DOI 10.15302/J-SSCAE-2021.03.005

Technical Countermeasures for Security Risks of Artificial General Intelligence[1]

Liu Yuqing, Zhang Yuhuai, Duan Peiqi, Shi Boxin, Yu Zhaofei, Huang Tiejun, Gao Wen

Department of Computer Science and Technology, Peking University, Beijing 100871, China

Abstract: Human beings may face significant security risks after entering the era of artificial general intelligence (AGI). By summarizing the difference between AGI and traditional artificial intelligence, we analyze the sources of the security risks of AGI from the aspects of model uninterpretability, unreliability of algorithms and hardware, and uncontrollability over autonomous consciousness. Moreover, we propose a security risk assessment system for AGI based on the aspects of ability, motivation, and behavior. Subsequently, we discuss defense countermeasures in the research and application stages. During the research stage, the theoretical verification should be improved to develop interpretable models, the basic values of AGI should be rigorously constrained, and technologies should be standardized. During the application stage, man-made risks should be prevented, motivations should be selected for AGI, and human values should be given to such intelligence. Furthermore, it is necessary to strengthen international cooperation and the education of AGI professionals to prepare for an unknown AGI era.

Keywords: artificial general intelligence (AGI); security risk; risk assessment; coping strategy

1 Introduction

Artificial general intelligence (AGI)[2] refers to the ability of a set of systems to process various intelligent behaviors, in contrast to narrow artificial intelligence, which requires an independent system for each intelligent behavior. It is impossible to implement AGI solely by improving the narrow AI algorithms (while ignoring systematic updates) [1]. In terms of cognitive theory, the concept of AGI emphasizes the existence of consciousness and highlights systems of values and worldviews. It posits that an intelligent artificial agent can have biological instincts. AGI is not necessarily in a humanoid form. It may be similar in appearance to human beings (sharing a common way of life) or may look extremely different (forming a new way of life). In terms of thought, AGI can share a set of thinking modes and moral standards with human beings or have its own unique reasoning method and thus become a kind of "machine with a soul." Research analyzing the brains of human beings and animals in more detail than in the artificial neural networks widely applied today might be realized within the next 20 years, allowing the construction of an architecture for future neural networks. The resulting neurocomputer could become a physical implementation platform for AGI [2].

Despite the human desire to develop AGI, doing so may result in substantial problems owing to social manipulation, new kinds of wars, and changes in power dynamics. Initially, AGI will obey the commands of human beings, but over time, it will trend toward making autonomous decisions. Whether such decisions will affect the interests of human beings,

[1]**Received date:** April 07, 2021**; Revised date:** April 25, 2021

Corresponding author: Shi Boxin, Assistant Professor and Research Professor of Department of Computer Science and Technology, Peking University. Major research field is computational photography and computer vision. E-mail: shiboxin@pku.edu.cn

Funding program: CAE Advisory Project "Strategic Research on the New Generation Artificial Intelligence Security and Self-control" (2019-ZD-01)

Chinese version: Strategic Study of CAE 2021, 23(3): 075–081

Cited item: Liu Yuqing et al. Technical Countermeasures for Security Risks of Artificial General Intelligence. *Strategic Study of CAE*, https://doi.org/10.15302/J-SSCAE-2021.03.005

[2]Translator's note: In English, the definitions of "strong AI" and "artificial general intelligence" are not commonly agreed. Some people use the terms interchangeably, while others view them as distinct concepts. In the original Chinese article, the authors use 强人工智能 (*qiang rengong zhineng*), which translates literally as "strong AI," abbreviating it to the English "AGI" throughout the body of the text. This translation adopts "AGI" throughout and refers the reader to the first paragraph for explanation of what the authors mean by the concept. Accordingly, 弱人工智能 (*ruo rengong zhineng*) is translated as "narrow artificial intelligence" rather than the more literal "weak artificial intelligence."

including their survival and the security[3] of their property, remains unknown. Intense discussions in the scientific community regarding AGI research are ongoing. Viewpoints include:

- The basic methods of existing AI are flawed, and we have to move toward AI with the ability to comprehend. The development of true AI is still very far off [3]
- It will require several decades for scientists to develop autonomous intelligence (i.e., AGI). We are currently faced with foundational problems, which are still essentially mathematical challenges [4]
- Current progress in AI technology comes from narrow AI. Mainstream academia does not consider AGI as a direction for development; it recommends not carrying out active research on AGI owing to various concerns [5]
- Humanity cannot remain satisfied with a weak version of AI. The evolution of intelligence cannot be halted. The ultimate scientific problems, such as the mysteries of consciousness, remain to be solved [6]

Therefore, given that AGI research presents both risks and opportunities, effective safeguards and codes of conduct for AI researchers and program developers must be formulated.

