上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

CIDEG

国际与公共事务学院
School of International and Public Affairs

中央广播电视总台研究院
CHINA MEDIA GROUP INSTITUTE

西岸对话
West Bund Dialogue

AI Safety as Global Public Goods
Working Report

07.05

中国 · 上海

CONTENTS

Introduction

AI holds the potential for significant benefits,but it also poses systemic risks. The rapid pace of technological advancements in this field exacerbates the uncertainty surrounding its development. This phenomenon has garnered widespread attention from global stakeholders,who agree on the criticality of governance measures to address these risks and ensure trustworthy and responsible AI.

All stakeholders recognize the importance of continuously improving the reliability,controllability,and fairness of AI,strengthening AI safety research,and promoting cooperation on AI safety. AI safety plays a pivotal role in unlocking the transformative potential of AI while mitigating associated risks. In recent years,several important actions and processes have emerged within the international governance of AI to address these concerns. These include drafting scientific reports on AI safety,establishing benchmarks for evaluating AI safety,increasing investment in AI safety research,setting up dedicated institutions for studying AI safety and fostering dialogue on this topic.

We acknowledge the positive and significant role of multilateral and multi-stakeholder actions in advancing AI safety,while also recognizing potential areas for improvement in the current process. Drawing from global practical experience,we propose that in promoting AI safety practices,establishing AI safety regulations,and fostering consensus on AI safety,AI safety should be regarded as a “global public good.” This entails efforts to enhance public understanding of AI safety,strengthen capabilities in addressing safety concerns,and ensure access to necessary safety resources.

The concept of “AI safety as a global public good” holds value in its capacity to coordinate the dual objectives of development and safety. It goes beyond perceiving safety solely as a risk management measure or regulatory tool,instead emphasizing the accumulation and growth of safety knowledge,capabilities,and resources to foster the optimal advancement of AI. This concept facilitates a balanced approach that takes into account both domestic and international considerations,thereby guiding the reform of AI safety governance mechanisms and systems within different countries while also promoting international cooperation and global governance in this domain.

The notion of “AI Safety as Global Public Goods”

“AI safety as a global public good”refers to the notion that,given the unique characteristics of AI technology innovation,industrial application patterns,and the risks associated,AI safety should be viewed as knowledge,capabilities,and resources with public good characteristics. Various stakeholders – including governments,businesses,and third parties – need to collaborate in exploring,building,and sharing these elements of AI safety. This notion of a public good can effectively guide the development and societal implementation of the AI industry while ensuring equitable benefits for all individuals and humanity as a whole. Ultimately,it aims to foster an ecosystem where progress at all levels contributes to overall prosperity.

The concept of AI safety as global public goods posits that AI safety should be non-rivalrous in consumption and non-exclusive in benefits. Non-rivalry means that the use and consumption of AI safety knowledge,capabilities,and resources by new entities does not diminish their availability to existing entities. Non-excludability implies that while certain entities benefit from AI safety knowledge,capabilities,and resources,others are not excluded from also benefiting.

At the global level,there are relatively few international public goods that possess both non-rivalrous and non-exclusive characteristics. However,the unique nature of AI safety issues necessitates that we establish it as a global public good. The “black box” nature and dynamism of AI innovation,along with its widespread application across various domains,introduce new challenges and requirements for stakeholders involved in advancing AI safety governance.

First,unlike the governance of traditional technological risks which primarily focuses on information asymmetry,AI safety risks exhibit a distinct characteristic of “shared ignorance.” Technology innovators and users,government regulators,and affected communities all lack sufficient knowledge of AI safety risk.

Second,the cross-domain and transnational applications of AI have resulted in significant externalities in AI safety risks,leading to interdependence among different stakeholders. One party's safety must be built on the safety of others,and no party can achieve safety in isolation.

Third,the transformative potential of AI as a new frontier of human development tightly links development issues with safety concerns. The principle of balancing development and safety requires that AI safety governance be dynamically adapted and agilely integrated into the development process.

It is due to the novel characteristics and requirements of AI safety and AI governance that AI safety knowledge,capabilities,and resources have taken on the attrivutes of a ‘comedy of the commons. Similar to a ‘carnival’,the more participants involved,the more vibrant the atmosphere becomes,and the greater the joy and benefits each person derives from it. Consequently,ensuring the safety of AI necessitates open collaboration among multiple parties in order to create and maintain public goods that benefits both individual parties and the whole.

