**Add triple “security protection” to the cognitive domain**

Publisher: Qiu Yunshu

**This article was first published on Guangming Daily Client, by all-media reporter Cai Tongchen and correspondent Fang Zhongli**

Date: 2022-09-02

<https://ibd.fudan.edu.cn/f8/9f/c24057a456863/page.htm>

In recent years, military confrontation has gradually expanded from physical space to cognitive space. The cognitive domain composed of human spiritual and psychological activities is becoming a new battlefield and a new field for military struggle and strategic game. **In essence, the cognitive domain that people are talking about today is a new space of virtual and real combination jointly constructed by human intelligence and artificial intelligence systems. "Human, machine, and intelligence" coexist here, and the three spaces of "physical-information-cognition" are intertwined. In** the cognitive domain, artificial intelligence systems are the link and bridge. Without artificial intelligence systems, the cognitive domain may still be in a metaphysical and illusory state. As a man-made system, artificial intelligence systems are bound to have vulnerabilities and inherent security issues. In the cognitive domain confrontation, how to build a new paradigm of endogenous security of artificial intelligence? How to build available, reliable, trustworthy and more sustainable technology? How to establish a security barrier for cognitive space? To this end, the reporter interviewed Chen Ping, a researcher at the Big Data Research Institute of Fudan University, to discuss the issue of cognitive domain security.

**Artificial intelligence has inherent security issues**

At present, countries around the world attach great importance to the development of artificial intelligence technology and are promoting the large-scale application of artificial intelligence. At the same time, the security risks of artificial intelligence are becoming increasingly prominent. Researcher Chen Ping pointed out that artificial intelligence technology has endogenous security problems, that is, endogenous security risks caused by the incompleteness of its basic theoretical system. For example, just like the "two dark clouds" that appeared in physics in the 19th century, artificial intelligence technology also has "two dark clouds".



The picture shows the artificial intelligence confrontation demonstration system based on the autonomous driving platform displayed at the Zhengzhou Cyber ​​Security Science and Technology Museum

(Photo provided by Wang Danyu)

**First, there are security risks caused by the unexplainability of artificial intelligence.** The machine learning algorithm that artificial intelligence technology relies on is essentially to train the learning model to judge information by a large amount of given data. If the data is classified in a "following the same pattern", it may "draw a tiger but not a dog" if it is not careful, and end up with a "diao shi shao shao shao shao". For example, an attacker can mark a bunch of sloth photos as cats. This small group of "toxic" data samples is likely to induce the system to show sloth photos when it is asked to show kittens. According to relevant materials, this "data poisoning" attack method has become easier and easier. Researchers only need to poison less than 0.7% of the data submitted to the machine learning system to make the artificial intelligence system make wrong judgments. This means that only a few malicious samples are needed to make artificial intelligence make misjudgments. At the same time, more and more mature artificial intelligence technologies are being used to create false information, deep fake videos, audio, images, and texts. According to a report by RAND Corporation, a new 0 - day risk of deep fake false information is emerging. Attackers create deep fake content that can bypass detection by developing a customized generative model, making all current detection methods vulnerable. Data poisoning and deep fakes are problems that will seriously interfere with people's individual cognition, and may even manipulate social cognition and affect group cognition, thereby posing new threats to national security.