Evaluating and formulating strategies for coping with potential AGI security risks, and finding measures that will ensure that AGI is beneficial to humanity rather than harmful to society, have become research topics worldwide. For instance, in 2016, the U.S. lab OpenAI analyzed potential security problems that might arise in the development of AI [7]. In 2018, the U.S. government established the National Security Commission on Artificial Intelligence [8]. In addition, the EU set up the High-Level Expert Group on AI to help it strive for discourse and rule-making power in technological development [9]. AI has also become a subject of significant attention in the field of national defense. For example, AI is being adopted to improve the capability of defense systems, and AI anomaly detection technology is being developed to prevent malicious tampering of private data. AI theories and technologies, including algorithms integrating multiple disciplines, self-adaptive situational awareness, and human-machine trust, are also being studied [10].

It should be noted that, in terms of research related to AGI security issues, there is a gap between China and the international frontier of progress. Chinese academic and industry circles are paying more attention to the development of AI and less to the value of and need for AGI security. In this study, the sources of AGI risks are analyzed from three aspects: model uninterpretability, unreliability of algorithms and hardware, and uncontrollability of autonomous consciousness. Security risks are evaluated from the dimensions of ability, motivation, and behavior, and suggestions are proposed to reduce security risks at both the theoretical and application levels.

2 Sources of AGI security risks

2.1 Model uninterpretability

In traditional AI, the deceptive effect of deepfakes [11] is widely recognized, and specialized research has been conducted on gradient-based attacks and defense. Given that the basic processing unit of a convolutional neural network (CNN) is texture, the essence of a gradient-based attack is to conduct different operations on different responses generated by texture. There also exists a phenomenon of target bias in the training of generative adversarial networks (GANs) [12], which can be explained through the motto, “if something is too difficult, it will not be generated” [13].

When a system cannot be explained, it is impossible to confirm whether its objective is affected by other factors during an operation. For example, if a diagnostic system based on a neurocomputer makes a diagnosis after analyzing the patient data, the reliability of the diagnosis can only be estimated from a statistical perspective. If the factors considered for the diagnosis are unknown, it becomes difficult to completely trust the results deduced by the machine. A neurocomputer is one of the basic routes for implementing AGI [14]. Spikes are the carriers of signals in a neurocomputer, but when analyzing a neurocomputer, it is not known whether noise affects the classification results (the premise here is that the phases of the peaks and troughs do not change after a superposition). Similar problems may also arise in the training of AGI. If mode collapse occurs, there is a risk of malicious use. Hence, model uninterpretability is a potential security risk in an AGI system.

[3] Translator’s note: In English, “safety” and “security” typically refer to protection against unintended and deliberate harms respectively. In Chinese, the word 安全 (*anquan*) encompasses the meaning of both “safety” and “security”. Rather than try to select one of these terms in each case according to the context, which would involve a high degree of subjective judgement, 安全 (*anquan*) is consistently translated as “security” throughout the piece.

DOI 10.15302/J-SSCAE-2021.03.005

2.2 Unreliability of algorithms and hardware

The development and application of AGI will greatly influence the AGI industry as well as people's lifestyles; however, current AGI algorithms and hardware cannot yet meet the requirements of security or reliability, or act in accord with expectations. When designing an algorithm, an immature design scheme, such as the failure to consider all possible situations or software and hardware compatibility, may cause a system breakdown. A European carrier rocket failed because the high-precision data surpassed the number of digits that the hardware could support [15].

