United Nations Educational, Scientific and Cultural Organization
联合国教育、科学及文化组织

IKCEST
International Knowledge Centre for Engineering Sciences and Technology under the Auspices of UNESCO
国际工程科技知识中心 由教科文组织支持

Silk Road Sciences and Technology Knowledge Service
丝路科技知识服务

# Current Situation and Countermeasures of Artificial Intelligence (AI) Security Problems：A Research Report

**IKCEST Silk Road Science and Technology Sub-platform Team**

2022.04

IKCEST Sci-Tech Trend Report

http://ikcest.org/tidings_article-304900.htm

# Current Situation and Countermeasures of Artificial Intelligence (AI) Security Problems: A Research Report

The Silk Road Branch of IKCEST (International Knowledge Centre for Engineering Sciences and Technology under the Auspices of UNESCO)

2021.12

# Table of Contents

Artificial Intelligence (AI) leads a new round of scientific and technological revolution and industrial transformation, playing as the impetus of accelerating economic and social development. However, in recent years, traffic accidents caused by automatic driving failure, deep forgery technology used for counterfeiting, the discrimination posed by intelligent recommendation system against skin color and gender, etc., show that AI is facing a series of safety issues involving in technical vulnerability, malicious application and social ethics. General Secretary Xi Jinping pointed out clearly at the Ninth Group Study Session of the Political Bureau of the 19th CPC Central Committee that "multidisciplinary forces should be integrated to strengthen research on legal, ethical and social issues related to AI, laws, regulations, institutional systems and ethics shall be established and improved to ensure the healthy development of AI." Safety, especially ethics, is an important part of AI technology research and industrial application, and we shall spare no efforts to seize the major strategic opportunities presented by the development of AI, to speed up the pace in building China into an innovative and strong country in science and technology, and to ensure that scientific and technological advances benefit all the people.

## I. AI Security Issues

AI security has been concerned by the government, the military and industries. The existing research mainly analyzes the security threats from the imperfection and improper use of AI technology. Nevertheless, AI has been gradually deeply integrating into human production and daily life, and the ethical safety of AI turns out to be its unique security problem.

### (I) Security Issues Caused by Vulnerable AI Technology

AI relies on various algorithms to learn and to make decisions. The imperfection and incorrectness of algorithm design and implementation will directly affect its security. Taking anti-sample attack as an example, machine learning algorithm can make wrong prediction results by adding a small disturbance to normal sample which is hard to be detected by human eyes. For example, as for image classification task, **only a few disturbed pixel, difficult to be detected by naked eyes, is enough for deep neural network in yielding possible false judgments [1].** Another example is backdoor attack, it refers to inserting specific neurons into a neural network to generate models with backdoor; when the input sample has a specific trigger, the

model is available to output the target category specified by the attacker. For instance, by adding a specific pattern to a right-turn traffic sign, a backdoor can be triggered to predict it as a left-turn sign [2].

AI is contributed by data mining and training; as a result, data security and privacy are basic requirements of AI. Setting data poisoning as an example, it means that carefully constructed abnormal data is added to training data, which destroys the probability distribution of original training data and leads to decision errors in the AI model. The purpose of data poisoning is to train AI systems that are insensitive to abnormal data, such as online recommendation systems and spam detection systems. Taking model and private data theft as an example, AI model involves with algorithm structure and parameters, which are important intellectual property rights. Direct theft refers to copying or modifying model files; while indirect theft refers to attacker speculates model parameters and training data information by querying and analyzing the input and output of AI system and other information [6].

AI's developing and operating is supported by various software platforms, and the security risks of third parties will directly affect the normal services of AI. In terms of **software framework**, the existing AI platforms are all consist of open source deep learning frameworks and components, which have not undergone strict security evaluation and may have vulnerabilities or even backdoor. These open source frameworks are built on numerous basic libraries and components, and the security vulnerability of any component will threaten the application system supported by the whole framework [7]. On the other hand, in accordance with hardware facilities, AI platform physical devices are composed of GPU/CPU servers and other infrastructure. Once an attacker has access to these corresponding hardware devices, he can forge and steal data, and then even destroy the integrity of the entire system; or he also is able to apply the platform's computing power for "illegal mining" that severely impacted normal business and service.