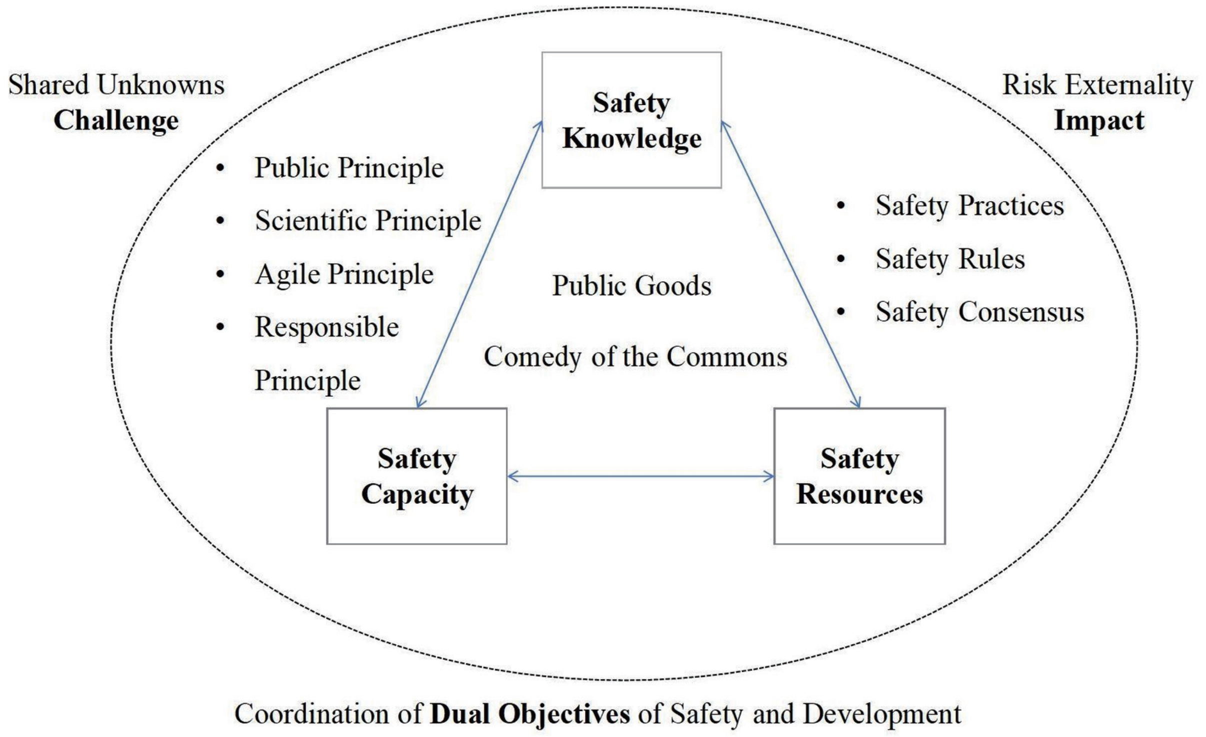

Fig.1 The Conceptual Framework Diagram of “AI Safety as Global Public Goods”

The principle of “AI Safety as Global Public Goods”

To achieve the goal of “AI safety as global public goods”,we should fully leverage the guidance,bridging,and catalytic roles of governments,the United Nations,and other intergovernmental international organizations,as well as other public institutions in the international community,on the basis of respecting multilateral and multi-party autonomous decision-making. We should promote the formation of a trustworthy governance environment in the field of AI safety,characterized by joint capacity building,shared risk bearing,collaborative framework consultation,and knowledge sharing. Such an environment aims to overcome collective action dilemmas and realize the sustainable production,maintenance,and reproduction of AI safety as a global public good.

The key principles for realizing the notion of “AI safety as global public goods” encompass four fundamental aspects: the public principle,scientific principle,responsibility principle,and agility principle.

Public Principle

This principle refers to the adherence to global public-oriented principles by multilateral and diverse stakeholders. In the process of AI governance,this principle emphasizes openness,inclusiveness,and cooperation. It involves cooperatively establishing credible governance institutions or cross-institutional collaborative mechanisms,and jointly sharing knowledge and experiences related to AI governance. This principle aims to promote the development of a safety management framework or tool that serves public interests,and to assist small and medium-sized enterprises as well as developing countries in enhancing their AI safety capabilities AI safety through various means.

Scientific Principle

This principle refers to following and respecting for scientific laws governing AI technology innovation and application. It involves mobilizing stakeholders such as enterprises,scientists,technology communities,regulatory agencies,and the public to collectively explore,learn,and accumulate knowledge on AI safety during the process of technological advancement. This collaborative effort aims to gradually enhance the level of consensus on AI safety while fostering capabilities and resources throughout this process.

Responsibility Principle

This principle entails the establishment of diverse governance frameworks and institutional arrangements for responsibility,each with varying levels of enforceability. It also involves developing corresponding requirements for responsibility systems based on specific safety risk levels. This principle encourages stakeholders to actively engage in constructing AI safety public goods that align with responsible concepts,ensuring that these goods are not abused or misused. Ultimately,it aims to foster a positive ecosystem in which responsible practices prevail.

Agility Principle

This principle pertains to the commitment of multilateral and diverse entities to form interconnected and interdependent governance relationships. Based on broad consensus,the entities should respond promptly to address AI safety risks while adapting to the dynamic and uncertain progress in AI development. This is achieved through feedback interaction and iterative innovation within the safety governance process.