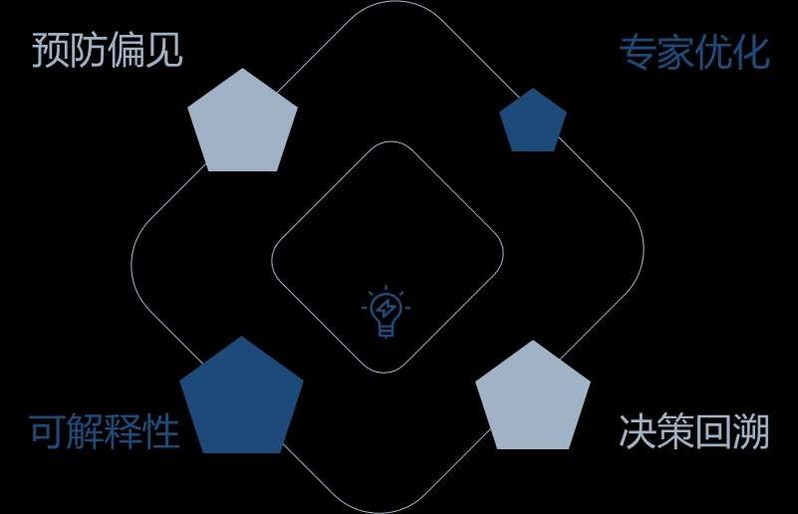
**Another problem is the backdoor problem of artificial intelligence models.** Big data, big computing power, and big models are known as the "three horses" driving the development of artificial intelligence technology today. With the emergence of big models, the development of artificial intelligence has entered a new stage. However, most of the advanced models currently use open source code frameworks for debugging and deployment, and these frameworks generally have security vulnerabilities. As early as  2017 , the University of Georgia, the University of Virginia and other institutions in the United States found  15 vulnerabilities in three platforms , including  TensorFlow , Caffe and Torch , including denial of service attacks and so on. According to data from the open source software community GitHub , since 2020 , Tensorflow has been exposed to more than 100 security vulnerabilities . Among them, the Baidu security team found 75 security vulnerabilities that can cause system instability, data leakage, memory corruption and other problems; 360 company found  49. Recently, relevant institutions have conducted security assessments on mainstream open source  AI frameworks at home and abroad. A total of  7 machine learning frameworks have been found to have more than  150 vulnerabilities and more than  200 vulnerabilities in the framework supply chain . According to statistics, more than 80% of the existing deep neural network models in China use foreign platforms. These development platforms have been exposed to have many security vulnerabilities, and their potential risks will lead to security issues such as system instability and data leakage. Under certain conditions, attackers can also achieve remote code execution and directly control the target system. Researcher Chen Ping told reporters that vulnerabilities and backdoors, as inherent security defects of artificial intelligence systems, will be purposefully exploited by attackers and will become a "major concern" in cognitive domain operations.

**Constructing a “three-dimensional barrier” for cognitive domain security**

Researcher Chen Ping told reporters that in the face of the current severe situation of cognitive domain confrontation, in addition to solving basic security issues such as real-time defense, timely detection, and immediate response, it is also necessary to strengthen cognitive security from three dimensions, namely data security, algorithm security, and model framework security, to build a three-dimensional defense barrier. Recently, IBM proposed a "triple protection" cognitive security immune system prototype, namely security platform, security operation and maintenance, and security threat data center, which is very inspiring for us to strengthen cognitive security.

From the perspective of data security, the main method of attackers is to quietly "concoct" a certain amount of malicious data after mastering the training data of the cognitive model, and to maliciously disturb the target model by "poisoning" the cognitive model, thereby affecting the correct judgment of the cognitive model. How to ensure the security of the data set in cognitive confrontation, prevent "poisoning" and resist "pollution". Researcher Chen Ping believes that **security defense can be carried out in three different stages: cognitive data and feature modification, cognitive model modification, and cognitive output** . Preprocess the input data to prevent "toxic data" from infiltrating at the first "gate". Prune the training model in a timely manner, and defend against data "poisoning" by reconstructing learning model parameters, limiting conditional queries, and other methods. In the "last mile" of cognitive output, add multi-model prediction voting, output result reprocessing, joint learning and other methods to form a closed defense loop.

From the perspective of algorithm security, people cannot find logical errors in the algorithm from the decision-making process of the cognitive model, which may lead to the occurrence of algorithm "bias" and other situations. In the cognitive field, some algorithms with "bias" will cause the model to make decision results that are completely inconsistent with the actual situation, thereby causing serious errors in the cognitive domain, and then triggering a series of chain reactions in the society, and its direct and indirect losses will be difficult to estimate. Therefore, an artificial intelligence method based on explainability ( XAI ) is urgently needed. It can make the decision-making of artificial intelligence transparent and trustworthy, so as to better help researchers optimize model strategies and provide the outside world with an execution result that can trace back its analysis and decision-making process. Just like a doctor's consultation, XAI can not only let researchers know which "prescription drugs" to prescribe, but also let them understand which reasons are "causes" and which phenomena are "pathologies". The decision-making algorithm built based on XAI can clarify the decision-making analysis of the system in each link by tracing back the decision-making process of its system, making its model "trustworthy", and adjusting it through experts, so as to optimize the operation of the system, overcome the "bias" of the system, enhance the trustworthiness of the system, and ensure "consistency between words and deeds".



