Vol.4 No.4(2019)

INTERNATIONAL SECURITY AND STRATEGY STUDIES

REPORT

Principles and Pivots of Artificial Intelligence Governance

清华大学战略与安全研究中心
CENTER FOR
INTERNATIONAL SECURITY AND STRATEGY
TSINGHUA UNIVERSITY

Principles and Pivots of Artificial Intelligence Governance

Fu Ying[1] Li Ruishen[2]

The wide application of Artificial intelligence (AI) not only has significantly facilitated the production and life of mankind today, but also will potentially bring about disruptive changes in the future. At the same time, risks and challenges related to AI are also a rising concern worldwide. In January 2015, hundreds of entrepreneurs and AI experts, including Stephen Hawking, the well-known physicist, cosigned an open letter to the public, warning that AI technologies should be effectively regulated; otherwise "it will spell the end of the human race." The letter sparked off people's fears and concerns over AI, which has been frequently discussed by the public and the media. Many countries and organizations have therefore started to think about establishing an AI security governance mechanism.

In 2017, industry leaders worldwide developed Asilomar AI Principles, which serve as self-disciplined regulations requiring technological advances to "benefit human beings". The European Commission also issued AI Ethics Guidelines. In 2019, the Organization for Economic Cooperation and Development (OECD) officially approved the first intergovernmental policy guidelines on AI, which aims to ensure AI system design can be righteous, safe, fair and trustworthy to meet the international standards.

[1] Chairperson for Center of International Security and Strategy, Tsinghua University

[2] Adjunct fellow at the Center for International Security and Strategy, Tsinghua University

The G20 also released G20 AI principles, proposing the application, research and development of AI to "respect the rule of law, human rights and democratic values". China's National Professional Committee on New Generation AI Governance outlined Principles of New Generation Artificial Intelligence Governance to encourage the development of responsible AI.

1. Six Principles of Governance

In July 2018, the AI Governance Project Group at Tsinghua University put forward "Six Principles of AI Governance" at the World Peace Forum (WPF)[3], which provides a general framework on comprehensive governance of AI: a. Principle of well-being. AI development shall serve the well-being and benefits of human beings; AI designs and application shall abide by the ethics of human society and respect the dignity and rights of human beings. b. Principle of security. AI shall not bring any harm to human beings; AI system shall be secure, applicable and controllable, and be able to protect privacy and prevent data disclosure and abuse. AI algorithm shall be traceable, transparent and free from algorithm discrimination. c. Principle of sharing. Economic prosperity brought by AI shall serve the entire human race. A rational AI mechanism shall be developed to benefit and facilitate more people and avoid digital divide. d. Principle of peace. AI shall serve the purpose of peace, devote to the enhancement of transparency and confidence-building measures, encourage AI application in a peaceful way and prevent arms races of lethal autonomous weapons. e. Principle of law. AI shall be applied in

[3] World Peace Forum was established by Tsinghua University in 2012 and has held eight forums as of today. It is the only non-governmental high-level forum on international security in China. The Forum aims to provide a platform for global strategists and think tank leaders to discuss and seek constructive solutions to international security issues.

accordance with the Charter of the United Nations and basic principles of modern international law, namely, equal sovereignty for all countries, peaceful resolution of disputes, prohibition on the use of force and non-intervention of each other's internal affairs. f. Principle of cooperation. All countries shall promote exchange on AI technologies and talents, and regulate technology advances in an open and inclusive environment.

The Six Principles shed light on the discussion and consensus on AI governance. In the World Internet Conference at the end of 2018 and the World Congress of Peace this year, many scholars and entrepreneurs showed their interests and attention to AI governance, and a large number of organizations expected further cooperation and discussion. Nowadays, the industry has witnessed attempts with self-discipline. For instance, no-fly zone avoidance strategy has been coded into drones for better control; data masking has been practiced in medical and transportation industries to protect personal information and form a virtuous cycle of data use. The task at the moment is to promote the implementation of those principles in the international community for the establishment of a more practical and feasible governance mechanism.

2. The Key to the Governance Mechanism

The international governance mechanism means not only consensus and rules, but also organizations and capabilities to ensure the implementation of the rules, as well as the relevant social, political and cultural environment. The Center for International Strategy and Security of Tsinghua University is working with scholars, former politicians and entrepreneurs from different countries to discuss relevant issues. Based on what is happening now, an effective mechanism for international governance of AI should include at least the following five keys:

(1) Capability for Dynamic Update

Research, development and application of AI have witnessed rapid progress, but there are still a lot of uncertainties towards application scenarios and security challenges in the future. Therefore, AI governance requires full consideration of changes in technology and its application, as well as the establishment of a dynamic and open governance mechanism with automatic update capability.

For example, it is necessary to provide the society with a specific definition of “malicious use” of AI, which shall be observable and distinguishable in the production and life of mankind, as well as measurable and calibratable by technology. More importantly, it should be continuously updated. Only the governance mechanism with dynamic update capability can play a role in the rapid development of AI technology.

That means that while making progress in governance, it is necessary to accept uncertainties of AI technology and be ready to adjust the thinking. Einstein once said, “we can’t solve problems by using the same kind of thinking we used when we created them.” The conflict between disruptive innovation and conventional thinking will be everywhere in AI governance. The governance mechanism in this case should also be inclusive and adjustable to the intertwining and recurrent opinions. This mechanism will assist human beings to adapt to the endless challenges of AI technology. In this regard, establishing a dynamic governance mechanism that updates with the continuous technology development is probably more significant than setting the rules of governance.

