



# The Global Governance of Artificial Intelligence: Safety and Security Perspective





# **The Global Governance of Artificial Intelligence: Safety and Security Perspective**

Zou Quancheng & Lu Chuanying

**360 Dipper Research**  
**Research Center for Global Cyberspace Governance (SIIS)**

July 2022



# The Global Governance of Artificial Intelligence: Safety and Security Perspective

## Supporting organization

Tianjin Institute for Digital Security on Smart City

360 AI Security Lab

National Engineering Research Center of Big Data Collaborative Security Technology

Business Research and Development Center of ICBC

Finance Laboratory of Citic Foundation

China Telecom Corporation Limited Research Institute

Shanghai Association for AI and Social Development


Shanghai Cyber Research Institute



# Foreword

“The Global Governance of Artificial Intelligence: Safety and Security Perspective” is the first research report jointly published by 360 Dipper Research and the Research Center for Global Cyberspace Governance, the Shanghai Institutes for International Studies. Through the collaboration between academia and industry, both parties hope to explore a governance approach that can maximize technological benefits while controlling security risks in the rapid development of AI technology.

Security governance is a critical procedure that can direct the future development of AI. Our two research groups analyzed the challenges, governance models, and industry practices of AI security governance from policy formulation and industry practice. Our findings suggest that artificial intelligence is a critical emerging technology actively developed in China, the United States, Europe, and other countries or regions; meanwhile, the security challenges arising from its development process are also complex and diversified. At the policy level, major countries and regions have prioritized security governance in their AI strategies; At the industry level, as the developers of AI technology, enterprises are important participants



in AI security governance. By participating in the formulation of standards and launching products and services, enterprises strive to explore the technical paths and solutions of AI security governance. Many domestic and foreign enterprises, including Qihoo 360, have provided various solutions of technology empowerment, industry regulation and platform monitoring through their own practices.

AI security is considered fundamental to public security in the digital age. While there is currently no mature solution for AI security governance, only best practices call for extensive collaboration and open-source innovation. In order to meet the challenges faced by AI security, 360 leverages the advantages of leading enterprises. It fulfills social responsibilities, undertaking the task of building a national-level open innovation platform. For example, since 2019, 360AI Security Lab has been devoted to constructing the “Security Brain - National New Generation Artificial Intelligence Open Innovation Platform”, dedicated to reducing the security risks of artificial intelligence, improving the innovation environment, and adequately implementing the innovative achievements in cybersecurity. Furthermore, the platform will empower small and medium-sized start-up security enterprises, vertical industries, and the artificial intelligence industry to improve the country’s overall AI security defense capability and build a “safe base” for AI. We believe industrial development will surely help artificial intelligence become a strategic technology combining technological innovation and security controllability.



It is worth mentioning that this report gets the participation and assistance of many units and benefits from the cooperation of the industry, the university, and the research institute. We would like to express our gratitude to Tianjin Institute for Digital Security on Smart City, 360 AI Security Laboratory, and National Engineering Research Center for Big Data Collaborative Security Technology for providing first-hand research materials. We also want to thank Business Research and Development Center of ICBC, Finance Laboratory of Citic Foundation, China Telecom Corporation Limited Research Institute, Shanghai Association for AI and Social Development, and Shanghai Cyber Research Institute for providing the research team with favorable opportunities for research. I believe that through the close integration of policy community, academia, and industry adhering to the principle of “technology for good”, AI technology will undoubtedly continue to benefit mankind.

In addition, there remain flaws in our report, and public criticism is deeply appreciated. We also look forward to conducting research collaborations with more partners.

Dr. Du Yuejin, the Vice President and Chief Security Officer of 360

June 2022



# Table of Contents

## Abstract

The Global Governance of Artificial Intelligence: Safety and Security Perspective.....	1
---	---

## I. Security Governance Emerged as a Priority for AI Strategy/2

(i) AI security governance in the US.....	3
(ii) AI security governance in the EU.....	3
(iii) AI security governance in China.....	5

## II. Challenges to AI Security Governance/7

(i) challenges to AI's own security.....	7
(ii) the coexistence of opportunities and risks of AI-enabled cybersecurity.....	9
(iii) challenges to AI-derived security.....	11
Case 1: challenges to digital city development brought by AI.....	12

## III. Risk Governance of AI/15

(i) AI risk governance philosophy.....	15
(ii) AI risk governance model.....	18
(iii) AI extraterritorial governance experience.....	21

## IV. Industry Practice of AI Security Governance/25

(i) technological solutions.....	25
(ii) industry standards.....	27
(iii) application cases.....	28

## V. Conclusion/32



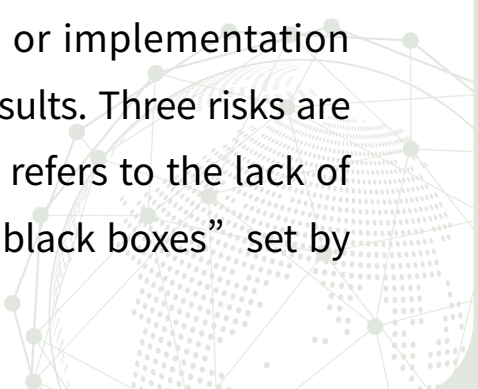
# Abstract

Security governance is critical to AI's future development. Nowadays, governments worldwide have taken security governance as an indispensable sector for their AI strategies. Globally, the exploration of ethical norms, risk frameworks, and governance concepts and models for AI has become a priority in academia and policy circles.

## 1.Challenges to AI Security Governance

Overall, there are ten aspects of AI risks, including cybersecurity, corporate compliance, interpretability, privacy, reputation and ethics, job prospects, fairness, personal safety, national security, political stability, etc. In the current stage of the technology development, the disputes are basically related the above ten aspects that can be further explained in three ways.

The first part is the safety of artificial intelligence, also known as AI system security, mainly includes the safety of AI framework, infrastructures, and algorithm, as well as the security of training datasets and protection of model files. Among them, algorithm security is especially worth noticing, for errors in design or implementation may generate unexpected, or worse, hazardous results. Three risks are of concern: First, the issue of “transparency” that refers to the lack of supervision and review arising from “algorithmic black boxes” set by





businesses or individuals who own decision-making algorithms, which not only poses security risks, but also is irreconcilable with the principle of interpretability of AI, and further, even becomes a controversy concerning political legitimacy at a higher level; Second, the “bias”, which refers to biased data sets and decision rules that lead to errors in AI training results; Third, “automation,” which refers to the pervasiveness of algorithms as a governance instrument that raises common concerns about its impact on human agency and autonomy.

The second part is challenges to AI-derived security. The application of AI technologies is expected to have a substantial impact on the economy, politics, military, and social matters. In Military field, “As AI technology matures, it will be used more and more extensively in military, including weapons systems, military strategies, military organizations; even the implications of war may change substantially. Human society will likely witness a different military security environment upon the coming of the AI era.” In politics, AI technologies and the big data and algorithms underpinning them can subtly influence human behavior, interfering directly with domestic political behavior and even influencing the substance and pattern of global competition; In economy, under the influence of AI technologies, the role of capital and technology in economic activities gains an overall increase, while the value of labor is severely diminished, leading to the risk of institutional unemployment, wealth polarization, and inequality.

The third part is the coexistence of opportunities and risks of AI-enabled cybersecurity. With the development of AI, such technology that can provide real-time monitoring based on massive data training seems to



offer new access to solve cybersecurity problems. At the same time, AI can bring new challenges to cybersecurity if used for cybersecurity attacks.

## 2.AI risk governance model

Faced with such a complex problem, AI security can not be achieved overnight.

Given that, a scientific and practical risk assessment and identification model should be established in three ways:

The first one is the impact assessment model based on future risk prevention. It refers to preventive supervision of potential hazards through multi-stakeholder consultation, participation, and consideration of the technological content that may have significant safety risks, and impact assessment before proceeding to practice; The second is one meta-regulation, which refers to those regulatory agencies, such as the government, do not set a strict compliance framework for enterprises, but leave it to the enterprises to justify the compliance of development activities, and the government agencies will investigate and penalize accordingly; Finally, we can also apply AI system vigilance model to early spot the problems and malfunctions and prompt error corrective work through systematic vigilance and transparent tracking of adverse incidents.

