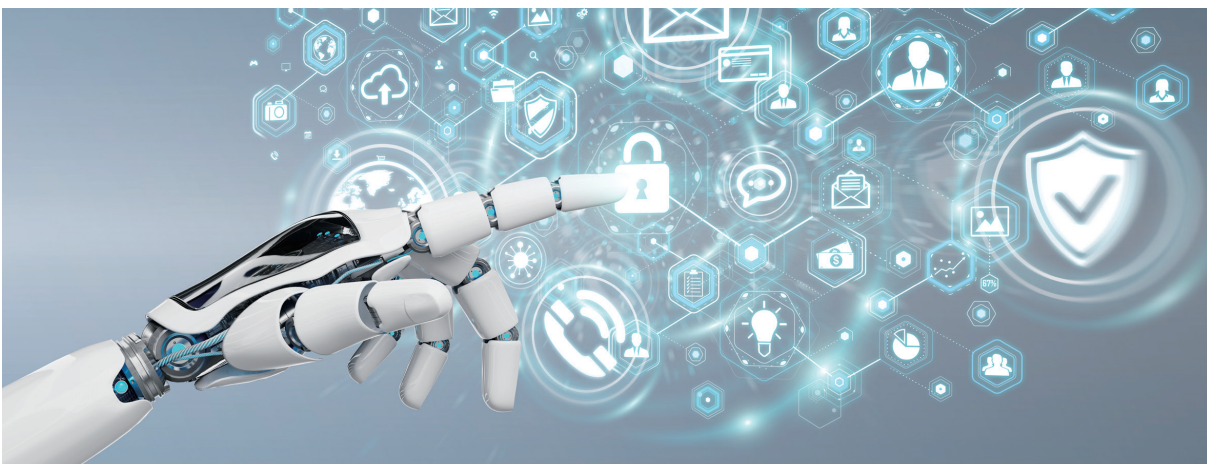
02 Five Challenges to AI Security

AI has great potential to build a better, smarter world, but at the same time faces severe security risks. Due to the lack of security consideration at the early development of AI algorithms, attackers are able to manipulate the inference results in ways that lead to misjudgment. In critical domains such as healthcare, transportation, and surveillance, security risks can be devastating. Successful attacks on AI systems can result in property loss or endanger personal safety.

AI security risks exist not only in theoretical analyses but also in AI deployments. For instance, attackers can craft files to bypass AI-based detection tools or add noise to smart home control command to invoke malicious applications. Attackers can also tamper with data returned by a terminal or deliberately engage in malicious dialogs with a chat robot to cause a prediction error in the backend AI system. It is even possible to apply small stickers on traffic signs or vehicles that cause false inferences by autonomous vehicles.

To mitigate these AI security risks, AI system design must overcome five security challenges:

- **Software and hardware security:** The code of applications, models, platforms, and chips may have vulnerabilities or backdoors that attackers can exploit. Further, attackers may implant backdoors in models to launch advanced attacks. Due to the inexplainsibility of AI models, the backdoors are difficult to discover.
- **Data integrity:** Attackers can inject malicious data in the training stage to affect the inference capability of AI models or add a small perturbation to input samples in the inference stage to change the inference result.
- **Model confidentiality:** Service providers generally want to provide only query services without exposing the training models. However, an attacker may create a clone model through a number of queries.
- **Model robustness:** Training samples typically do not cover all possible corner cases, resulting in the insufficiency of robustness. Therefore the model may fail to provide correct inference on adversarial examples.
- **Data privacy:** For scenarios in which users provide training data, attackers can repeatedly query a trained model to obtain users' private information.



03 Typical AI Security Attacks

3.1 Evasion

In an evasion attack, an attacker modifies input data so that the AI model cannot correctly identify the input. Evasion attacks are the most extensively studied attacks in academia. The following three types are representative.

Adversarial examples: Studies show that deep learning systems can be easily affected by well-crafted input samples, which are called adversarial examples. They are usually obtained by adding small perturbations into the original legitimate samples. These changes are not noticeable to human eyes but greatly affect the output of deep learning models. Szegedy [2] et al. first proposed adversarial examples in 2013. After that, scholars proposed other methods for generating adversarial examples. For example, the CW attack proposed by Carlini et al. can achieve a 100 percent attack success rate using small perturbations and successfully bypass most defense mechanisms.