(2) Technology Governance from the Source

AI application is essentially the application of a technology, so AI governance should be based on its technical nature. In particular, it is more effective to govern AI security from the source. For example, the current focus is on deep learning technology with data, algorithms and computing power as key elements. Therefore, it is better to start the governance from data flow control, algorithm audit, and computing power control.

With the rapid development of AI technology, there may be different intelligent technologies in the future, such as few-shot learning, unsupervised learning, Generative Adversarial Networks (GAN), and even brain-computer interface technology. Different technical principles mean that the latest and most important sections and tools of governance should be find out from the source of technology, and be incorporated into the mechanism for the sustainability of governance.

Another important section of technology governance is to endow AI with the “meritorious application” gene in technology. On the issue of weaponization of AI, for instance, whether there is possibility, like “Three Laws of Robotics” formulated by the novelist Asimov, to constrain AI in technology, and code the “principle of distinction” from Law of Armed Conflicts and International Humanitarian Law for prohibiting any attacks on civilian facilities. This is indeed a tough challenge. Paul Scharre[4], who used to work in the Office of the US Secretary of Defense and played a leading role in policy-making of autonomous systems, said: “It is very hard for machines of today to meet these standards (principle of distinction, principle of proportionality and avoidance of unnecessary pain). Whether it can be achieved depends on the goals pursued, the surrounding environment and future technological predictions.”

[4] Paul Scharre, *Army of None: Autonomous Weapons and the Future of War*, World Knowledge Press, 1st edition, June, 2019

(3) Multidimensional Characterization

International governance of AI requires the establishment of diversified governance ecology that involves all stakeholders. Scholars and experts are the main force driving technological development; politicians are the major state decision-makers; people's consumption demands are the core incentive to boost the progress of all parties. Sufficient communication and discussion among these groups underpins the foundation of AI governance. Enterprises are the core of technology transfer application; academic organizations are the core of industry self-discipline; governments and military forces are the core of AI security governance. Communication among these organizations is the key to truly implement AI governance mechanisms.

In this ecosystem, different groups should achieve deeper interpretations of AI governance rules from their own perspectives. For example, an article published in August this years, written by Henry Kissinger, Eric Schmidt, and Daniel Huttenlocher, stated that considering the impact of AI on philosophical awareness, it may be necessary to ban intelligent assistants from answering philosophical questions; human beings shall be asked to participate in influential identification activities; it is needed to "audit" AI and correct it when it violates human values.[5]

If bringing together the governing norms from different groups, there will be the wisdom of multiculturism, which guide human beings to tackle challenges brought by AI collectively. Many a little makes a mickle. Philosophers' concerns about truth and reality are as important as the public's fear for privacy disclosure. Only by carefully delineating the

[5] The Atlantic, August 2019, p.23

details of AI governance can confusion and fear turn into curiosity and hope.

(4) Effective Attribution Mechanism

In the international governance mechanism of AI, clear definition is the norm and start of governance; technology governance from the source is the key path; and participation of multiple stakeholders lays the foundation for governance. Attribution and imputation are the heart in the entire governance mechanism. If no one is responsible for this, then all governance efforts will ultimately be meaningless.

A major obstacle to the current AI governance is the difficulty of attribution: when it comes to human-machine relationship, does the greater responsibility of human beings in AI application bring more deterrent effect on malicious use and a bigger possibility of effective governance? In terms of social relations, in the case where various stakeholders all presume that AI has the possibility of "self-evolution", who is responsible for the consequences of the "self-evolution" of the program? Should that person be the creator, the owner or the user?

From a technical point of view, all machines in the world can have faults. Just like no one is perfect in this world. AI is doomed to cause property damage or even casualties sooner or later. Should machines really be endowed "personality" and be responsible? If so, does it mean that human beings give away their final adjudication power to machines to some extent?

(5) Reasonable Division of Scenarios

Before AI develops into “general intelligence,” an effective way to implement governance is to divide the scenarios and process them individually. The recent development only presents limited AI application scenarios. At the World Peace Forum in July 2019, many scholars believed that it is time to start from a few specific scenarios as soon as possible to accumulate governance experience and to achieve effective governance gradually.

Dividing scenarios helps us understand what and when AI can do. This can, on the one hand, relieve people’s fear due to the insufficient understanding about AI, and on the other hand, remove any exaggeration on AI functions. For example, even Robert O. Work, former US Deputy Secretary of Defense and active advocate of AI weaponization, had to admit that AI shouldn’t be extended to nuclear weapons in the context of nuclear weapons command and control, because it may cause disastrous consequences.[6]

Effective scenario division should be as close as possible to the actual physical and social scenarios, and should pay attention to the impact of data on the scenario. This is because the current AI technology is heavily associated with data. Different data may result in different scenarios. The scenarios should at least be subdivided into three categories: physical scenario, social scenario and data scenario.

[6] Breaking Defense website, August 29, 2019

3. Conclusion

For the human race, any new technology is a double-edged sword. Almost every major technological innovation would bring about discomfort and pain to people at that time. However, scientific advances and people’s thriving livelihood today prove that human beings have enough wisdom in the governance of new technologies. Any subsequent new threats can be resolved with good use and scientific governance of technologies. We believe that the international community will be able to form a well-functioned governance mechanism and enjoy a more prosperous and safer world enabled by AI technology. (End)

Center for International Security and Strategy of Tsinghua University

Address：Room 217,Mingzhai,Tsinghua University

Tel：+010-62771388

E-mail：ciss_thu@163.com