### (II) Security Issues by Abusing AI Technology

Misuse of AI technology will also pose serious security threats. AI technology can be directly used in cyber-attacks as it can greatly automate the writing and distribution of malware, so that **attackers produce samples of malware that bypass detection systems by making use of GAN [8]**. Using botnets, attackers communicate and exchange with each other on the network, allowing controlled objects to execute commands autonomously when no instructions are required, and

automatically attack multiple targets.

For social security, images, audio and video that look like the real things could be created by deep forgery, and then further been used for online frauds. In Zhejiang, Hubei and other provinces of China, there are many criminal cases of using speech synthesis technology to pretend as the relatives of victims to carry out frauds, resulting in huge economic losses and bad social influences. Recommended information content based on users' online behavior data spoils users by intelligent recommendation algorithms in modelling users' preferences. The abuse of such technologies will bring serious threats to political and social stability.

In terms of national security, AI can be used to influence public political ideology and to threaten national political security. Foreign technology companies analyze the political orientation of users on social media and accurately push campaign news, so as to intervene and guide voters to vote. It has been proved to affect major political events such as the US presidential election and the Brexit referendum.

When it comes to military security, AI weapons with high efficiency, precision and lethality are the key technologies for all the countries, exacerbating a new round of arms race.

### (III) Ethical Problems of AI

AI technology has caused a huge impact on the existing social ethics and moral system, mainly for security and control, fairness and justice, privacy protection, rights and responsibilities, and mental health [12].

Security and control are the biggest ethical problems for AI technology. Since AI is featured by weak robustness, uninterpretability and vulnerability, it brings great security risks to use AI technology and threatens the security of human beings. For example, due to the lack of mature self-driving technology, Tesla and other brands have suffered major traffic accidents occasionally. In addition, when the decision analysis ability of intelligent agent exceeds that of human, how to ensure the ultimate control of human to the system is very important. For example, the Ethiopian airlines Boeing 737MAX crash in 2019, the autopilot system deprived the pilot of the right to control the plane.

According to McKinsey, AI technology will create global economic imbalances and increase the wealth disparity between countries, companies and individuals. Technologically advanced countries can use their advanced AI technology to further

improve social productivity and widen the national power gap with other countries. The company that takes the lead in AI technology can gain a larger market share, or even a monopoly position. And individuals engaged in labor-intensive work are likely to encounter wage reduction and unemployment in the future [11], and even social unemployment crisis will be triggered.

The successful application of AI technology is inseparable from the accumulation of big data, which leads to data leakage and abuse. Data security is closely related to personal, corporate and even national security. The disclosure of personal data such as physiological characteristics will lead to the infringement of personal privacy, the disclosure of corporate data will bring serious legal consequences and huge economic losses, and the disclosure of national infrastructure data such as road network and power grid will directly threaten national military security.

AI intellectual property touches on algorithm, data and creation. At present, there has been great progress in the patent protection of AI algorithm and open protection of AI data, but the research on how to intellectualize AI works is still in the early stage. As for legal responsibility, the legal subject of AI is vague and the definition of AI responsibility is complex, which is a difficult problem for the effective supervision of laws and regulations involved in AI.

People attach AI with severe emotional dependence, which shocks the existing concept of interpersonal relationship. The companionship, consideration and obedience of intelligent robots greatly satisfy human beings psychologically; as a result, people are unwilling to have normal heterosexual communication, or even unwilling to set up families, and ultimately harming the benign development of the whole society. AI also creates severe mental dependency in humans since it further liberates productivity, helps human beings to analyze and make decisions, leading to the stagnation of human intelligence.

## II. Coping Strategies for AI Security Issues

Enhancing algorithm security, protecting data privacy AI platforms so as to solve traditional AI security problems; however, the ethical issues of AI are more dependent on the relevant security systems.