Moreover, innovative governance models are required. There are two patterns worth mentioning: the first one is the participatory design model, which is also called participatory governance, is the inclusion of stakeholders (e.g., expected end-users) in the design process, working together with professional designers and researchers, and participating



in decision-making; The second one is agile governance which aims at building an integrated AI governance ecosystem with all stakeholders through systematic government integration and to promote the iteration of governance policies through a sensitive, prompt, and continuous “consultation-feedback” mechanism to redress the lag of information in public governance and form a forward-looking assessment and governance of AI risks. In principle, agile governance consists of the three elements, including two “consultation-feedback” paths and a “dynamic evaluation” mechanism, collectively working on governance policy updates.

### 3. Industry Practice of AI Security Governance

All parties have listed the AI industry as a key sector for the development of strategic emerging industries and intensified support for the development of AI industry. Businesses are both developers of AI technologies as well as important actors in AI security governance. They enthusiastically take part in setting standards, introducing secure products and services, and taking AI security governance as an opportunity for industrial development.

In the future, all stakeholders such as the government, technology companies, academic institutions and users need to collectively adhering to the concept of community with a shared future for mankind, establishing positive ecological rules, strengthening multilateral interactive cooperation, building an open ecosystem of artificial intelligence, and directing the AI development with “AI for Good” principle to benefit human beings.



# **The Global Governance of Artificial Intelligence: Safety and Security Perspective**

As a strategic technology, artificial intelligence (AI) is demonstrating its profound impact on economic development, social progress and human life, making itself a matter of great interest for all countries strategically. AI, being a digital technology, is faced with security threats and potential risks. Moreover, as engineering-enhanced, scenario-contingent, platform-based AI becoming a reality, AI security demands more than a simple answer which only addresses technological matters. Naturally, AI security governance is a prioritized agenda of major countries and regions around the globe for their AI strategies, by which all participants hope to find a road that leads to a balance -- leverage the strengths of AI technology without loosening the grip on security risks.

## **I. Security Governance Emerged as a Priority for AI Strategy**

Security governance is regarded as a top priority in national AI strategies across the globe, with major countries seeking to avoid the security risks and challenges that accompany the encouragement of AI progress. “Globally, 444 AI development plans have been released by central governments in 61 countries, according to the Organization for Economic Cooperation and Development (OECD)<sup>1</sup>. Major global powers, with their actual needs on industrial development in mind, are reorienting the policy mixes to the build-up of ethics on technologies and regulations.” Early starters in formulating AI strategies, including the US, the EU, and China, have embarked on the exploration in the world of security governance.

---

<sup>1</sup> OECD, “OECD AI’s Live Repository of over 260 AI Strategies & Policies,” <https://oecd.ai/en/dashboards>.

**Table 1-1 Major countries' national AI strategies**

Document Name	Main Content	Released at	Released by
Ethical Principles for Artificial Intelligence	Responsible, Fair, Traceable, Reliable, and Controllable; calls for increasing DoD investment in AI research, training, ethics, and evaluation.	October 2019	DoD (USA)
Guidance for Regulation of Artificial Intelligence Applications	The public trust in AI, public participation, scientific integrity and information quality, risk assessment and management, flexibility, benefits and costs, fairness and nondiscrimination, openness and transparency, safety and security, and inter-institutional collaboration.	January 2020	U.S. White House
Ethics guidelines for trustworthy AI	Compliance with laws and ethics; respect for human freedom and agency; human regulation; avoidance of harm, fairness, stability and reliability; protection of privacy; transparency and interpretability; auditability; accountability; and ensuring social well-being.	April 2019	AI HLEG
AI in the UK: ready, willing and able	Ensure the common good of humanity, guarantee fairness, easy understanding, protect privacy, popularization, and avoid harming and deceiving humans.	April 2018	House of Lords, UK
The Japanese Society for Artificial Intelligence Ethical Guidelines	Contribute to humanity, abide by the law, respect privacy, be fair and just, ensure security, and be socially responsible.	February 2017	JSAI
The German Ethics Code for Automated and Connected Driving	Ensuring the safety of traffic participants, official approval for regulation required for driving systems, prohibition of personal attributes as evaluation criteria, prohibition of quantifying the value of life, and shared responsibility.	August 2017	Federal Ministry for Digital and Transport, Germany
The National Strategy for Artificial Intelligence	Transparency in algorithms, responsibility, establishment of AI ethics committee, organizing of public debates on ethics.	March 2018	French Government

**Source:** Collated from the OECD's website on AI National Policies and Strategies.

## (i) AI Security Governance in the US

The U.S. approach to AI security governance is to put the whole process, including arrangement, application and monitoring of AI technologies under validation and supervision with a supporting regulatory system.

In the arrangement phase, emphasis is placed on improving the interpretability and transparency so that imperfect decisions due to technical barriers can be reduced, and allows staff who are not tech-savvy to understand how it works and propose suggestions for improvement. In the meantime, a trusted input database is established to lower the bias of AI in its decision-making process.

In the application phase, the emphasis is placed on ensuring its verifiability and confirmability; meeting formal specifications and operational needs of users, and allowing them to operate the extensive and complex AI systems in a visible manner; output in a user-acceptable form, and run according to expectations of users. By doing so, AI will form a transparent, trustworthy, and reliable interaction method.

In the regulation phase, emphasis is placed on the establishment of targeted development standards and evaluation methods to properly verify AI.<sup>2</sup> In this regard, the emphasis is also on continuous updating of the technology, i.e., enhance the security and optimization of AI through self-monitoring, restrictive policies, and value learning so as to create auditable and recoverable AI systems. Based on this effort, it can address potential noise pollution and “anti-machine learning” to prevent other countries from trying to hinder the accurate identification of a target by “polluting” training data, tampering with algorithms or resorting to other means, by which they cannot get away with harming AI systems.<sup>3</sup>

## (i) AI security governance in the EU

Unlike the US, the EU is more inclined to place hopes on leveraging a regulatory

---

<sup>2</sup> Networking & Information Technology Research and Development Subcommittee and The Machine Learning & Artificial Intelligence Subcommittee of The Nation AI Science & Technology Council, “Artificial Intelligence And Cybersecurity: Opportunities And Challenges,” 2020, p.1-4, <https://www.nitrd.gov/pubs/ai-cs-tech-summary-2020.pdf>.

<sup>3</sup> National Science and Technology Council, “The National Artificial Intelligence Research And Development Strategic Plan,” 2016, p.27-30, [https://www.nitrd.gov/PUBS/national\\_ai\\_rd\\_strategic\\_plan.pdf](https://www.nitrd.gov/PUBS/national_ai_rd_strategic_plan.pdf).

framework and trust system to regulate the AI security, and the human rights consideration carries more weight in such regulation. The EU believes that it must see through critical changes in 7 dimensions in the European AI development, and has proposed independent directions for development in such 7 dimensions respectively.

**Table 1-2 Major initiatives and contents of the EU AI strategy**

Key requirements for trustworthy AI	Targeted measures
Human agency and oversight	Ensure that AI does not compromise human autonomy
Technical robustness and safety	Integrated security design mechanism
Privacy and data governance	Ensure privacy and data protection while using quality AI systems
Policy transparency	Require that AI systems are traceable
Diversity, non-discrimination and fairness	Establish diversified design teams and create mechanisms to ensure citizen participation
Societal and environmental well-being	Encourage sustainability and ecological responsibility of AI systems
Accountability	Establish mechanisms to ensure responsibility and accountability for AI systems and their outcomes

**Source:** Collated based on the EU’s Building Trust in Human-Centric Artificial Intelligence. <sup>4</sup>

In February 2020, the EU published a White Paper on Artificial Intelligence: A European Approach to Excellence and Trust, which draws up a risk-based approach to regulation. An AI application will be regarded as “high risk” when it meets two criteria: a major risk event is expected when applying it to a certain industry; the way to apply such AI application may bring significant risks. Additionally, when a case involves employee rights/intrusive surveillance technologies, it will always be regarded as “high risk” and subject to such level of regulation. Specifically speaking,

<sup>4</sup> European Commission, “Building Trust in Human-Centric Artificial Intelligence,” 2019, <https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52019DC0168&qid=1650694295419>.

the risk-based regulatory approach focuses on training data, the way of recording and keeping data, applying the required information, the accuracy of AI, and the level of human supervision, with some extra requirements for certain AI applications. This approach provides sufficient protection for the AI trust system without being too rigid, in which case it may impose an unnecessary burden on businesses.<sup>5</sup>

Meanwhile, the EU holds the view that the advent of AI makes the current legal system inadequate in terms of its enforcement, applicability, allocation of operators' responsibility, and security ideas. In view of this, the EU has made two adjustments to the existing legislative framework on AI. First, the current legislation on product safety should be extended to prevent such products from generating various risks, and to encourage innovation while protecting users. Second, in terms of the regulatory framework, a more flexible one should be established so that it can respond to latest technological updates and provide the necessary legal certainty.