When an AI expert system serves society, the assumptions underlying the system may become invalid in certain circumstances, resulting in a system breakdown. In the Wall Street flash crash, an incorrect assumption led to a serious error in stock pricing, causing a loss of over a trillion dollars and severely affecting the American securities market [16].

The information security of algorithms and hardware is becoming an important pillar for maintaining security in the economy and society. There have been news reports of hackers exploiting system loopholes to steal personal information and private data from institutions and companies, with adverse effects on society. From this, it is reasonable to infer that AGI will be attacked by hackers and malicious software once it becomes widely used, resulting in data leakage and even endangering public security.

2.3 Uncontrollability of consciousness

The construction of an initial intelligent agent and effective principles for evolution are key to the design of an AGI system that can conduct self-development and self-iteration. Although human beings can control the initial intelligent agent well, AGI can design rules of evolution autonomously, possibly much more efficiently than human beings. After it undergoes recursive self-improvement, AGI will have a higher development efficiency in the subsequent stages and will surpass the cognition of human beings by a long way through recursive self-improvement.

AGI with autonomous consciousness carries potential risks. Unlike those of the human brain, the computational and analytical abilities of AGI are theoretically limitless. AGI has efficient data collection, processing, and analysis abilities and can understand all the information it sees, hears, and receives. Once it achieves consciousness, AGI will be able to share and exchange information through communication and significantly improve its understanding of the world and the efficiency through which it can transform reality. Accordingly, AI may gradually conduct various human activities. With the emergence of consciousness, the legal status of AGI becomes unclear: Is it a subject with consciousness or personal property? This may lead to disagreements at the legal, ethical, and political levels, and cause unexpected consequences.

3 AGI security risk assessment

3.1 Capability risk assessment

Moravec proposed the "landscape of human competence," [17] which described the development of human and computer capabilities and how difficult it is for them to address various problems. Within this landscape, the "altitude" of a task represents how difficult it is for a computer, and the ever-rising "sea level" represents tasks the computer can perform at present. The critical point is reached when the computer is able to design intelligence independently. Prior to this critical point, the algorithm design was mainly controlled by human beings. After the critical point is reached, humans will be replaced by computers in the research and development of intelligence, which will result in quantitative to qualitative leaps and radical changes to productivity and living standards. Whereas humans construct algorithms in accord with certain principles and experience, algorithms designed by AI cannot be permanently guaranteed to be reliable. For users, the current AI is like a "black box," whose internal operating logic and basis for decision-making cannot be probed (or are very difficult to probe).

3.2 Motivation risk assessment

Human intelligence and its products are extremely precious to the development of civilization. We can trust that life will be better if AI is efficiently utilized to improve productivity and create new tools. Some disruptive technologies are derived from small improvements or innovations, but they play a notable role in increasing productivity. However, the question of how to make rational use of increased productivity and technological leaps brought about by AI without causing new social problems is one that humanity should pay significant attention to. For example, how can we build a

robust AGI? How can we control AI weapons and avoid a harmful arms race? How can the application of AI to productivity avoid aggravating inequality in social distribution?

There is no need to worry that AI might cause harm to human beings when it is weak and can be controlled by them. However, once AI completely surpasses humans in all abilities and possesses consciousness, it will become difficult to assess whether AI will necessarily continue to obey the orders of human beings. This situation has been called the "treacherous turn" [18]. Although the questions of whether AI has human consciousness and how it will realize human-like consciousness remain unanswered, they are worthy of attention and research.

3.3 Behavior risk assessment

The supervision and control of AGI behaviors can be regarded as a "principal-agent" problem, where humans are the principals and AGI systems are the agents. However, this differs from the current "principal-agent" problems between people, because AGI can formulate differential strategies and actions based on its analytical capabilities and knowledge reserves. Therefore, the monitoring of AGI behavior during testing at an early stage of research and development cannot support humans in making rational inferences about the future reliability of AGI. This means that behaviorist methods may fail.

4 Risk management strategies in theoretical and technical research stages

4.1 Improvement in theoretical foundation verification and exploration of model interpretability

Improving the verification of theoretical foundations and exploring the interpretability of models constitute the foundations of AGI accuracy and the formal guarantees of AGI security.