### (I) Coping Strategies for Traditional AI Security Problems

#### 1. Algorithm Security Enhancement of AI

The security enhancement of AI algorithm can be improved from abnormal sample detection, algorithm ability improvement and malicious behavior detection. Among them, abnormal sample detection refers to detecting deep forged sample and adversarial sample; the improvement of algorithm capability indicates the improvement of generalization ability and interpretability; and malicious behavior includes maliciously implanted functional defects such as backdoor.

**Detecting adversarial samples**: To check whether the input data is an adversarial sample added with malicious disturbance, and to discover in time whether the application is under adversarial sample attack. One detection is based on data feature layer difference, which is achieved by modeling the feature difference between normal input data and adversarial samples [13]; the other detection is contributed by using models to predict differences in outcomes, observing the difference of prediction results between adversarial samples and normal data [14].

**Detecting deep forged samples**: Deep forgery detection is a technology to detect forged audio, image and video. At present, it is mainly based on data-driven learning forgery features to identify the authenticity of images by studying the representational difference between deep forgery content and real content. In order to improve the generalization ability of the algorithm, existing studies increase the diversity degree of training data and use Bayesian Learning Framework to estimate the uncertainty of new sample prediction.

**Enhancement of generalization ability**: Enhancing generalization ability is helpful to practice AI algorithm. First, data enhancement, using various data enhancement methods to increase the number and diversity of training sets. Second, robust feature learning enables the model to learn the rules hidden behind the data and to enhance model robustness. Third, model regularization, including model integration and other empirical regularization methods, as well as parametric regularization methods, alleviating the over-fitting problem of models.

**Interpretability of models**: To interpret the decision logic and working mechanism of AI algorithms in a way that humans can understand. Before modeling, the influence of different features on the final decision is analyzed to help understand the decision basis of the model. In modeling, build interpretable AI systems, such as combining neural networks with decision trees [15]. After modeling, the model logic is verified by analyzing the dependence between the input and output of the model and the intermediate information [16].

**Backdoor attach defense**: Backdoor attack defense can be implemented through network pruning and backdoor detecting. The principle of pruning is to properly cut off the neurons of the original model, and reduce the function possibility of the backdoor neurons when ensuring the consistency of normal functions [17]. Backdoor detection is based on the "Minimum Attack Cost" assumption [18] , that is, less modification is required for the misclassification of the attacked category in the model than for other categories that are not attacked.

### 2. Protecting AI Data Security

Data security protection mainly aims at data privacy security and consistency of data use process. Data privacy security includes privacy protection of data itself and model intellectual property rights. Consistency during data use mainly depends on data tracing.

**Protecting data privacy**: **Differential privacy** can avoid privacy disclosure caused by small changes in data sources, and overall data can be studied and analyzed without disclosing individual data privacy [19]. Federated learning refers to the process of data joint training and the establishment of shared machine learning model through parameter exchanging in encryption mechanism without data of all participants being exported locally [20].

**Data tracing**: Data tracing is to verify the correctness of data in use. The security label (encrypted meta information) of the data is hidden in original data to determine whether the data is tampered during use. Block chain technology can also be used to store data identification, collection source, time, provider, processing behavior and other data traceability information in the block chain, so as to trace every data processing behavior.

**Protecting model intellectual property**: By embedding watermark into the model file during training, the model is avoided from being stolen [21]. By adding a special task-independent watermark module behind the target model, a uniform and invisible watermark can be embedded in the output. When the attacker uses the model with watermark output training in order to substitute the model, the invisible watermark will be embedded into the alternative model to trace the use of the target model.

### 3. Protecting AI Platform Security

Software framework protection: It refers to vulnerability mining and model file

verification. **Vulnerability mining** uses static code audit technology to detect security vulnerabilities and code compliance, conducts fuzzy tests on code modules, and establishes a fast security response mechanism. **Model file verification** is to detect the security problems in model file before loading by checking the attribute information of the model file to prevent malicious algorithm model file from being loaded.

Hardware protection: It represents device encryption and device detection. **Device encryption** protects encrypted internal data from leakage [7]. **Device detection**: Necessary tests should be carried out on the devices used in AI to ensure that they are not be hijacked by attackers [7].