### (iii) AI security governance in China

On July 8, 2017, Chinese government released *A [New] Generation Artificial Intelligence Development Plan*. In 2019, China released the *Governance Principles for a New Generation of Artificial Intelligence: Develop Responsible Artificial Intelligence*, which clarifies its AI governance framework and action guidelines. China's AI security governance strives for a whole-process security mechanism<sup>6</sup> that includes research and development, management, application, covering the development of fundamental framework, basic security principles, practical guidelines for supply chain management, security service capabilities, and standards development for application.

For cybersecurity, China has built a classification system for AI security threats,

<sup>5</sup> European Commission, "WHITE PAPER On Artificial Intelligence - A European approach to excellence and trust," 2020, <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0065&qid=1650694043168>.

<sup>6</sup> Standardization Administration of China (SAC), Cyberspace Administration of China (CAC), National Development and Reform Commission, National Development and Reform Commission (NDRC), Ministry of Science and Technology, and Ministry of Industry and Information Technology: "Guide to the Construction of a National New Generation Artificial Intelligence Standard System," July 2020.



developed a security assessment criteria system for AI systems, and incorporated various security factors into consideration, including interpretability and privacy to deal with possible attacks and pollution.<sup>7</sup> It also incorporates the key points of AI technology into the security governance. For algorithms, data and models, China has clear rules, which encompasses the regulations on utilization management of anonymous user data, AI data security, AI data annotation security, AI algorithm model trustworthiness and others.

China also attaches great importance to the social and ethical implications of AI. Code of Ethics for New-generation Artificial Intelligence, released in September 2021, underlines the integration of ethics into the whole life cycle of AI with the aim of promoting fairness, justice, harmony and safety; avoiding problems such as prejudice, discrimination, and privacy and information leakage; providing ethical guidelines for natural persons, legal entities and other related institutions who/which are engaged in AI-related operations.<sup>8</sup>

---

<sup>7</sup> Ministry of Industry and Information Technology of the People's Republic of China: "Three-Year Action Plan for the High-Quality Development of the Cybersecurity Industry. (2021-2023). (Draft for Solicitation of Opinions)," July 2021, [https://www.miit.gov.cn/cms\\_files/filemanager/1226211233/attach/20217/0e5071815ec641be9e2154566c09fe33.wps](https://www.miit.gov.cn/cms_files/filemanager/1226211233/attach/20217/0e5071815ec641be9e2154566c09fe33.wps).

<sup>8</sup> Ministry of Science and Technology of the People's Republic of China: "Code of Ethics for New-generation Artificial Intelligence" , September 26, 2021, [http://www.safea.gov.cn/kjbgz/202109/t20210926\\_177063.html](http://www.safea.gov.cn/kjbgz/202109/t20210926_177063.html).

## II. Challenges to AI Security Governance

Still a kind of digital technology, AI is obviously a “double-edged sword”. We have seen an upsurge in promoting the development of AI internationally while the various security risks it may contain are also taken seriously. In a word, AI security covers three dimensions, including the AI Safety, related security issues arising from AI, and AI-enabled cyber security.

### (i) challenges to AI Safety

The safety of artificial intelligence, also known as AI system security, mainly includes the safety of AI framework, infrastructures, and algorithm, as well as the security of training datasets and protection of model files.

Main manifestations of AI infrastructure risks are malicious or incomplete development from developers, immature product service systems, insufficient response capabilities, and imbalanced global infrastructure capacity building among different regions. Taking algorithmic frameworks as an example. An algorithmic framework refers to a deep learning based framework that AI rely on for development, whose security risks are manifested in its vulnerabilities and backdoor loopholes. The OpenCV (Open Source Computer Vision) developed by Intel, which supports facial recognition technology development for Google, Yahoo, Microsoft and other tech giants, has been detected to have two buffer overflow vulnerabilities in its version 4.1.0.<sup>9</sup> As these vulnerabilities are embedded in the underlying infrastructure framework, they have a huge impact on the following R&D and commercialization. As a result, they must be recognized and addressed quickly. In addition, there are also security risks in hardware, such as in CPU, and in cloud platforms which all fall under the infrastructure security of AI.

In terms of algorithm security, errors in design or implementation may generate

---

<sup>9</sup> “OpenCV XML Persistence Parser Buffer Overflow Vulnerability,” Talos Vulnerability Report, 2020, [https://talosintelligence.com/vulnerability\\_reports/TALOS-2019-0852](https://talosintelligence.com/vulnerability_reports/TALOS-2019-0852).

unexpected, or worse, hazardous results. Three risks are of concern:

First, the issue of “transparency” that refers to the lack of supervision and review arising from “algorithmic black boxes” set by businesses or individuals who own decision-making algorithms, which not only poses security risks, but also is irreconcilable with the principle of interpretability of AI, and further, even becomes a controversy concerning political legitimacy at a higher level; Second, the “bias”, which refers to biased data sets and decision rules that lead to errors in AI training results. For example, MIT researchers and Microsoft scientists have conducted tests on the face recognition systems of Microsoft, IBM and Megvii, and they have found that the error rate for white male subjects is less than 1%, while the error rate for black female subjects is as high as 21%-35%.<sup>10</sup> Third, “automation”, which refers to the pervasiveness of algorithms as a governance instrument that raises common concerns about its impact on human agency and autonomy.<sup>11</sup>

AI demands massive data sets as a resource to train machines with algorithms, creating risks to data privacy. The tension between the "data thirst" of AI and the “privacy awareness” of mankind is hard to avoid, and the concern of whether AI technology will compromise data and privacy has become a public worry. In a survey done by Microsoft, 41% of respondents said they do not trust smart voice assistants and think their privacy has been intruded upon through automated real-time voice collection. About 52% said they are worried that their personal information is not secured.<sup>12</sup>

Cooperation in AI research and development across institutions has increased further with the intensified data sharing between countries and the application of new AI technologies, such as federal learning and transfer learning. In recent years, in particular, the rapid rise of new security attack techniques such as adversarial examples attacks, backdoor attacks with algorithms, model stealing attacks, feedback

---

<sup>10</sup> The Technology and Standards Research Institute of China Academy of Information and Communications Technology (CAICT), Artificial Intelligence Security White Paper 2018, September 2018.

<sup>11</sup> Hildebrandt, Mireille, “The New Imbroglia – Living with Machine Algorithms,” 2016, p.55–60, <https://doi.org/10.25969/mediarep/13395>.

<sup>12</sup> Microsoft Advertising, “The 2019 Voice Report,” 2019, <https://about.ads.microsoft.com/en-us/insights/2019-voice-report>.

model misdirection, data reversion, and member inference attacks, individual privacy becomes more vulnerable to mining and exposure.<sup>13</sup> Cambridge Analytica, the main initiator of the data breach event that led to the Facebook market value evaporation of more than \$36 billion, was able to obtain a large amount of personal information on U.S. citizens exactly through correlation analysis, which was used to carry out various political campaigns and illegal profit-making activities. On Nov. 2, 2021, Facebook Inc. announced its plan to shut down its 10-year-old facial recognition system this month and delete the facial scan data of more than 1 billion users.<sup>14</sup>

## (ii) challenges to AI-derived security

The application of AI technologies is expected to have a substantial impact on the economy, politics, military, and social matters, resulting in the “Collingridge dilemma”<sup>15</sup> of technological development. It is vital for the future development of AI to be well positioned for potential upheaval before they are full-fledged, so that impact of the technological revolution will not backfire on humanity. “AI applications are reshaping existing threats and creating new types that can be exploited for activities such as community infiltration, data leakage, and disruption, deepening the threats presented by cyber attacks and disinformation campaigns.”<sup>16</sup>

Military-wise, “As AI technology matures, it will be used more and more extensively in military, including weapons systems, military strategies, military organizations; even the implications of war may change substantially. Human society

---

<sup>13</sup> Security Research Institute of China Academy of Information and Communications Technology: AI security framework, December 2020, <http://www.caict.ac.cn/kxyj/qwfb/ztbg/202012/P020201209408499730071.pdf>.

<sup>14</sup> Hill, Kashmir, and Ryan Mac, “Facebook, Citing Societal Concerns, Plans to Shut Down Facial Recognition System,” *The New York Times*, November 2, 2021, <https://www.nytimes.com/2021/11/02/technology/facebook-facial-recognition.html>.