The model design of AGI should be explored based on cognitive neuroscience, the discipline that studies the brain's structure and investigates the brain's mode of operation based on its biological structure and the cognitive ability of human beings. A suitable AGI model can be designed based on the structure and mode of operation of the human brain.

The implementation of AGI should be based on meta learning, a method of learning how to learn [19] that enables AI to think and reason. As a key research direction of deep learning, meta learning aims to learn relevant information from data and to endow present-day AI with the ability to learn new knowledge automatically. For today's AI, a new task generally means learning new knowledge from scratch, which is time-consuming and inflexible. Meta learning is experience-directed because it involves learning solutions for new tasks based on past experience. This can equip AI with more skills and allow it to adapt better to complex real-world environments. As one of the implementation methods of semi-supervised and unsupervised learning, meta learning is an important mathematical implementation for simulating human learning processes. Seeking methods for such simulations can improve the model interpretability, explore ways to enable AGI to "learn to learn," and develop consciousness similar to that of human beings.

The interpretability of deep learning should be explored from the perspective of mathematics. Currently, there are no generally accepted, systematic theoretical frameworks that explain deep learning, and the interpretability of relevant models is still regarded as a complex problem. Current methods for exploring the interpretability of deep learning from a mathematical perspective include information theory, structural expression, the ability to generalize, the principles of dynamics, and manifold learning. Exploring the functions and contributions of each component module of the models and conducting a pattern analysis of their structure and functions from a semantic perspective are areas that AGI interpretability research needs to closely follow going forward.

4.2 Strict control of underlying value orientation of AGI

The underlying value orientation of AGI must be constrained and monitored by corresponding rules and memory.

Explicit rules should be designed to limit the range of action of AI. In view of the complexity and uninterpretability of AI, it is difficult to constrain and monitor its value orientation using the source code. Constraining the value orientation of AI from a behavioral perspective and limiting the behavioral ability and action permissions of AGI using explicit rules are key research objectives. An underlying value network can be constructed during the process of meta learning to

DOI 10.15302/J-SSCAE-2021.03.005

accelerate inference training and guide the action network to take action.[4] The algorithm for the underlying value network is complex, and the dataset cannot be controlled, making it extremely difficult to adopt measures to limit the inference process of the network. For the action network, explicit rules can be manually added to ensure that each action is in line with the correct values (i.e., limiting the occurrence of incorrect behaviors for every independent action).

Trusted computing technology should be applied to monitor AI actions. Trusted computing is a mechanism for defending against malicious code and attacks and can be regarded as an "immune system" for computers. Additional supervision is introduced to build a complete, trustworthy, and quantifiable evaluation mechanism for various computer behaviors, and then judge whether these behaviors meet the expectations of human beings, thus preventing and handling actions that cannot be trusted. The operation process of AGI should be monitored and analyzed. A time series analysis can be used to determine if the current behavior has a reasonable value orientation. If it does not conform to such an orientation, an external intervention method should be adopted to interrupt the current action of the AGI and ensure that the AGI will not act contrary to values.

4.3 Implementation of technology standardization

AGI technology standardization involves four aspects: standardization of the model design, training methods, datasets, and security guarantees.

(1) *Standardization of the model design* - Currently, research in deep learning and AI has led to the creation of some extensively applied basic modules, for example, 3 × 3 convolution layers, rectified linear units, and batch normalization. Different basic modules can be used to construct differentiated neural networks. A standardized design of the basic modules is conducive to unifying the design of the interfaces and configuration documents. Using a general descriptive language to represent a neural network can facilitate the migration and deployment of the model. In addition, it supports the use of hardware chips and driver programs to achieve a targeted acceleration. Taking CNNs as an example, the compute unified device architecture (CUDA) and CUDA deep neural network library (cuDNN) developed on this basis have accelerated convolution operations, significantly improving the training and inference speeds.

(2) *Standardization of training methods* - Training is an essential part of AI. Different networks can solve network weights through different training hyperparameters, optimizers, and algorithms. Diversified training methods lead to poor model reproducibility and failure of the neural network optimizer to obtain hardware acceleration support during the iteration process. The key to training standardization is to design a set of reasonable training frameworks, abstract different optimizers as interfaces, and provide hardware acceleration support to uniform interfaces, thereby improving the training efficiency.