## (II) Coping Strategies for AI Ethical and Moral Problems

### 1. AI Ethics

Asilomar AI Principles, the most famous AI ethics, which is also a consensus of AI ethics in current academic circles. In January 2017, in Asilomar of California, the Beneficial AI Conference was held, nearly thousands of experts in the field of AI and robotics jointly signed the Asilomar AI 23 Principles, calling on the world to strictly abide by these principles while developing AI, and jointly safeguarding the ethics, interests and security of humanity in the future. Essentially, Asilomar AI 23 Principles are an extension of Isaac Asimov's "Three Laws of Robotic".

Major countries and organizations around the world are also accelerating the construction of AI ethical governance systems [22]. United States, the European Union, the United Kingdom and Japan have incorporated ethical governance into their AI strategies. China has set up a special committee on AI governance to further strengthen research on legal, ethical, standard and social issues related to AI, and participate in international exchanges and cooperation on AI governance.

In addition, countries and organizations worldwide have also issued ethical guidelines for AI to specifically guide the research and development of AI technologies. For example, the European Union has published *AI Ethics Guidelines*, listing seven principles for AI to be trusted, ensuring that AI applications are ethical and the technology is robust enough to maximize its benefits while minimize its risks. IEEE has published Ethically Aligned Design V2, which aims to design, develop and apply AI and autonomous technologies ethically.

### 2. AI Security Standards

AI security standards are important and necessary for industrial development, involving AI algorithm model, data, infrastructure, product and application related security standards.

Currently, AI security standards are mainly set by international organizations [3]. For example, ISO/IEC JTC1 established the AI Standardization Committee to develop standards such as AI credibility, robustness evaluation, algorithm bias and ethics. ITU-T has carried out the development of AI security-related standards, aiming to solve security problems in AI applications such as smart medicine, smart vehicles, junk information management and biometric recognition. At the same time, National Information Security Standardization Technical Committee of China has also formulated China's basic common standards for AI, as well as safety standards in special areas such as biometric identification, autonomous driving and data security.

### 3. Legal Specifications for AI Security

Compared with the rapid development of IT technology, its construction of laws and regulations is relatively lagging behind. For example, in May 2018, the European Union issued the *General Data Protection Specification (GDPR)*, requiring that AI algorithms should be interpretable and stipulating extremely strict conditions for legitimate applications of AI automated decision making. In 2019, US Senators proposed the federal *Algorithmic Accountability Act of 2019*, suggesting that the evaluation rules for "high-risk automated decision systems" be formulated as soon as possible. China has also recently issued *The Personal Information Protection Law* and *Data Security Law*, aiming at regulating data processing activities, ensuring data security, promoting data development and utilization, protecting the legitimate rights and interests of individuals and organizations, and safeguarding national sovereignty, security and development interests.

## III. Security Risks in AI Typical Applications

### (I) Automatic Driving

Automobile enterprises around the world are swarming for developing autonomous driving. And safety comes as the first principle in the autonomous driving industry, but in recent years, the frequent accidents of autonomous driving make its safety controversial.

(a) Tesla Autopilot Accident

(b) NIO Autopilot Accident

Fig. 1 Safety Risks of Automatic Driving

As Fig.1 (a) shows, in June 2020, a moving Tesla drove straight into a white truck that had flipped on its side on the highway. An investigation into the accident concluded that Tesla's autopilot system mistakenly recognized the white trunk of the truck as sky, so it crashed into it without slowing down. In Fig.1 (b), in August 2021, the driver of a NIO car had a traffic accident on a highway after the piloting assistance function was enabled, and died when he collided head-on with a construction vehicle.

**(II) AI Deepfake**

AI face-changing, also known as AI deepfake, is the most popular AI APP. However, the abuse of AI face-changing may bring personal loss of privacy, portrait rights or property, and may also cause social risks and political crises. In April 2018, a video of "Obama" criticizing Trump forged by deep forgery technology spread widely on Twitter. As shown in Fig. 2, the number of views reached millions within a few hours, causing an uproar in the American society. Moreover, deep forgery technology in pornographic videos, replacing the heroine in the video with some popular actresses, has seriously damaged the right of portrait, reputation and privacy of these women.