<sup>15</sup> The Collingridge dilemma is the idea that the social consequences of a technology cannot be envisaged early in its life circle. By the time such technology has had its adverse consequences, it has often become such a deeply embedded part of the economic and social fabric that keeping it under control is already very difficult. Collingridge, D, “The Social Control of Technology,” Milton Keynes, UK: Open University Press, 1980, pp. 16-17; cited in WEN, Chengwei, and WANG, Zijun : “Analysis on the Path of Resolving the Collingridge’s Dilemma of AI by Anticipatory Technology Ethics,” *Journal of Dialectics of Nature*, Vol.43, No.4, 2021, p.10.

<sup>16</sup> QIN, Hao, “Analysis of the U.S. Government's Artificial Intelligence Strategic Goals, Initiatives and Experiences,” *Journal of China Academy of Electronics and Information Technology*, Vol. 12, No. 2021, pp. 1243-1250.

is likely to witness a different military security environment upon the coming of the AI era.”<sup>17</sup> Being the latest amplifier of power in the history of human technology, AI has showcased the potential to outperform humans by a large margin and continue to grow in the military affairs. International actors applying AI are very unlikely to be defeated in military engagements by adversaries who are not yet to employ AI technologies. In the wake of such a wave of technological change, all major countries equipped with the technological infrastructure will seek to acquire the know-how by all means, and a new arms race with AI technology at its core will be hard to avoid. Moreover, the involvement of AI technology makes it possible to employ a great number of unmanned combat weapons in operations, and the extensive use of fully autonomous weapons will pose a huge controversy on military ethics.

In politics, AI technologies and the big data and algorithms underpinning them can subtly influence human behavior, interfering directly with domestic political behavior and even influencing the substance and pattern of global competition. Take facial recognition and deepfake technology as an example. Right before the 2020 U.S. presidential election, a video of House Speaker Nancy Pelosi's speech was manipulated by deepfake technology, which was seen as a smear campaign against Democratic politicians. It is worth mentioning that when the video was identified as faked, the US online platforms reacted in different ways - YouTube promptly took down the video, but Facebook refused to remove it.<sup>18</sup>

In this case, the biometric identification of an individual involves privacy issues, the dissemination of false information on social media involves cybersecurity of online contents and personal reputation, and actions against government leaders may undermine political stability and national security, whereas compliance issues need to be addressed when dealing with the Internet giants as well as social platform operators.

In economy, under the influence of AI technologies, the role of capital and

---

<sup>17</sup> FENG, Shuai and LU, Chuanying, “National Security in the Era of Artificial Intelligence: risks and governance,” *Information Security and Communications Privacy*, Vol. 10, No. 1, 2018, p. 36.

<sup>18</sup> “Faked Pelosi Videos, Slowed to Make Her Appear Drunk, Spread across Social Media,” *Washington Post*, May 24th 2019, <https://www.washingtonpost.com/technology/2019/05/23/faked-pelosi-videos-slowed-make-her-appear-drunk-spread-across-social-media/>.

technology in economic activities gains an overall increase, while the value of labor is severely diminished, leading to the risk of institutional unemployment, wealth polarization and inequality. Moreover, the global economic reshuffle brought about by AI technology will lead to a greater flow of global capital and talents to technologically advanced countries, thus leaving developing countries with more restricted opportunities in their pursuit of modernization.<sup>19</sup>

As the structural changes in the society's economy driven by AI technologies gradually take hold, the monopoly of capital and technological power is likely to combine to diffuse, to some extent, important powers such as in finance and information traditionally held by the nation states.<sup>20</sup> For example, the large corporations' power over data resources and AI technologies is creating a de facto monopoly. Such a monopoly will be deeply embedded in every aspect of the digital age, including the use of black boxes of algorithms to push the public with targeted information, implicitly hanging the way public goods and services are provided.

### (iii) the coexistence of opportunities and risks of AI-enabled cybersecurity

And with the development of AI, such technology that can provide real-time monitoring based on massive data training seems to offer new access to solve cybersecurity problems. At the same time, AI can also bring new challenges to cybersecurity if it is used for cybersecurity attacks.

Under the current technological features and progress, AI will function in promoting cybersecurity in the following five aspects: 1) identifying system vulnerabilities to improve system credibility; 2) conducting automated online operations and cyber defense to offset attacks; 3) enabling real-time collection and analysis of cybersecurity intelligence; 4) automated decision making; and 5) human-

---

<sup>19</sup> FENG, Shuai and LU, Chuanying, "National Security in the Era of Artificial Intelligence: risks and governance," *Information Security and Communications Privacy*, Vol. 10, No. 1, 2018, p. 33.

<sup>20</sup> FENG, Shuai and LU, Chuanying, "National Security in the Era of Artificial Intelligence: risks and governance," *Information Security and Communications Privacy*, Vol. 10, No. 1, 2018, p. 34.

computer interaction.<sup>21</sup> These functions of AI can not only help solve cybersecurity challenges, but also enhance our integration with cyberspace and maximize the value of the network.

However, AI technology abuse can also increase cyber risks and bring deeper security threats. One of the major new problems is the generation of generative adversarial networks, a technique that can bypass AI monitoring systems and output any result based on the intent of the user.<sup>22</sup> Besides, the development and application of AI technologies also rely on cyber security. Just imagine the devastating aftermath if, in the coming future, the programs of automated weapons are hacked to modify the targets. As such, cyber security is integral to the technological security of AI.

### Case 1: challenges to smart city development brought by AI

From the perspective of social development, currently, applications of AI technologies are mainly focused on city governance and public services. Globally, smart city investment reached \$114.4 billion in 2020, and investment in smart cities in China reached \$24.1 billion and is expected to exceed \$45 billion in 2024, according to IDC.<sup>23</sup> In 2020, the main customers in China's AI market came from urban governance and administration (e.g. public security, traffic police, justice, public services, administration, epidemic prevention and control, transport management, land resources, environmental protection, etc.), accounting for 49%, followed by the Internet and financial industries, with 18% and 12% respectively.<sup>24</sup>

Since 2016, with the deep integration of 5G, AI, big data and other

<sup>21</sup> Please refer to “Artificial Intelligence and Cybersecurity: A Detailed Technical Workshop Report,” The Networking & Information Technology R&D Program (NITRD), 2020, <https://www.nitrd.gov/pubs/AI-CS-Detailed-Technical-Workshop-Report-2020.pdf>.

<sup>22</sup> S. Mathew Liao, *Ethics of Artificial Intelligence*, Oxford University Press, 2020, p.221.

<sup>23</sup> IDC: “China's Smart City Development Continues with High Quality in a Stable and Controlled Pandemic Prvention,” December 2020, <https://www.idc.com/getdoc.jsp?containerId=prCHC47212520>.

<sup>24</sup> iResearch: “Serial Market Research on China's AI Industry, 2020”, 2020, <https://www.iresearch.com.cn/Detail/report?id=3707&isfree=0>

technologies into the building of smart cities, AI cities are increasingly becoming the new models of smart city development. Smart transport, smart security, smart health care, smart parks and other concepts have emerged and rapidly taken root. AI and “the city brain” – kernel of digitalized city governance -- is increasingly becoming a top priority in the building of smart cities. However, though being a great help in city development and governance, AI inevitably brings new risks and challenges.

**First, AI may lead to excessive surveillance.** With the widespread and all-around use of monitors, panoramic cameras and other sensors, surveillance at public places may reach every corner of a city. Around the globe, video surveillance generated about 18.1 PB of data in 2020, accounting for 83.1% of the IoT data volume in the same year, making up the majority of IoT data, according to IDC's calculation.<sup>25</sup> While AI provides city governance with the convenience of “God's perspective”, it may also create a regulatory dilemma called “panopticon”, making residents feel insecure that they are being “monitored 24/7” while the city is unprecedentedly well protected. As a result, the citizen’s happiness index and city vitality may drop.

**Second, AI may encroach on personal privacy.** The extensive integration of AI into the building of smart cities and “city brains”, represented by deep learning, is inevitably leading to an unlimited demand for large amounts of data, especially personal data. From personal identity information to virtual information, from life footprints to consumption records, from individual lifestyles to biometrics, people living in smart cities will certainly become the targets of established AI services, and AI technologies will definitely collect as

<sup>25</sup> IDC: The premier global market intelligence company, “IDC: New Horizon for Security - Safeguarding the Terminal-to-terminal Security of Video Surveillance Systems” , 2021, <https://www.idc.com/getdoc.jsp?containerId=prCHC47327821>.



much personal information as possible while satisfying personalized needs. In the cycle of collection, analysis and updating of massive personal information, it is hard to avoid data abuse, leakage, black boxes, and other invasions of personal privacy.