(3) *Dataset standardization* - This mainly refers to the release of open-sourced, standardized, and commonly recognized datasets put forward by various industries for use in model training and testing. The standardization of datasets can strengthen data security and improve data quality. It is important to promote the formulation of standardized datasets in various industries and offer open and high-quality benchmarks.

(4) *Standardization of security guarantees* - Security is an essential prerequisite for the use of AGI. General, clear, and executable standards should be developed to ensure the security of the AGI design, training, and operation. These standards should be scalable to adapt to the environmental complexity of AGI applications. Standardizing the security guarantees, designing methods appropriate for the characteristics of different stages, and ensuring the proper operation of AGI are the strongest assurances for responding to relevant risks.

5. Risk defense strategies for AGI during the application stage

5.1 Preventing anthropogenic AI security problems

The application of AI technology for fraudulent behavior, such as the quick creation of convincing deepfake videos, is increasingly attracting the public's attention. A study has summarized deepfake detection techniques, including traditional image forensics, biological signal analysis, traces of image tampering, and image features of GANs [20]. Although research is progressing in the detection of deepfake images, the emergence of new synthesis techniques has led

[4]Translator's note: The value network (价值观网络, *jiazhiguan wangluo*) and action network (行为网络 *xingwei wangluo*) referred to here are distinct from the value network (数值网络 *shuzhi wangluo*) and policy network (策略网络 *celue wangluo*) of reinforcement learning.

to significant difficulties in identifying deepfake content. Only by establishing as great a technical advantage as possible can those identifying deepfake videos defeat those producing them. In addition, supplementary means such as legislation and training for the news industry can be adopted to respond to security problems resulting from the application of such technology.

Potential errors in the algorithm design should also be taken seriously. Although the applicability of AI has been proven, the algorithm design inevitably contains flaws. As a result, security should be prioritized, particularly in fields directly related to human security, such as automated driving, telemedicine, and industrial manufacturing. Major incidents have resulted from errors in the autopilot systems of civilian aircraft, accompanied by an inability to switch to manual operation. Against the backdrop of further development of AI technology and its increasingly widespread application, security issues must be considered from the beginning to avoid the potentially severe consequences of malicious attacks on systems and data or interference from false signals.

The introduction of third-party components may also cause security problems. This is both an issue in the traditional security field and an important factor affecting AI security. Malicious third-party components may lead to problems such as a system breakdown and a misappropriation of system permissions.

5.2 Motivation selection for AGI

At the “treacherous turn” stage, AI has already developed cognitive abilities far exceeding those of human beings in various fields. This can be referred to as superintelligence [18]. Given the reasonable assumption that superintelligence may betray human beings, humans should select the motivation of intelligent agents in advance to fully prevent undesirable results and equip superintelligence with the innate wish to not harm human beings.

Four approaches to motivation selection have been proposed: direct specification, domesticity, augmentation, and indirect normativity [18]. Direct specifications can be either rule-based or consequentialist. A traditional illustration of a rule-based specification is the “three laws of robotics” concept [21]. Regarding the first law that a robot shall not harm a human, there are several unresolved questions. How can harm to humans be weighed? How can we define “harm” and “human”? Why do we not consider other sentient animals and digital minds? Producing a set of complex and detailed rules that are suitable for highly diverse situations and getting things right the first time seem impossible under the current conditions. Consequentialist specifications also face challenges because there are many different ways to achieve the same result, and the computer code must describe the goals precisely. For instance, if the goal of an AI is to make people smile, achieving that goal by making people happy is obviously different from doing so by stimulating their muscles. Domesticity can be considered a self-limitation [18]. As a special ultimate goal, we can attempt to constrain the ambition of AI by shaping its motivation such that it eventually limits its own behavior within a stipulated range. Augmentation refers to starting with an intelligent agent with an acceptable motivation and enhancing its intelligent behaviors through a transformation. As a weakness of this method, it is difficult to ensure that the motivation system will not be changed or destroyed once cognitive abilities are greatly improved. Unlike direct specification, indirect normativity involves specifying a process for deriving a normative standard and allowing AI to carry out this process.