Attacks in the physical world: In addition to perturbing digital image files, Eykholt [3] et al. describe modifying traffic signs to mislead an AI traffic sign recognition algorithm into identifying a "No Entry" sign as "45 km Speed Limit." In contrast to adversarial examples in the digital world, the physical world examples need to be scaling, cropping, rotation, and noise resistant.

Transferability and black-box attacks: To generate adversarial examples, attackers need to obtain AI model parameters, but these parameters are difficult to obtain in some scenarios. Papernot [4] et al. found that an adversarial example generated against a model can also deceive another model as long as the training data of the two models are the same. Attackers can exploit this transferability to launch black-box attacks without knowing AI model parameters. To do that, an attacker queries a model multiple times, uses the query results to train a "substitute model," and finally uses the substitute model to generate adversarial examples, which can be used to deceive the original model.



3.2 Poisoning

AI systems are usually retrained using new data collected after deployment to adapt to changes in input distribution. For example, an Intrusion Detection System (IDS) continuously collects samples on a network and re-trains models to detect new attacks. In a poisoning attack, the attacker may inject carefully crafted samples to contaminate the training data in a way that eventually impairs the normal functions of the AI system — for example, escaping AI security detection. Deep learning requires a large number of training samples, so it is difficult to guarantee the quality of samples.

Jagielski [5] et al. found that mixing a small number of adversarial examples with training samples can significantly affect the accuracy of AI models. The authors put forward three kinds of poisoning attacks: optimal gradient attack, global optimum attack, and statistical optimization attack. Demonstrations of these poisoning attacks targeted health care, loan, and house pricing data sets to affect the inference of AI models on new samples — affecting dosage, loan size/interest rate, and house sale price predictions, respectively. By adding 8 percent of malicious training data, attackers can cause a 75-percent change of the dosages suggested by models for half of the patients.

3.3 Backdoor

Like traditional programs, AI models can be embedded with backdoors. Only the person who makes backdoors knows how to trigger them; other people do not know of their existence nor can trigger them. Unlike traditional programs, a neural network model only consists of a set of parameters, without source code. Therefore, backdoors in AI models are harder to detect than in traditional programs. These backdoors are typically implanted by adding some specific neuron into the neural network model. A model with a backdoor responds in the same way as the original model on normal input, but on a specific input, the responses are controlled by the backdoor. Gu [6] et al. proposed a method of embedding backdoors in AI models. The backdoors can be triggered only when an input image contains a specific pattern and, from the model, it is difficult to find the pattern or even to detect whether such a backdoor exists. Most of these attacks occur during the generation or transmission of the models.

3.4 Model Extraction

In a model or training data extraction attack, an attacker analyzes the input, output, and other external information of a system to speculate on the parameters or training data of the model. Similar to the concept of Software-as-a-Service (SaaS) proposed by cloud service providers, AI-as-a-Service (AlaaS) is proposed by AI service providers to provide the services of model training, inference, etc. These services are open, and users can use open APIs to perform image and voice recognition. Tramèr [7] et al. developed an attack in which an attacker invoked AlaaS APIs multiple times to steal AI models. These attacks create two problems. The first is the theft of intellectual property. Sample collection and model training require a lot of resources, so trained models are important intellectual property. The second problem is the black-box evasion attack mentioned earlier. Attackers can craft adversarial examples using extracted models.

04 AI Security Layered Defense

As illustrated in Figure 4-1, three layers of defense are needed for deploying AI systems in service scenarios:

- **Attack mitigation:** Design defense mechanisms for known attacks.
- **Model security:** Enhance model robustness by various mechanisms such as model verification.
- **Architecture security:** Build a secure architecture with multiple security mechanisms to ensure business security.