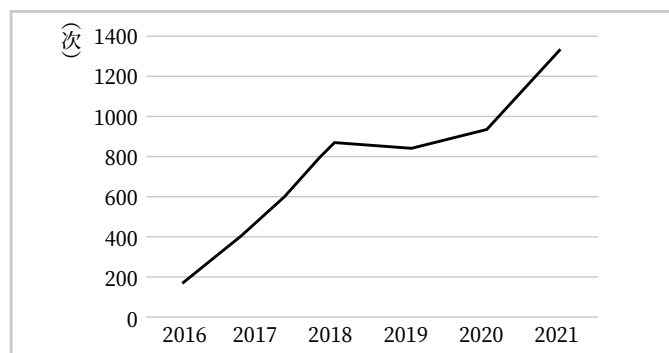
**Third, the security dilemma of smart city development.** With the extensive application of depth perception technology, AI constitutes the nervous system of the smart cities; and with the technical support of deep learning and big data analysis, it further forms the central “brain” in the process of digital transformation of cities. However, like any other system, it has vulnerabilities, and the more advanced a system is, the more vulnerabilities it may have. In the future, seizing a city may no longer need a long-lasting war in the streets; by only capturing and control the city’s AI system, it can control the operation of critical infrastructure, the release of every message, and even monitor the footprints of every individual’s actions.

### III. Risk Governance of AI

As a general-purpose revolutionary technology, the value of AI is evident in all aspects such as politics, economics, society, and culture, and hence there is risk in all aspects. Therefore, security governance is also crucial to the future development of AI. Overall, there are ten aspects of AI risks, including cybersecurity, corporate compliance, interpretability, privacy, reputation and ethics, job prospects, fairness, personal safety, national security, political stability, etc.<sup>26</sup> The controversial technologies that have emerged in the current AI technology development are largely contained within the above ten areas of concern.

#### (i) AI risk governance philosophy

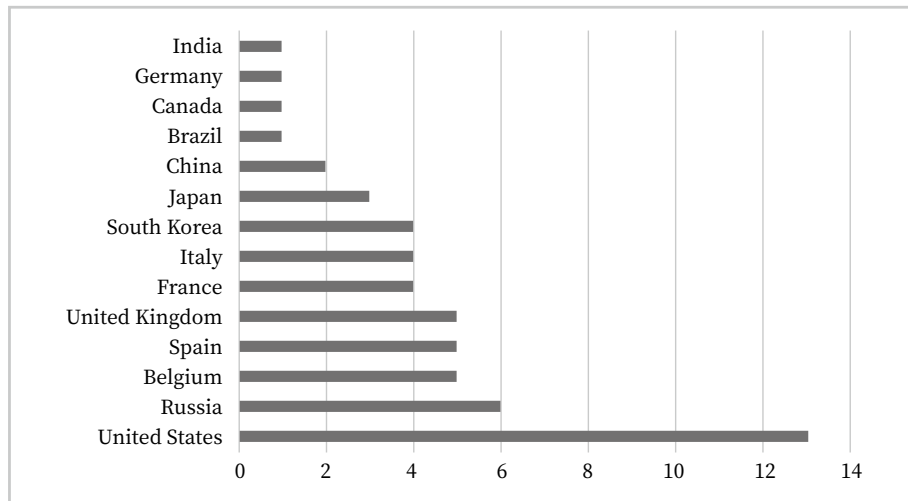
Among the many risks in the field of emerging technologies, privacy, personal safety and corporate compliance can all be regulated by specialized laws. The problem, though, is that it will be difficult to push ahead with a professional legislative process for AI in the short term. As presented in the figure, from 2016 to the present, nations are mentioning AI more and more in legislation (Figure 3-1), but the professionalized legislative process is still slow (Figure 3-2).



**Figure 3-1 Number of references to “artificial intelligence” in global legislations**

**Data source:** Stanford University “2022 AI INDEX REPORT” <https://aiindex.stanford.edu/report/>

<sup>26</sup> “Global Survey: The State of AI in 2020,” McKinsey, <https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/global-survey-the-state-of-ai-in-2020>.



**Figure 3-2 Number of professional AI acts in various countries**

**Data source:** Stanford University “2022 AI INDEX REPORT” <https://aiindex.stanford.edu/report/>

The legislative slowness is indicative of the fact that the current AI risk governance should be guided by regulation and operational principles for implementation; governance practices should be leveraged to build consensus and further facilitate the AI legislative process:

**A) impact assessment model based on future risk prevention**

Impact assessment refers to preventive supervision of potential hazards through multi-stakeholder consultation, participation and consideration of the technological content that may have significant safety risks, and impact assessment before proceeding to practice. The impact assessment mainly involves privacy impact assessment (PIA), data protection impact assessment (DPIA), social impact assessment (SIA) and ethical impact assessment (EtIA). Of these, the data protection impact assessment is introduced by Article 35 in the GDPR as a compliance binding article for data controllers and a new element of the GDPR accountability package concerned.<sup>27</sup> This impact assessment model balances the factor of data at the basic level with the social ethics at the higher level, in the hope of applying the impact assessment to set the “responsibility brake” on the users who are dealing with state-

<sup>27</sup> GDPR.eu. “Data Protection Impact Assessment (DPIA),” August 9, 2018, <https://gdpr.eu/data-protection-impact-assessment-template/>.

of-art AI technologies. It is a preventive measure of risks, but it, to a certain extent, relies on the self-awareness of enterprise and other entities.

### **B) meta-regulation model based on the principle of autonomy**

Meta-regulation refers to those regulatory agencies, such as the government, do not set a strict compliance framework for enterprises, but leave it to the enterprises to justify the compliance of development activities, and the government agencies will investigate and penalize accordingly.<sup>28</sup> This model basically runs in sync with activities concerning AI technology, while the government takes a backseat to ensure maximum corporate autonomy. However, the weakness is that the efficiency and effectiveness of governmental regulation, playing a secondary role, will also be limited.

### **C) AI system vigilance model based on transparent tracking**

AI system vigilance refers to the early spotting of problems and malfunctions and prompt error corrective work through systematic vigilance and transparent tracking of adverse incidents.<sup>29</sup> In fact, this behavior of vigilance occurs after the implementation of technological activities, which is based on a future-oriented risk management philosophy and is conducive to the long-term stable and healthy operation of technological products. However, the regulatory costs are also higher, as a result.

Generally speaking, each of the above three governance approaches have its own unique advantages, but also shortcomings that must be addressed. This is because any of these three governance approaches only addresses the specific risk control before/during/after the technology activities, and does not provide a sophisticated governance model through the entire process.

---

<sup>28</sup> Please refer to: Simon, F. C., *Meta-Regulation in Practice: Beyond Normative Views of Morality and Rationality*, London: Routledge, 2017, p.2-17.

<sup>29</sup> Please refer to Dubber, Markus Dirk, Frank Pasquale, and Sunit Das, *The Oxford Handbook of Ethics of AI*, Oxford University Press, 2020, p.88.

## (ii) AI risk governance model

In recent years, countries have embarked on exploring comprehensive AI governance models. Currently, the “participatory governance model” proposed by academia and entrepreneurs, and the “Agile Governance” model proposed by the World Economic Forum in 2018, are gradually being applied effectively in the AI governance practices around the world. These integrated governance models will be helpful in controlling AI risks, coordinating entities at all levels, and providing examples and ideas for global technology governance.

### **A) participatory design model converges the contributions of different stakeholders**

The kernel of the participatory design model, also called participatory governance, is the inclusion of stakeholders (e.g., expected end users) in the design process, working together with professional designers and researchers, and participating in decision-making. This requires that users are made fully informed of the possibilities and limitations of AI technology first in order to make design advice and decisions, and engage in the entire process of risk governance. The governance model of direct inclusion of user participation aims at reducing the risks of AI, such as algorithmic bias and black box decision making. Within the complete participatory governance model, there are mainly four steps as follows:<sup>30</sup>

**First**, identifying the specific steps of participatory design. Because some steps in AI technology activities still require expertise in designing the framework and government direction in setting goals and baselines, it will be counterproductive to include them in the framework of participatory governance.