Fig.2 "Obama" was Faked by AI Face-changing

**(III)Big Data Collection**

In the era of big data, user privacy security faces great challenges. There are certain risks in data collection, storage, analysis, transaction and discarding, which expose personal user data to extremely insecure network space at any time and places, causing users to face unprecedented privacy and security threats. In July 2020, the Beijing Internet Court made a judgment on a case in which Tik Tok violated the privacy of a user. The court held that Tik Tok violated the user's personal information rights and interests by processing the user's personal information without the user's consent.


Fig. 3 Big Data Collection of Didi Chuxing

Big data collection by Internet companies may even threaten national security. As Fig.3 shows, Didi, as a mobility platform, has personal travel, traffic and street view data for all Chinese cities. After analysis and processing, these geographic data can be directly used for military purposes, posing a serious threat to national security.

## References:

[1] Su J, Vargas D V, Sakurai K. One pixel attack for fooling deep neural networks[J]. IEEE Transactions on Evolutionary Computation, 2019, 23(5): 828-841.

[2] Chen H, Fu C, Zhao J, et al. DeepInspect: A Black-box Trojan Detection and Mitigation Framework for Deep Neural Networks[C], IJCAI, 2019.

[3] Artificial Intelligence Security Standardization White Paper (2019 Edition), Task Force on Big Data Security Standards of National Information Security Standardization Technical Committee, 2019.

[4] Buolamwini J, Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification[C], FACCT, 2018.

[5] K. Eykholt, I. Evtimov, E. Fernandes, et al. Robust physicalworld attacks on deep learning models[C], CVPR, 2018.

[6] F. Tramèr, F. Zhang, A. Juels, et al. Stealing Machine Learning Models via Prediction APIs[C],

USENIX Security Symposium, 2016.
[7] White Paper on Artificial Intelligence Security (2020 Edition), Data Security and Privacy Protection Laboratory, Zhejiang University-Ant Group Fintech Research Center, 2020.
[8] Hu W, Tan Y. Generating adversarial malware examples for black-box attacks based on GAN[J]. arXiv preprint arXiv:1702.05983, 2017.
[9] Fang Binxing, My Opinions on Artificial Intelligence- AI and Its Security, Chinese Society for Artificial Intelligence Newsletter, vol. 10, No. 4, 2020.
[10] White Paper on Artificial Intelligence Security (2018 Edition), Security Research Institute of China Academy of Information and Communication Technology, 2018.
[11] How robots change the world[R], Oxford Economics, 2019.
[12] Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines[J]. Nature Machine Intelligence, 2019, 1(9): 389-399.
[13] Ma X, Li B, Wang Y, et al. Characterizing adversarial subspaces using local intrinsic dimensionality[C], ICLR, 2018.
[14] Xu W, Evans D, Qi Y. Feature squeezing: Detecting adversarial examples in deep neural networks[C], NDSS, 2018.
[15] Wan A, Dunlap L, Ho D, et al. NBDT: Neural-Backed Decision Tree[C], ICLR, 2020.
[16] Liu X, Wang X, Matwin S. Improving the interpretability of deep neural networks with knowledge distillation[C], ICDMW, 2018.
[17] K. Liu, D.-G. Brendan and G. Siddharth. Fine-Pruning: Defending Against Backdooring Attacks on Deep Neural Networks[J]. arXiv preprint arXiv:1805.12185, 2018.
[18] B. Wang, Y. Yao, S. Shan, et al. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks[C], S&P, 2019
[19] Kairouz P, Oh S, Viswanath P. The composition theorem for differential privacy[C], ICML, 2015.
[20] Yang Q, Liu Y, Chen T, et al. Federated machine learning: Concept and applications[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2019, 10(2): 1-19.
[21] Zhang J, Chen D, Liao J, et al. Model watermarking for image processing networks[C], AAAI, 2020.
[22] Report on Global Artificial Intelligence Industry Governance System 2020, Institute of Policy and Economics, China Academy of Information and Communication Technology, 2020.