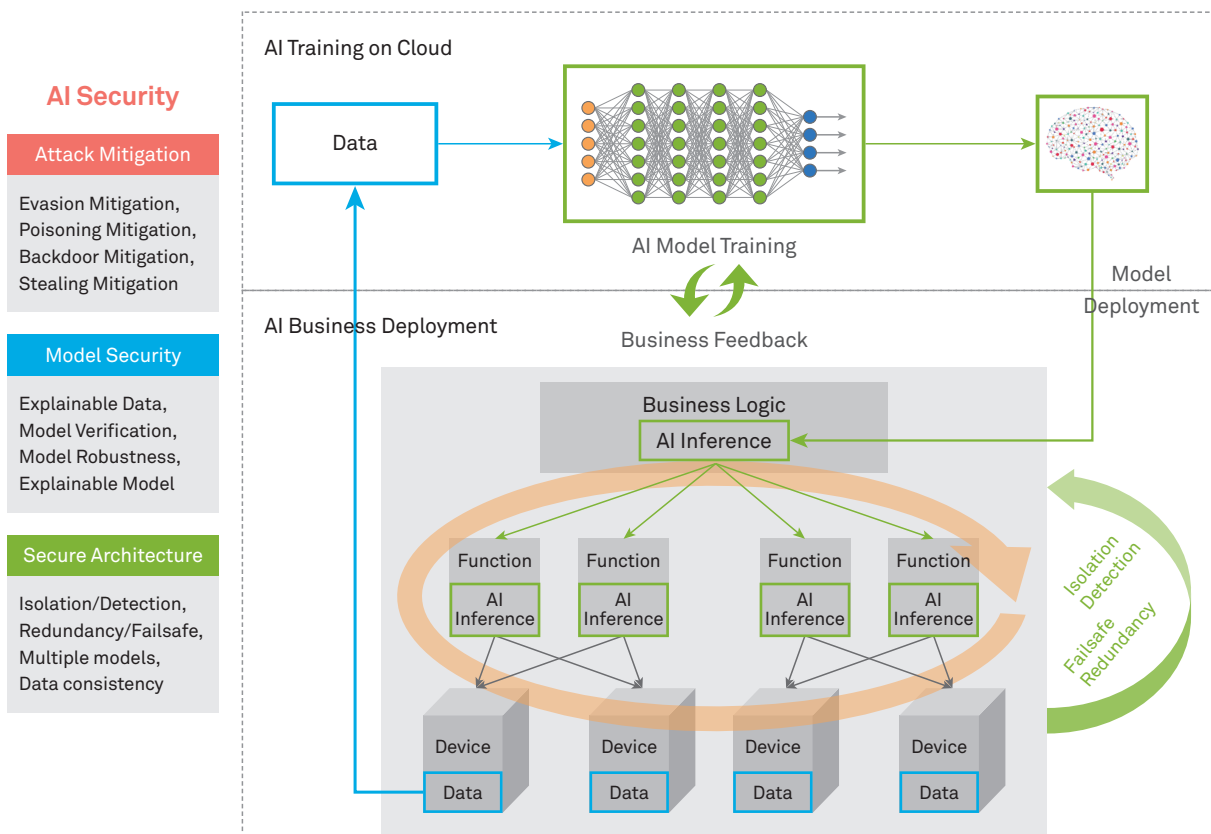


Figure 4-1 AI security defense architecture

4.1 Defense Against AI Security Attacks

Many countermeasures have been put forward in literature to mitigate potential attacks. Figure 4-2 shows the various defense technologies used by AI systems during data collection, model training, and model inference.

	Data Collection	Model Training	Model Inference
Evasion	Adversarial Samples	Network Distillation Adversarial Training	Adversarial Detection Input Reconstruction DNN Model Verification
Poisoning	Data Filtering Regression Analysis	Ensemble Analysis	
Backdoor		Model Pruning	Input Pre-processing
Stealing	Differential Privacy	PATE Model Watermarking	

Figure 4-2 AI security defense technologies

Defense technologies for evasion attack:

- 1. Network distillation:** These technologies work by concatenating multiple DNNs in the model training stage, so that the classification result generated by one DNN is used for training the next DNN. Researchers found that the transfer of knowledge can reduce the sensitivity of an AI model to small perturbations and improve the robustness of the model [8]. Therefore, they proposed using network distillation technologies to defend against evasion attacks. Using tests on MNIST and CIFAR-10 data sets, they found that distillation technologies can reduce the success rate of specific attacks (such as Jacobian-based Saliency Map Attacks).
- 2. Adversarial training:** This technology works by generating adversarial examples using known attack methods in the model training stage, adding the generated adversarial examples to the training set, and performing retraining one or multiple times to generate a new model resistant to attack perturbations. This technology enhances not only the robustness but also the accuracy and standardization of the new model, since the combination of various adversarial examples increases the training set data.
- 3. Adversarial example detection:** This technology identifies adversarial examples by adding an external detection model or a detection component of the original model in the model inference stage. Before an input sample arrives at the original model, the detection model determines whether the sample is adversarial. Alternatively, the detection model may extract related information at each layer of the original model to perform detection based on the extracted information. Detection models may identify adversarial examples based on different criteria. For example, the deterministic differences between input samples and normal data can be used as a criterion; and the distribution characteristics of adversarial examples and history of input samples can be used as the basis for identifying adversarial examples.
- 4. Input reconstruction:** This technology works by deforming input samples in the model inference stage to defend against evasion attacks. The deformed input does not affect the normal classification function of models. Input reconstruction can be implemented by adding noise, de-noising, or using an automatic encoder (autoencoder) [9] to change an input sample.
- 5. DNN verification:** Similar to software verification analysis technologies, the DNN verification technology uses a solver to verify attributes of a DNN model. For example, a solver can verify that no adversarial examples exist within a specific perturbation range. However, DNN verification is an NP-complete problem, and the verification efficiency of the solver is low. The efficiency of DNN verification can be improved by tradeoffs and optimizations such as prioritizing model nodes in verification, sharing verification information, and performing region-based verification.

These defense technologies apply only to specific scenarios and cannot completely defend against all adversarial examples. In addition to these technologies, model robustness enhancement can be performed to improve the resistance to input perturbations while keeping model functions to defend against evasion attacks. Multiple defense technologies can be combined in parallel or serially to defend against evasion attacks.

Defense technologies for poisoning attack:

- 1. Training data filtering:** This technology focuses on the control of training data sets and implements detection and purification to prevent poisoning attacks from affecting models. This method can be implemented by identifying possible poisoned data points based on label characteristics and filtering those points out during retraining, or comparing models to minimize sample data exploitable in poisoning attacks and filtering that data out [10].
- 2. Regression analysis:** Based on statistical methods, this technology detects noise and abnormal values in data sets. This method can be implemented in multiple ways. For example, different loss functions can be defined for a model to check abnormal values, or the distribution characteristics of data can be used.
- 3. Ensemble analysis:** This technology emphasizes the use of multiple sub-models to improve an ML system's ability to defend against poisoning attacks. When the system comprises multiple independent models that use different training data sets, the probability of the system being affected by poisoning attacks is reduced.

An ML system's overall ability to defend against poisoning attacks can be further enhanced by controlling the collection of training data, filtering data, and periodically retraining and updating models.

Defense technologies for backdoor attack:

- 1. Input pre-processing:** This technology aims to filter out inputs that can trigger backdoors to minimize the risk of triggering backdoors and changing model inference results [11].
- 2. Model pruning:** This technology prunes off neurons of the original model while keeping normal functions to reduce the possibility that backdoor neurons work. Neurons constituting a backdoor can be removed using fine-grained pruning [12] to prevent backdoor attacks.

Defense technologies for model/data stealing:

- 1. Private Aggregation of Teacher Ensembles (PATE):** This technology works by segmenting training data into multiple sets in the model training stage, each for training an independent DNN model. The independent DNN models are then used to jointly train a student model by voting [13]. This technology ensures that the inference of the student model does not reveal the information of a particular training data set, thus ensuring the privacy of the training data.
- 2. Differentially private protection:** This technology adds noise to data or models by means of differential privacy in the model training stage. For example, some scholars propose a method [14] for generating gradients by means of differential privacy to protect the privacy of model data.
- 3. Model watermarking:** This technology embeds special recognition neurons into the original model in the model training stage. These neurons enable a special input sample to check whether some other model was obtained by stealing the original model.