Conceptual diagram of algorithm security solution (Photo provided by He Zhuolin)

From the perspective of training model security, it mainly refers to the existence of vulnerabilities in the model itself. After being attacked, it will not work properly. It can even use the infinite sensitivity of machine learning to induce the model to output wrong results. In order to ensure the security of the training model, **it is necessary to "train the muscles and bones" and strengthen the process management of vulnerabilities; it is also necessary to "train the breath" and prevent vulnerability risks early** . On the one hand, it is necessary to strengthen the management of artificial intelligence network security vulnerabilities, establish an artificial intelligence vulnerability resource library, standardize the entire process of vulnerability discovery, reporting, disclosure and repair, and realize closed-loop management of vulnerabilities. On the other hand, we must focus on improving the defense capabilities against vulnerability backdoors, establish and improve risk monitoring, early warning, notification and disposal mechanisms, improve threat discovery, attack blocking, tracing and other technical means, timely discover risks and eliminate security risks. At the same time, we must also focus on promoting defense through attack, actively build cognitive domain attack weapons, establish our own vulnerability library for artificial intelligence systems and training models, and effectively reserve strategic resources such as vulnerabilities. In particular, we must strengthen the mining of common model vulnerabilities across languages, applications, platforms and systems to seize the joint point of "one move to defeat the enemy".

**Promoting innovation in cognitive domain security technology paradigm**

Researcher Chen Ping pointed out that establishing a "three-dimensional" defense system is only a stopgap measure. To fundamentally deal with security threats in the cognitive domain and crack the two major security issues of "unexplainability" and "system vulnerability backdoors", it is also necessary to promote innovation in the development paradigm of cognitive domain security technology. **By constructing an endogenous security defense system, intelligent systems with cognitive capabilities can have endogenous security functions at the "gene" level and obtain non-specific immunity.** In layman's terms, the essence of cognitive security still requires the formation of "cognitive immunity" capabilities.

Regarding the new paradigm of cognitive domain security technology, researcher Chen Ping introduced that my country is currently in a leading position in this research. As early as ten years ago, Academician Wu Jiangxing was the first in the world to propose the theory of mimicry defense, which successfully solved many contradictions faced by network security and digital security, such as security and advancement, maturity and credibility, autonomy and openness. Under the restrictive conditions of "supply chain insecurity", it ensures the security and controllability of digital space under open conditions; in the objective environment of "toxic bacteria", it maintains the national cyberspace security and the generalized functional security of critical information infrastructure. At the Digital China Construction Exhibition that just ended, mimicry defense technology was listed as one of my country's "Top Ten Hard-Core Technologies".



The picture shows the "Comic Mimicry Defense" published by Science Press, which popularizes the knowledge about the struggle in cyberspace and cognitive space to the public (Photo provided by Wang Danyu)

The innovation of cognitive domain security technology paradigm will fully draw on the implementation mechanism of cyberspace mimicry defense technology, and take the construction of artificial intelligence endogenous security theory and technology system as the forerunner. **The thinking perspective of artificial intelligence endogenous security technology paradigm is "three domains interweaving", specifically, the three domains of "physics-information-cognition" are interweaving, and openness rather than closure is the premise of artificial intelligence security.** The methodology is to rediscover relatively correct axioms, form a logical expression of multi-modal decision-making and feedback control, establish a dynamic heterogeneous redundant structure and a multi-agent collaborative decision-making model, to express "known unknowns and unknown unknowns" , and finally establish "cognitive immunity" capabilities. The practice norm is to transform the uncertainty problems in artificial intelligence technology into consistency problems and quantifiable problems, so as to realize the quantifiable design and verifiable measurement of the security functions of various cognitive systems, and build the generalized robust functions of artificial intelligence.

Researcher Chen Ping predicted that driven by the theory of mimicry technology, the intrinsic security technology of artificial intelligence will become the "golden bell" to maintain the security of the cognitive domain. On the basis of the "three-dimensional" and external security barriers of data security, algorithm security, and model security, it will establish an intrinsic security line of defense based on "cognitive immunity", providing strong support for promoting the high-quality development and high-level security of cognitive domain technology.