**Second**, inspiring the design ideas. Because of the diversity of entities involved in the participatory design and the absence of expertise of many users, this step

---

<sup>30</sup> For steps and features, please refer to: Zytka, Douglas, Pamela J. Wisniewski, Shion Guha, Eric P. S. Baumer, and Min Kyung Lee. “Participatory Design of AI Systems: Opportunities and Challenges Across Diverse Users, Relationships, and Application Domains.” CHI Conference on Human Factors in Computing Systems Extended Abstracts, 1–4. New Orleans LA USA: ACM, 2022. <https://doi.org/10.1145/3491101.3516506>.

usually requires some visits or workshops to explain the logic of the AI technology to the various designers, who in turn, design the AI technology conceptually and functionally by following an impact assessment approach similar to the one mentioned above.

**Third**, the selection and implementation of design philosophy. The participatory design model, in effect, provides a “field trip” for AI development and application, but the trip needs to be followed by selection and implementation of solutions by professionals.

**Fourth**, the assessment and review of results. After implementation, this AI technology activity then goes to the final step. Typically, the evaluation also follows the participatory design philosophy, in which users and professionals work together to assess whether the technological products meet the design philosophy and uses of the requirements, and whether it poses potential risks. On the other hand, regulatory authorities may review the technology activities from another angle for compliance with the national AI development direction and security requirements.

In summary, the participatory design model guides the entire process of technology activities. The user-centric governance model deals with the risks arising from interpretability, privacy, reputation and ethics and fairness, which are mentioned previously, making it a scientific model for future AI risk governance.

### **B) agile governance addresses technological uncertainty**

Compared to the user-centric governance model of participatory governance, agile governance incorporates the entire AI risk governance process with the government at its core.<sup>31</sup>

A key feature of the agile governance model is to build an integrated AI governance ecosystem with all stakeholders through systematic government integration, and to promote the iteration of governance policies through a sensitive, prompt, and continuous “consultation-feedback” mechanism to redress the lag

---

<sup>31</sup> XUE Lan, ZHAO Jing, “Towards Agile Governance: An Inquiry into the Development and Regulatory Model of Emerging Industries”, Chinese Public Administration, Vol. 8, 2019, pp. 28-34.

of information in public governance and form a forward-looking assessment and governance of AI risks. In principle, the agile governance consists of the three elements, including two “consultation-feedback” paths and a “dynamic evaluation” mechanism, collectively working on governance policy updates.<sup>32</sup>

The first element is the “consultation-feedback” path based on the AI “governance principles” outlined by the government and the “factual characteristics” of the AI technology used by the stakeholders. In the agile governance model, instead of clear and strict AI governance rules, it depends on a framework of principles and works with multi-stakeholders who have market and technology information through this “consultative-feedback” path so as to ensure that the governance principles are set in a scientific manner.

The second element is the “consultation-feedback” path based on AI “governance goals” of the government and “market demand information, values and fairness” of the stakeholders. Similar to the governance principles, the second element that the government needs to design under agile governance is the AI governance objectives so as to outline the path and direction of the overall governance activities. Moreover, it needs to meet the essential market needs and spiritual pursuits of all stakeholders to be sustainable.

The third element is a dynamic assessment mechanism. Apart from two “consultation-feedback” paths, the government needs to establish a dynamic evaluation mechanism for assessing in advance whether the AI technology activities are in keeping with the principles, whether goals are met after the technology activities, and integrate the whole process. Furthermore, the agile governance model creates an iteration of policies by combining pilot technologies designed by companies with dynamic assessments by regulatory agencies so as to respond to technological and environmental changes while taking into account strategic goals. This provides a strong guide for dealing with the “uncertain” risks of AI technology itself.

---

<sup>32</sup> Please refer to “Agile Governance Reimagining Policy-Making in the Fourth Industrial Revolution,” World Economic Forum, January 2018. [https://www3.weforum.org/docs/WEF\\_Agile\\_Governance\\_Reimagining\\_Policy-making\\_4IR\\_report.pdf](https://www3.weforum.org/docs/WEF_Agile_Governance_Reimagining_Policy-making_4IR_report.pdf).

### (iii) AI governance experience in other countries

#### A) AI risk management framework in the US

In March 2022, the National Institute of Standards and Technology (NIST) released a draft Artificial Intelligence Risk Management Framework (AI RMF)<sup>33</sup> that addresses risks in the design, development, deployment, and assessment of AI systems to promote the development and deployment of trustworthy and responsible artificial intelligence (AI) technologies and systems.

AI RMF is developed through a consensus-driven, open, transparent, and collaborative development process by working in cooperation with the private and public sectors.<sup>34</sup> The framework has two objectives, one is to facilitate development and innovation that improve the trustworthiness of AI, including features such as accuracy, interpretability, reliability, privacy, robustness, security, resilience, and the mitigation of unintended and/or harmful biases and deleterious uses.

This framework follows a three-tiered category of characteristics. In order to identify and manage risks involved in AI systems, the integrated approach need to focus on three factors, including technological characteristics, socio-technological characteristics, and guiding principles. Among them, technological characteristics are factors under the direct control of AI system designers and developers, which can be measured with standard evaluation criteria, such as accuracy, reliability, and resilience. Socio-technical characteristics refer to the ways AI systems are used and perceived by individuals, groups, and the society, such as "interpretability", privacy, security, and management of discrimination. Through AI RMF classification method, guiding principles refer to broader social norms and values, reflecting social priorities such as equity, accountability and transparency.

---

<sup>33</sup> "AI Risk Management Framework Concept Paper." The National Institute of Standards and Technology (NIST), December 2021, [https://www.nist.gov/system/files/documents/2021/12/14/AI%20RMF%20Concept%20Paper\\_13Dec2021\\_posted.pdf](https://www.nist.gov/system/files/documents/2021/12/14/AI%20RMF%20Concept%20Paper_13Dec2021_posted.pdf).

<sup>34</sup> Federal Register. "Artificial Intelligence Risk Management Framework," July 2021, <https://www.federalregister.gov/documents/2021/07/29/2021-16176/artificial-intelligence-risk-management-framework>.



## **B) EU's AI regulatory framework**

The EU's AI Act follows a risk-based approach with four categories of specific risks: unacceptable risk, high risk, limited risk, and minimal risk.

**First**, unacceptable risk refers to those required to go over an explicit prohibition of harmful AI practices that pose a clear threat to human safety, livelihoods and rights, such as the exploitation of specific vulnerable groups, social scoring, and real-time biometrics for law enforcement purposes.

**Second**, high risk is assessed not only on the functionality of the AI system, but also on the specific purposes and means of using the system. The EU has identified two main categories of high-risk AI systems: AI systems that have been previously assessed and qualified by third-party agencies for use in security; and AI applications that are involved in biometrics, infrastructure management, education and vocational trainings, recruitment and workforce management, access to public services and essential private services, law enforcement, immigration and border control, administration of justice and democratic processes. These high-risk AI systems are granted access to the European market subject to certain mandatory requirements and prior conformity assessment.

**Third**, limited risk refers to those involve interaction with humans, detecting emotions or biometric-based analysis, generating or manipulating relevant content through technology, and the EU requires that such AI systems are obliged to inform people of the existence of such a situation. When an AI system is used to generate audio and video contents that are almost identical to real contents, it is also obliged to disclose that the contents are automatically generated under legitimate purposes.

**Fourth**, low/minimal risk refer to other AI systems that have low-risk and can be exempted from additional EU regulations. However, the EU is also creating codes of conduct to encourage such AI systems to voluntarily comply with mandatory requirements of the AI Act for their high-risk counterparts, or to create their own mandatory codes of conduct in order to protect the environment, ensure diversity, provide accessibility, etc.

The EU's AI Act enables AI to efficiently strike a balance in data protection and

utilization, thus aligning with the European General Data Protection Regulation (GDPR). The rapid development of AI would be impossible without the effective use of data. The enactment of the AI Act means that AI can play an active role in the collection and use of personal data, thus allowing for a more sensible use of data while protecting it more holistically. On the one hand, the employment of AI will help boost the collection and use of personal data, better comply with GDPR requirements, and advance the protection of personal data. On the other hand, regulating the uses of personal data within the framework of GDPR will also help to make the most of AI technologies.