4.2 AI Model Security

As described above, adversarial ML exists extensively. Evasion attacks, poisoning attacks, and all kinds of methods that take advantage of vulnerabilities and backdoors are not only accurate, but also have strong transferability,

leading to high risks of misjudgment by AI models. Thus, in addition to defense against known attacks, the security of an AI model itself must be enhanced to avoid the damage caused by other potential attacks. Figure 4-3 illustrates some considerations on the enhancements.

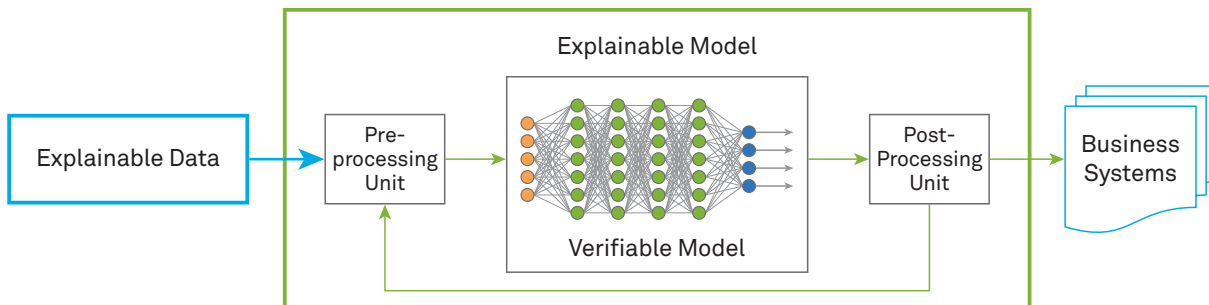


Figure 4-3 Model security analysis

Model detectability: Like traditional program analysis in software engineering, AI models could also be checked with some adversarial detection technologies such as black-box and white-box testing methods to guarantee some degree of security. However, existing test tools are generally based on open data sets having limited samples and cannot cover many cases in realistic deployments. Moreover, adversarial training technologies would cause high performance overhead due to re-training. Therefore, when AI models are being deployed in any system, a large number of security tests need to be performed on DNN models. For instance, a pre-processing unit could be used to filter out malicious samples prior to feeding into the training model, or a post-processing unit could be added to check the integrity of the model output to further reduce false positives. With these methodologies, we could possibly enhance the robustness of AI systems before deployment.

Model verifiability: DNN models work surprisingly better than traditional ML techniques (for example, providing higher classification rates and lower false positive rates). Thus, DNN models are widely used in image and voice recognition applications. However, caution needs to be taken when applying AI models in security and safety-sensitive applications such as autonomous driving and medical auto-diagnosis. Certified verification of DNN models can ensure security to some extent. Model verification generally requires restricting the mapping between the input space and output space to determine whether the output is within a certain range. However, since statistical optimization-based learning and verification methods typically cannot traverse all data distributions, corner cases like adversarial samples would still exist. In this situation, it is relatively difficult to implement specific protection measures in actual deployments. Principled defense can only be resolved when the fundamental working principle of DNN models is fully understood.

Model explainability: At present, most AI models are considered to be complicated black-box systems whose decision-making process, justification logic, and inference basis are hard to fully interpret. In some applications, such as playing chess and machine translation, we need to understand why machines make this or that decision to ensure better interaction between humans and machines. Nevertheless, the inexplainability of AI systems does not bring big problems in these applications. A translation machine can continue to be a complete complex black-box system as long as it provides good translation results, even though it does not explain why it translates one word into another. However, in some use cases, inexplainability tends to bring legal or business risks. For example, in insurance and loan analysis systems, if an AI system cannot provide the basis for its analysis results, it may be criticized as being discriminative; in healthcare systems, to accurately perform further processing based on AI analysis results, the basis of AI inference needs to be known. For instance, we hope that an AI system can analyze whether a patient has cancer, and the AI system needs to show how it draws the conclusion. Moreover, it is impossible to effectively design a secure model when its working principle is unknown. Enhancing the explainability of AI systems helps in analyzing their logic vulnerabilities or blind spots of data, thereby improving the security of the AI systems.