5.3 Endowing AGI with human values

Although motivation selection improves the effectiveness of human control over AI, compared to limiting its ability, several problems still exist. For example, AI may face an infinite number of situations, making it impossible to discuss solutions for every situation, and it is infeasible for human beings to continuously monitor the motivation of AI. In this case, one feasible solution is to endow AGI with human values (by loading them into the AGI), thereby allowing it to consciously execute actions that will not pose a threat to human beings. It is impossible to fully represent the motivation systems present under all situations in a table (which would lead to an infinitely large table). Such systems can only be expressed in a more abstract manner, using formulae, rules, and other factors.

The use of evolutionary algorithms is a feasible route for value loading. With this method, rules are produced randomly, and candidates are screened through an evaluation function (by removing candidates with low scores and retaining those with high scores). Reinforcement learning methods can maximize the cumulative reward of intelligent agents such that the agents accumulate values as they learn to deal with various problems.

However, the accumulation of human values is the result of our genetic mechanism evolving over millions of years and imitating or reproducing this process would be extremely difficult. Because this mechanism has adapted to the neural

DOI 10.15302/J-SSCAE-2021.03.005

cognitive architecture of human beings, it can only be realized through whole-brain emulation [22]. As the premise of whole-brain emulation, the brain is a computer that can be simulated. However, it faces three challenges: scanning, translation, and simulation [18]. The required precision can only be achieved using high-throughput microscopy and supercomputing systems.

5.4 International cooperation for AGI

AGI research has become a subject of international attention. Only by concentrating the scientific and technological strengths of the whole of humanity can we ensure that AGI better serves society. The process of AGI research and its gradual application involve many unknown problems. Strengthening international AGI cooperation and promoting the sharing of research will be necessary to improve the ability to respond to emerging situations and guarantee the implementation and expansion of AGI applications.

Importance has already been placed on international AGI cooperation, and some countries and regions have provided policy support for such cooperation through legislative and other methods. In 2018, 25 EU member states signed the *Declaration of Cooperation on Artificial Intelligence* in Norway [23]. The declaration committed to promoting dialogue and cooperating on an aligned approach to AI research and application cooperation. EU member states have also used joint statements and other methods to encourage legislative cooperation in high-priority fields, including key issues such as data protection, ethical standards, and data rights. These are all useful practices for China to refer to in carrying out international cooperation in the area of AGI.

5.5 AGI talent cultivation

Talent cultivation is the basis of scientific research. Because AGI is a direction for frontier technology development, the scale, speed, and quality of the corresponding talent cultivation are clearly unable to meet the development requirements of the field. The strengthening of talent cultivation, particularly local talent, is urgently needed. In the technical field, we should optimize the mechanisms and environment for talent education, cultivation, and growth and quickly develop professionals with specialized research and development knowledge. In the management field, we should emphasize the cultivation of entrepreneurs and operational talent who show aptitude for commercial promotion and demand expansion. Through cooperation between industry and academia, and between research and application, we can provide the talent necessary for the healthy and stable development of AGI.

6 Conclusion

The intelligence and behavior of AGI cannot simply be equated to those of human beings. The motivation for creating AGI is to benefit human society. However, to protect the privacy of individuals and society as a whole, AGI should be controlled such that it only serves human beings passively, rather than learning on its own initiative. If there is an intelligence explosion once AI has evolved to a certain level, the default result will inevitably be catastrophic. In view of such potential threats, humanity should continue to monitor the risks and search for countermeasures to avoid the occurrence of this default ending. Humanity should design a controlled intelligence explosion and set the proper initial foundations, all while achieving humanity's desired results and ensuring that all consequences remain within an acceptable range.

In the future, we recommend paying close attention to the technological evolution of AGI and proposing dynamic strategies for responding to potential security risks. We should examine international discussions and drafting of AGI policies, integrate cutting-edge legal and ethical findings, and explore the elements of China's AGI policymaking in a deeper and more timely manner.

Acknowledgement

We would like to thank Brian Tse for his contribution to this translation. Brian Tse is the founding director of Concordia Consulting, Policy Affiliate at the Centre for the Governance of AI (GovAI), and one of the founding members of the Beijing Academy of AI's AI4SDGs Cooperation Network. Cynthia Chen, Yawen Duan, Kwan Yee Ng, and Julia Chen, in particular, contributed to this translation.