### **C) UNESCO's first ever global agreement on the ethics of AI**

United Nations Educational, Scientific, and Cultural Organization (UNESCO) adopted the Recommendation on the Ethics of Artificial Intelligence (hereinafter short for “Recommendation” in this section) on November 25, 2021. With the Recommendation, UNESCO aims to provide a research framework that can be globally implemented in terms of values, principles, and actions, that can guide the private sectors and civil society, that can make the whole life cycle of AI consistent with ethics, and that can facilitate a pluralistic dialogue among global stakeholders to achieve progress in AI for the good of all nations and social classes. The Recommendation sets forth the following values:

#### **a) respect, protection and promotion of human rights and fundamental freedoms and human dignity**

The Recommendation argues that "No human being or human community should be harmed or subordinated, whether physically, economically, socially, politically, culturally or mentally during any phase of the life cycle of AI systems",<sup>35</sup> that human rights and fundamental freedoms need to be guaranteed, and that AI should promote better protection of such rights.

#### **b) environment and ecosystem flourishing**

---

<sup>35</sup> “Recommendation on the Ethics of Artificial Intelligence” , UNESCO, November 25, 2021, [https://unesdoc.unesco.org/ark:/48223/pf0000380455\\_chi](https://unesdoc.unesco.org/ark:/48223/pf0000380455_chi).

The Recommendation requires AI activities to be conducted under the principle of protecting the environment and promoting sustainable development. Governments need to evaluate the impact of AI on the environment during its life cycle and minimize its impact on climate change and environmental risks to prevent environmental and ecosystem degradation. AI, in turn, should benefit the environment and promote the environmental protection campaign.

#### c) ensuring diversity and inclusiveness

The Recommendation requires all governments to ensure that AI complies with international law and respects, protects, and promotes diversity and inclusion. Also, all countries should vigorously promote international cooperation, thereby bridging the gap in infrastructure, education, skills, and law vacuums brought about by the AI technology divide; countries should not take advantage of the current state of technological underdevelopment in other countries to commit acts damaging to the interests of other countries.

#### d) living in peaceful, just and interconnected societies

The Recommendation believes that AI actors should play a participative and enabling role to ensure peaceful and just societies, which is based on an interconnected future for the benefit of all. AI should promote a fair and inclusive interconnected environment that enables societies to form organic, immediate, uncalculated bond of solidarity.

The Recommendation continues with detailed policies in multiple fields including ethical impact assessment, ethical governance and management, data policy, development and international cooperation, environment and ecosystems, gender, culture, education and research, communication and information, economy and labor, health and social well-being; it also proposes holistic ethical principles to regulate AI. Member States have also been required to develop scientific, effective methods for monitoring and evaluating AI policies, plans and mechanisms within a broad, credible and transparent framework.

## IV. Industry Practice of AI Security Governance

All parties have listed the AI industry as a key sector for the development of strategic emerging industries and intensified support for the development of AI industry. Businesses are both developers of AI technologies as well as important actors in AI security governance. They enthusiastically take part in setting standards, introducing secure products and services, and taking AI security governance as an opportunity for industrial development.

### (i) technological solutions

AI application is related to healthcare, finance, retail, government governance, etc., with the technology penetration rising. From 2018, Google, Microsoft, IBM, Tencent, Ali, Baidu and other Chinese and foreign companies have rolled out their own AI governance solutions.

IBM has developed a collection of trusted AI critical technologies, such as the AI Fairness 360 (AIF360) that can be used to test and mitigate bias in machine learning models; the Adversarial Robustness 360 (ART) that can be used to quickly create and analyze attacks and defense methods for machine learning models; and AI Explainability 360 (AIX360) that can be applied to support the interpretability of machine learning models and algorithms.<sup>36</sup>

Microsoft has teamed up with MITRE, Bosch, IBM and many other corporations to launch a threat matrix to combat machine learning. Counterfit, an algorithm security attack and defense tool developed by Microsoft that can attack multiple AI models on a large scale, has become the main tool for security tests of Microsoft's AI operations.<sup>37</sup>

Tencent designed a set of automated attack methods for lane line system attacks

<sup>36</sup> IBM Research Teams, “Trusted AI,” February 2021, [https://research.ibm.com/teams/trusted-ai?\\_ga=2.264991951.1171655595.1652318757-354424505.1652318757](https://research.ibm.com/teams/trusted-ai?_ga=2.264991951.1171655595.1652318757-354424505.1652318757).

<sup>37</sup> Microsoft Security Blog, “AI Security Risk Assessment Using Counterfit,” May 2021, <https://www.microsoft.com/security/blog/2021/05/03/ai-security-risk-assessment-using-counterfit/>.

(i.e., deploying jamming information on the roads, causing vehicles to make wrong judgments about lane lines when passing by, resulting in vehicles driving into the reverse lane) based on black box testing and optimization algorithms in 2019.<sup>38</sup> In 2020, Tencent released the industry's first AI security attack matrix and showcased research results on AI model backdoor attacks. In 2021, Tencent released sound anti-cloning research results as well as a new idea to combat Deepfake, i. e. MagDR.<sup>39</sup>

In 2021, Ali released CAA, an automated AI counter-attack platform, and held the sixth session of CVP2021 Challenger Program jointly with Tsinghua University and UIUC; released AI security assessment benchmark platform jointly with Tsinghua University and RealAI.

In 2018, Baidu released AdvBox, an anti-attack open source toolkit, submitted 86 TensorFlow vulnerabilities in 2021, and held the first Autonomous Driving CTF Contest.

In 2020, Huawei released MindArmour, providing security and privacy protection capabilities for the MindSpore framework.

Qihoo 360 Group built the “Security Brain - National New Generation Artificial Intelligence Open Innovation Platform” in 2019, which was funded by “Science and Technology Innovation 2030 - New Generation Artificial Intelligence Major Project”, a project launched by Ministry of Science and Technology in June 2020, and has been cooperating with Tsinghua University, Institute of Automation of Chinese Academy of Sciences, Institute of Information Technology of Chinese Academy of Sciences, Beijing RealAI Intelligence Technology Co. The platform mainly consists of six parts, namely, multi-source heterogeneous security big data integration and governance, AI-enabled security, AI's own security, openness and sharing, open source community and APP STORE, security big data attack and defense competition and scientific research projects, and is dedicated to provide technological reserve for building AI open source software and hardware technology platform and intelligent

---

<sup>38</sup> Jing, Pengfei, Qiyi Tang, Yuefeng Du, Lei Xue, Xiapu Luo, Ting Wang, Sen Nie, and Shi Wu, “Too Good to Be Safe: Tricking Lane Detection in Autonomous Driving with Crafted Perturbations,” 3237–54, 2021. <https://www.usenix.org/conference/usenixsecurity21/presentation/jing>.

<sup>39</sup> Chen, Zhikai, Lingxi Xie, Shanmin Pang, Yong He, and Bo Zhang. “MagDR: Mask-Guided Detection and Reconstruction for Defending Deepfakes.” ArXiv:2103.14211 [Cs], March 25, 2021. <http://arxiv.org/abs/2103.14211>.

security detection platform.<sup>40</sup>

In particular, 360 AI framework vulnerability threat awareness system is able to continuously carry out security risk research on machine learning framework on multiple fronts, such as algorithm realization, vulnerability type, compilation optimization, etc., and adopt a combination of dynamic and static analyses to conduct comprehensive and systematic tests on different vulnerability types in different languages such as Python, C++ and Go. The detection process focuses on the security risks existing in the process from training to inference, from data to model, and from cloud to terminals.

### (ii) industry standards

In the aspect of industry governance, Chinese and foreign trade associations have developed industry standards and made them an essential part of industry self-regulation.<sup>41</sup>

**Table 4-1 Examples of Industry Standards for Artificial Intelligence**

Trade association' s name	Standards and related events
International Organization for Standardization (ISO)	In October 2017, it approved the establishment of the subcommittee of JTC 1/SC 42 Artificial Intelligence which focuses on standardization research in terminology, reference frameworks, algorithm models and computation methods, security and trustworthiness.
International Telecommunication Union (ITU)	Mainly dedicated to solving security problems in AI applications such as smart healthcare, smart cars, spam management, biometric identification, etc.

<sup>40</sup> State-level AI innovations' debuts! AIExpo2020 'New Generation Artificial Intelligence Open Innovation Platform' Summit Forum successfully held, Netease, August 16, 2020, <https://www.163.com/dy/article/FK5OQLU705118HA4.html>.