In the literature, researchers are actively exploring the explainability of AI models. For example, Strobel et al. put forward the visualization analysis of hidden activation functions [15]. Morcos et al. proposed the use of statistical analysis methods to discover semantic neurons [16]. Selvaraju et al. offer saliency detection for image classification [17]. Model explainability can also be implemented in three phases:

- 1. "Explainable data" before modeling:** Because models are trained using data, efforts to explain the behavior of a model can start by analyzing the data used to train the model. If a few representative characteristics can be found from the training data, required characteristics can be selected to build the model during training. With these meaningful characteristics, the input and output of the model can be explained.
- 2. Building an "explainable model":** One method is to supplement the AI structure by combining it with traditional ML. This combination can balance the effectiveness of the learning result and the explainability of the learning model and provide a framework for explainable learning. A common important theoretical foundation of traditional ML methods is statistics. This approach has been widely used and provides sound explainability in many computer fields, such as natural language processing, voice recognition, image recognition, information retrieval, and biological information recognition.
- 3. Explainability analysis of established models:** This approach seeks to analyze the dependencies between the input, output, and intermediate information of AI models to analyze and verify the models' logic. In the literature, there are both general model analysis methods applicable to multiple models, such as Local Interpretable Model-Agnostic Explanations (LIME) [18], and specific model analysis methods that can analyze the construction of a specific model in depth.

When an AI system is explainable, we can effectively verify and check it. By analyzing the logical relationship between modules of the AI system and input data, for example, we can confirm that the customer reimbursement capability is irrelevant to the gender and race of the customer. The explainability of an AI system ensures clearer logical relationships between input data and intermediate data; this is the other advantage of AI system explainability. We can identify illegitimate or attack data, or even fix or delete adversarial examples based on the self-consistency of data to improve model robustness.

The EU General Data Protection Regulation (GDPR) mandates that AI system decisions must not be based on user racial or ethnic origins, political positions, religious beliefs, etc. Explainable AI systems can ensure that the analysis results they produce comply with the GDPR requirement, preventing users from becoming victims of "algorithm discrimination." In most AI systems, prejudice does not often lie in the algorithm itself, but in the data provided to machines. If the inputs contain biased data — for example, when a company's HR department applies a subtle prejudice against female job seekers — the number of cases in which female job seekers are rejected will increase in the model, resulting in a gender imbalance. Even when gender is not an important characteristic of model training data, the data can amplify human prejudice in the AI model analysis conclusion. Governments usually need to verify the security, reliability, and explainability of AI-enabled systems. Only an explainable, verifiable, robust AI system can build confidence and trust in the public.

4.3 Security Architecture of AI Services

When developing AI systems, we must pay close attention to their potential security risks; strengthen prevention mechanisms and constraint conditions; minimize risks; and ensure AI's secure, reliable, and controllable development. When applying AI models, we must analyze and determine the risks in using AI models based on the characteristics and architecture of specific services, and design a robust AI security architecture and deployment solution using security mechanisms involving **isolation, detection, failsafe, and redundancy** (Figure 4-4).

In autonomous driving, if an AI system incorrectly decides on critical operations such as braking, turning, and

acceleration, the system may seriously endanger human life or property. Therefore, we must guarantee the security of AI systems for critical operations. Various security tests are surely important, but simulation cannot guarantee that AI systems will not go wrong in real scenarios. In many applications, it may be difficult to find an AI system that can give 100 percent correct answers every time. This underlying uncertainty makes the security design of the system architecture more important. The system must enable fallback to manual operation or other secure status when unable to make a deterministic decision. For example, if an AI-assisted medical system cannot provide a definite answer about the required medicine and dosage or detects possible attacks, it is better for the system to reply "Consult the doctor" than to provide an inaccurate prediction that may endanger patient health. For safety's sake, correct use of the following security mechanisms based on business requirements are essential to ensure AI business security:

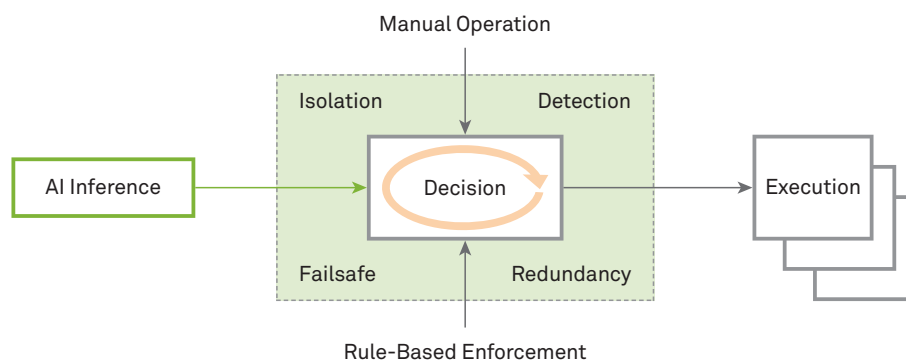


Figure 4-4 Security architecture for AI with business decision making

- 1. Isolation:** To ensure stable operation, an AI system analyzes and identifies the optimal solution and sends it to the control system for verification and implementation. Generally, the security architecture must isolate functional modules and setup access control mechanisms between modules. The isolation of AI models can reduce the attack surface for AI inference, while the isolation of the integrated decision module can reduce attacks on the decision module. The output of AI inference can be imported into the integrated decision module as an auxiliary decision-making suggestion, and only authorized suggestions can enter the decision module.
- 2. Detection:** Adopting continuous monitoring and with an attack-detection model in the main system architecture, it is possible to comprehensively analyze the network security status and estimate the current risk level. When the risk is high, the integrated decision system can reject the suggestion coming from the automatic system and hand over control to a person to ensure security under attacks.
- 3. Failsafe:** When a system needs to conduct critical operations such as AI-assisted autonomous driving or medical surgery, a multi-level security architecture is required to ensure the entire system security. The certainty of the inference results provided by the AI system must be analyzed. When the certainty of the result is lower than a certain threshold, the system falls back to conventional rule-based technologies or manual processing.
- 4. Redundancy:** Many business decisions and data are associated with each other. A feasible method to ensure the security of AI models is to analyze whether the association has been ruined. A multi-model architecture can be set up for critical applications, so that a mistake in one model does not keep the system from reaching a valid decision. In addition, the multi-model architecture can largely reduce the possibility of the system being fully compromised by a single attack, thereby improving the robustness of the entire system.

Amodei et al. further describe a number of security challenges that AI systems may encounter during deployment [19]. For example, the authors describe how to avoid potential negative side effects during task execution and possible reward hacking during objective fulfillment, as well as safe exploration of AI systems. Fundamental research should be conducted on making AI systems more secure in future deployments.

05 Collaborating for a Safe and Smart Future

Various AI sub-disciplines, such as computer vision, voice recognition, natural language processing, cognition and reasoning, and game theory, are still in the early stages of development. Deep learning systems that leverage big data for statistical analysis have expanded the boundaries of AI, but they are also generally considered as "lacking common sense," which is the biggest obstacle to current AI research. AI could be both data driven and knowledge driven. The breakthrough of the next-generation AI lies in knowledge inference.

Whatever the next steps, the large-scale popularization and advancement of AI require strong security assurance. This document focuses on two types of AI attacks and defenses. First, affecting the correctness of AI decisions: An attacker can sabotage or control an AI system or intentionally change the input so that the system makes a decision desired by the attacker. Second, attackers may steal the confidential data used to train an AI system or extract the AI model. With these problems in mind, AI system security should be dealt with from three aspects: AI attack mitigation, AI model security, and AI architecture security. In addition, the transparency and explainability of AI are the foundation of security. An AI system that is not transparent or explainable cannot undertake critical tasks involving personal safety and public safety.

AI also requires security investigations in a wide range of domains, such as laws and regulations, ethics, and social supervision. On September 1, 2016, the One Hundred Year Study on Artificial Intelligence (AI100) project of Stanford University released a research report titled "AI and Life in 2030"[20], pointing out the profound changes brought by AI technologies and calling for regulatory activities that are more appropriate and that "will not stifle innovation." In the coming years, since AI will be deployed in the fields of transportation and medical care, the technology "must be introduced in ways that build trust and understanding, and respect human and civil rights." In addition, "policies and processes should address ethical, privacy, and security implications." To this end, international communities should cooperate to drive AI to evolve toward a direction that benefits mankind.