References

[1] Chen J B, Gao Y F. Artificial intelligence and human intelligence from the perspective of system theory [J]. Studies in Dialectics of Nature, 2019, 35(9): 99–104. Chinese.

[2] Huang T J, Yu Z F, Liu Y J. Brain-like machine: Thought and architecture [J]. Journal of Computer Research and Development, 2019, 56(6): 1133–1148. Chinese.

[3] Zhang B. Towards the real artificial intelligence [J]. Satellite & Network, 2018 (6): 24–27. Chinese.

[4] Xu Z B. AI and math go together towards the era of autonomous intelligence [EB/OL]. (2020-06-08) [2021-02-15]. http://news.sciencenet.cn/htmlnews/2020/6/441057.shtm. Chinese.

[5] Zhou Z H. Views on artificial general intelligence [J]. Communication of the CCF, 2018, 14(1): 45–46. Chinese.

[6] Huang T J. Different views on artificial general intelligence [J]. Communication of the CCF, 2018, 14(2): 47–48. Chinese.

[7] Amodei D, Olah C, Steinhardt J, et al. Concrete problems in AI safety [EB/OL]. (2016-07-25) [2021-02-15]. https://arxiv.org/abs/1606.06565.

[8] Congress of the United States. H.R.5356-National security commission artificial intelligence act of 2018 [EB/OL]. (2018-03-20) [2021-02-15]. https://www.congress.org/bill/115th-congress/housebill/5356.

[9] China Academy of Information and Communications Technology. Global AI governance report [EB/OL]. (2020-12-30) [2021-02-15]. https://pdf.dfcfw.com/pdf/H3_AP202012301445361107_1.pdf?1609356816000.pdf. Chinese.

[10] Jin J, Qin H, Dai Z X. Top-level strategy of artificial intelligence security and the research status of key institutions in the United States [J]. Civil-Military Integration on Cyberspace, 2020 (5): 45–48. Chinese.

[11] Whyte C. Deepfake news: AI-enabled disinformation as a multilevel public policy challenge [J]. Journal of Cyber Policy, 2020, 5(2): 1–19.

[12] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks [J]. Advances in Neural Information Processing Systems, 2014, 3(11): 2672–2680.

[13] Bau D, Zhu J Y, Wulff J, et al. Seeing what a GAN cannot generate [C]. Seoul: 2019 IEEE/CVF International Conference on Computer Vision, 2019.

[14] Huang T J. Imitating the brain with neurocomputer a “new” way towards artificial general intelligence [J]. International Journal of Automation and Computing, 2017, 14(5): 520–531.

[15] Qu J, Zhang L Y. Statistical analysis of foreign rocket launch and failure [J]. Aerospace China, 2016 (2): 13–18. Chinese.

[16] Xing H Q. Research on the legal regulatory framework of high frequency trading in securities and futures market [J]. China Legal Science, 2016 (5): 156–177. Chinese.

[17] Tegmark M. Life 3.0: Being human in the age of artificial intelligence [M]. New York: Penguin Random House LLC, 2017.

[18] Bostrom N. Superintelligence: Paths, dangers, strategies [M]. Oxford: Oxford University Press, 2015.

[19] Vilalta R, Drissi Y. A perspective view and survey of meta-learning [J]. Artificial Intelligence Review, 2002, 18(2): 77–95.

[20] Li X R, Ji S L, Wu C M, et al. Survey on deepfakes and detection techniques [J]. Journal of Software, 2021, 32(2): 496–518.

[21] Asimov I. I, robot [M]. Louisville: Spectra Press and Promotions, 2004. Chinese.

[22] Huang T J. Can human build “super brain”? [N]. China Reading Weekly, 2015-01-07(5). Chinese.

[23] Ministry of Science and Technology of the People’s Republic of China. 25 European countries sign the *Declaration on Artificial Intelligence Cooperation* [EB/OL]. (2018-07-18) [2021-02-15]. http://www.most.gov.cn/gnwkjdt/201807/t20180718_140708.htm. Chinese.