<sup>41</sup> "Two works of 360 selected as Representative Cases of Artificial Intelligence Security Practices", 360 Government and Enterprise Security, October 13, 2021, <https://www.360.net/about/news/article61679316eec939004a2dfae0>

Institute of Electrical and Electronics Engineers (IEEE)	Carried out several studies on AI ethics, published several AI ethics standards and research reports, and developed the IEEE 7000™ standards for ethical regulation of AI systems.
National Institute of Standards and Technology (NIST)	Published guidance on how governments can develop AI technology standards and related tools in 2019.
China National Information Technology Standardization Network	It has carried out standardization work mainly in AI terminology, human-computer interaction, biometric recognition, big data, and cloud computing.
China Academy of Information and Communications Technology	In the ITU-T SG16 plenary session, led the creation of a new topic - AI multimedia, deployed AI and multimedia integration research and standards output.
China Communications Standards Association (CCSA)	Published the Safety Capability Requirements for Smart Home Terminals, Safety Technical Requirements for Smart Home Network Security System and other standards.
Artificial Intelligence Industry Alliance (AIIA)	Published the Guidelines for Trusted AI Operations in 2019 and announced the first batch of trustworthiness assessment results for commercial AI systems, involving 16 AI systems from 11 companies, providing reference for user choices.
AI Open Source Software Development League (AIOSS)	Released evaluation criteria for products or services such as machine translation and smart assistants, as well as reliability assessment criteria for deep learning algorithms.

**Source:** collated from official documents of various trade associations. Search date: May 11, 2022.

### (iii) application cases

#### **A) automated cybersecurity AI company Darktrace's partnership with Microsoft**

In May 2021, Microsoft partnered with Darktrace, whose self-learning AI helps users of Microsoft Mail, Microsoft 365, Azure and other services cope with cyber threats, and enable integration with Microsoft Sentinel to help enterprises build security capabilities in multi-cloud and multi-platform environments.

Darktrace is a leading automated cybersecurity AI company and the founder

of automation-related technology. The technology integration between Microsoft and Darktrace will help Microsoft detect Zero-Day vulnerabilities and improve Web protection by parsing malicious IP addresses, domains and URLs. Specifically, self-learning technologies are employed to empower email security, components of Microsoft 365 and Azure, and are well integrated with the SIEM platform Sentinel.<sup>42</sup>

① AI Email Security. By learning the “behavior patterns” of each user in Microsoft 365 it identifies abnormal behavior in the environment. Darktrace is able to identify new email threats including sophisticated phishing, business email compromise (BEC) and supply chain attacks (or vendor email leaks).

② Microsoft 365. It serves as a self-learning technology that identifies cybersecurity threats in Microsoft 365 product components, including credentials leakage, administrator abuse, and malicious insiders. When Darktrace detects a network incident in Microsoft 365, it classifies, interprets and reports on the incident to help an organization quick respond to threats.

③ Microsoft Azure Cloud Security. Through self-learning technology, it provides insight into behavior in Azure cloud environments, places behavior in context, and detects deviations from “behavior patterns” to identify threats. Self-learning AI can automatically connect the dots among anomalous behavior in different infrastructures, making sure that cloud security is not isolated from monitoring in other parts of the organization.

④ Integration with Security Information and Event Management (SIEM). Darktrace's job logs allows security teams to send and visualize alerts and network events in SIEM. In addition, these can be grouped by activity type, and users can return to the Darktrace, the threat visualization tool, with a single click for further inquiry.

## **B) RealAI’s deepfake detection platform DeepReal**

RealAI was incorporated in July 2018 and incubated from the Institute for AI

---

<sup>42</sup> Please refer to Darktrace, “World-Leading AI for Cyber Security,” <https://www.darktrace.com/en/>.



Industry Research of Tsinghua University. Its deepfake detection platform, DeepReal, relies on the third-generation AI technology. It can quickly and precisely identify the authenticity of image, video and audio contents by identifying the representational differences between fake contents and real contents, and digging for the consistency features of deepfake contents from different generation channels so as to effectively crack down on property fraud, illegal business, false propaganda, evidence forgery and other illegal acts.<sup>43</sup>

This platform features “high performance and high accuracy detection capability of real network” and “cooperation in building a deepfake governance ecology”, and it is applied to scenarios such as network content compliance detection, face verification security, image evidence authenticity detection, and anti-infringement fraud.

Through deepfake detection algorithms, the platform detects contents such as images, videos and audios, monitors and identifies the contents to provide interpretable descriptions, and produces multi-dimensional detection reports. Technically, it is designed and developed based on Bayesian deep learning, multi-feature fusion and multi-task learning; it is trained based on more than 10 million datasets, with advanced accuracy and good robustness; its detection efficiency is up to 30 milliseconds per frame.

### **C) 360 Group’s digital twin-based IoT security attack and defense platform**

As a demonstration project under the critical technology category published the Ministry of Industry and Information Technology (MIIT) in April 2022, 360 Group’s “digital twin-based IoT security attack and defense platform” was listed.<sup>44</sup> The platform is designed to effectively tackle the cyber threats that may be met by the smart cities’IoT in the future; to improve the overall defense and emergency response

---

<sup>43</sup> “DeepReal, a deepfake detection platform” , RealAI, <https://www.real-ai.cn/products/9.html>. Accessed May 12, 2022. <https://www.real-ai.cn/products/9.html>.

<sup>44</sup> Ministry of Industry and Information Technology of the People's Republic of China: “Public Notification of IoT Demonstration Projects in 2021,” April 13, 2022, [https://www.miit.gov.cn/zwgk/wjgs/art/2022/art\\_81b88cf50fd144f19267faeca3a35c2c.html](https://www.miit.gov.cn/zwgk/wjgs/art/2022/art_81b88cf50fd144f19267faeca3a35c2c.html).

capability of cities, and escort the digital security of smart cities.

Smart cities are where all kinds of IoT gather; they are hubs of critical information infrastructure, industrial Internet, intelligent transportation, etc. They are also the preferences of cyber attacks and the main positions of digital security risk prevention. The platform has two major innovations, one is establishing a digital twin network of smart city IoT through the combination of digital twin technology and AI technology, thus supporting a more agile, more accessible and larger scale simulation of smart city IoT network, and realizing smart verification of the simulation attack and defense capabilities of regional smart city IoT device clusters.<sup>45</sup>

Second, to prompt the breakthrough in the IoT device automation vulnerability mining technology. Through intensive research on key vulnerability mining technologies, the platform is able to conduct static scanning and dynamic simulation of relevant devices' firmware and equipped with the capabilities of intelligent vulnerability mining in smart city scenarios; it supports mainstream processor architectures such as ARM, MIPS, X86, etc., supports mainstream network protocols. It keeps on with continuous research in smart automated vulnerability mining, and its comprehensive code coverage rate of fuzzy testing is leading in the industry.

---

<sup>45</sup> 360 Attack and Defense Platform Selected as a 2021 IoT Demonstration Project by the Ministry of Industry and Information Technology, Safeguarding Digital Security in Smart Cities” , cww.net, April 2022.

## V. Conclusion

As a critical emerging technology being developed in China, the United States and Europe, it is natural the security challenges arising from the development of AI are more complex and manifold. To address the challenges of AI regarding its own safety, empowerment security, derivative security, and digital city development security, the AI risk governance model has emerged as a result. Guided by operability as the main principle, it encompasses three major governance ideas: impact assessment, meta-regulation and AI system alertness, and, more specifically, it breaks down two paths: user-centered participatory design and agile governance with the government as the main force. Under the governance practices of all countries and organizations, the above models have, in turn, developed their own official regulatory frameworks for AI. With this, a number of leading AI companies, including 360 Group, have also set themselves as industry security establishment exemplars with their own practices. They have embarked on a variety of paths of technology empowerment, industry regulation and platform monitoring. We believe that all countries will ramp up the development of AI security industry and continue to help AI become a strategic technology that integrates technological innovation and security control.

© 2022 Research Center for Global Cyberspace Governance (RCGCG) & Beijing Qihoo Technology co., LTD  
. All rights reserved.

Qihoo Technology & RCGCG don't take institutional positions on public policy issues; the views represented herein are those of the author(s) and do not necessarily reflect the views of Qihoo Technology & RCGCG. No part of this publication may be reproduced or transmitted in any form or by any means without permission in writing from Qihoo Technology & RCGCG

## Beijing Qihoo Technology co., LTD

---

Email: [dipperresearch@360.cn](mailto:dipperresearch@360.cn)

Address: Build 2, 6 Haoyun, Jiuxianqiao Road, Chaoyang District, Beijing

Phone: 010-58781000

Fax: +86-10-56822000

## Research Center for Global Cyberspace Governance

---

Address: 195 -15 Tianlin Road, Xuhui District, Shanghai

Phone: +86-21-54614900

Fax: +86-21-64850100