References

- [1] I. Stoica, D. Song, R. A. Popa, D. Patterson, M. W. Mahoney, R. Katz, A. D. Joseph, M. Jordan, J. M. Hellerstein, J. Gonzalez, K. Goldberg, A. Ghodsi, D. Culler and P. Abbeel, "A Berkeley View of Systems Challenges for AI," University of California, Berkeley, Technical Report No. UCB/Eecs-2017-159, 2017.
- [2] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow and R. Fergus, "Intriguing properties of neural networks," arXiv preprint arXiv:1312.6199, 2013.
- [3] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno and D. Song, "Robust physical-world attacks on deep learning models," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [4] N. Papernot, P. McDaniel and I. Goodfellow, "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples," arXiv preprint arXiv:1605.07277, 2016.
- [5] M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru and B. Li, "Manipulating machine learning: Poisoning attacks and countermeasures for regression learning," in *IEEE Symposium on Security and Privacy (S&P)*, 2018.
- [6] T. Gu, B. Dolan-Gavitt and S. Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain," in *NIPS MLSec Workshop*, 2017.
- [7] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter and T. Ristenpart, "Stealing Machine Learning Models via Prediction APIs," in *USENIX Security Symposium*, 2016.
- [8] N. Papernot, P. McDaniel, X. Wu, S. Jha and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *IEEE Symposium on Security and Privacy (S&P)*, 2016.
- [9] S. Gu and L. Rigazio, "Towards deep neural network architectures robust to adversarial examples," in *International Conference on Learning Representations (ICLR)*, 2015.
- [10] R. Laishram and V. Phoha, "Curie: A method for protecting SVM classifier from poisoning attack," arXiv preprint arXiv:1606.01584, 2016.
- [11] Y. Liu, X. Yang and S. Ankur, "Neural trojans," in *International Conference on Computer Design (ICCD)*, 2017.
- [12] K. Liu, D.-G. Brendan and G. Siddharth, "Fine-Pruning: Defending Against Backdooring Attacks on Deep Neural Networks," arXiv preprint arXiv:1805.12185, 2018.
- [13] N. Papernot, A. Martín, E. Ulfar, G. Ian and T. Kunal, "Semi-supervised knowledge transfer for deep learning from private training data," arXiv preprint arXiv:1610.05755, 2016.
- [14] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar and L. Zhang, "Deep learning with differential privacy," ACM SIGSAC Conference on Computer and Communications Security, 2016.
- [15] H. Strobelt, S. Gehrmann, H. Pfister and A. M. R. Lstmvis, "A tool for visual analysis of hidden state dynamics in recurrent neural networks," *IEEE transactions on visualization and computer graphics*, vol. 24, no. 1, pp. 667-676, 2018.
- [16] A. S. Morcos, D. G. Barrett, N. C. Rabinowitz and M. Botvinick, "On the importance of single directions for generalization," arXiv preprint arXiv:1803.06959, 2018.
- [17] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," arXiv preprint arXiv:1610.02391, 2016.
- [18] M. T. Ribeiro, S. Singh and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in *ACM international conference on knowledge discovery and data mining (KDD)*, 2016.
- [19] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman and D. Mané, "Concrete Problems in AI Safety," arXiv preprint arXiv:1606.06565, 2016.
- [20] S. Peter, B. Rodney, B. Erik, C. Ryan, E. Oren, H. Greg, H. Julia, K. Shivaram, K. Ece, K. Sarit, L.-B. Kevin, P. David, P. William, S. AnnaLee, S. Julie and etc, Artificial Intelligence and Life in 2030, One Hundred Year Study on Artificial Intelligence: Report of the 2015–2016 Study Panel, Stanford University, 2016.

Copyright © Huawei Technologies Co., Ltd. 2018. All rights reserved.

No part of this document may be reproduced or transmitted in any form or by any means without prior written consent of Huawei Technologies Co., Ltd.

Trademark Notice

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

General Disclaimer

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.

HUAWEI TECHNOLOGIES CO., LTD.

Huawei Industrial Base
Bantian Longgang
Shenzhen 518129, P.R. China
Tel: +86-755-28780808

www.huawei.com