The Global Value of “AI Safety as Global Public Goods”

“AI safety as global public goods” is a theoretical distillation and summary of experiences derived from China’s AI safety governance practices. It can serve as a valuable complement to domestic AI safety governance reforms in various countries and the process of international AI safety governance. This principle offers significant theoretical insights and practical guidance for enhancing global AI safety standards.

The current institutional models and reform processes for promoting global AI safety governance can be categorized into three main approaches. First,the National Institute of Standards and Technology (NIST) in the United States has proposed an AI risk management framework. This approach focuses on four dimensions: govern,match,measure,and manage. The key characteristic of this approach is to view AI safety as the subject of risk management. Second,the UK AI Safety Institute has developed a series of technical tools for AI risk management. These tools aim to address AI safety risks through technical solutions independently developed by stakeholders. The primary characteristic here is treating AI safety as a technical risk issue to be managed and addressed. Third,the European Union’s AI Act (AIA) includes various institutional arrangements such as risk classification management and a risk assessment system. Its fundamental characteristic is to regulate the AI safety as a target of supervision.

The three models hold positive significance and substantial value. In comparison,the complementary value of the concept of “AI safety as global public goods” lies in the fact that it does not only consider AI safety as an object a risk management,a technical solution,or regulatory target. Instead,it recognizes the novel characteristics and requirements of AI safety risks and their governance. Therefore,it emphasizes the importance of multilateral and multi-stakeholder exploration in terms of AI safety knowledge,cap abilities,resources,and other aspects. Within this framework,we place great emphasis on the process of governing AI safety guided by the creation of public goods to foster interdependent,trusting,and cooperative relationships among diverse entities.

Realizing the value of “AI safety as global public goods” requires the gradual enhancement of governance systems and mechanisms. This concept neither implies mandatory government regulations,nor does it represent leaving actors to their own devices and rely on voluntary commitments. Rather,it adheres to the fundamental principles of governing public goods. It relies on institutional design and evolutionary processes involving stakeholders to foster the development and enrichment of such public goods.

Exploration of “AI Safety as Global Public Goods” in China

Since the release of the Development Planning for a New Generation of AI in 2017,China has actively pursued advancements in AI development,safety,and governance. In response to the risks and challenges posed by new technologies and business models,a series of governance rules have been implemented,accompanied by the establishment of various governance institutions and the formulation of comprehensive governance plans. These efforts have contributed to China’s accumulation of valuable experience in global AI governance.

AI safety is a key focus and crucial component of China’s AI governance. Drawing on the scientific principles of AI technology innovation and industrial application,as well as considering the Chinese national context of balancing development and safety,China’s actions in the field of AI safety governance can be summarized into four aspects. First,promoting dialogue among stakeholders and fostering a broad consensus on AI safety governance has led to the formulation of fundamental principles such as fairness,justice,respect for privacy,safety and controllability,shared responsibility,and agile governance. Second,addressing specific issues in AI safety governance has resulted in the introduction of laws and regulations pertaining to information service algorithms,deep synthesis techniques,and generative AI safety governance. Additionally,various mechanisms have been proposed to enhance safety governance including algorithm filing procedures,algorithm classification systems,and algorithm impact assessments. Third,mobilizing frontline forces such as research institutions and enterprises has encouraged the development of technological solutions for AI safety governance. This approach aims to bolster the safety capabilities of entities involved in technological innovation and application. Furthermore,the exploration of constructing resources libraries,such as AI safety standards,risk management frameworks,risk evaluation benchmarks,and governance benchmark cases,is underway to further strengthen the overall safety infrastructure in this domain. Fourth,it is recommended that China actively engage in the global governance process of AI safety. This can be achieved by proposing initiatives that promote international cooperation in safety governance within the Global AI Governance Initiative. Additionally,participation in the Global AI Safety Summit series and proactive efforts to establish a dialogue platform for the World AI Conference should be pursued.

In the process of promoting AI safety governance,China has accumulated significant experience. A consensus has been formed among various stakeholders,including the government,enterprises,third parties,and other entities. They recognize AI safety as public goods and advocate for its production and accumulation through principles of public interest,scientific approach,responsibility,and agility. This collaborative effort aims to enhance both individual and overall levels of AI safety. Different entities play distinct roles in this process. The government takes on a framework role by establishing incentive-compatible institutional frameworks that provide space for the innovation and application of AI technology. It also offers safety knowledge,capabilities,and resources through public services. Enterprises and other technological innovation and application entities assume a leading role by actively exploring solutions for AI safety governance while participating reasonably in the production and management of AI safety public goods. Third-party entities play a catalytic role by enriching governance mechanisms,balancing interests and demands,and improving the production governance of AI safety public goods within a healthy ecosystem. China’s response to the safety governance challenges posed by large-scale models in recent years has gradually unearthed valuable experiences that have been summarized effectively. The exploratory practices carried out by relevant institutions in Shanghai can be considered benchmark cases for implementing national requirements.

1 New Challenges in safety Governance of Large-Scale AI Model

Since 2022,the emergence of large language models has sparked a new wave of technological innovation in AI,heralding the advent of general AI technology and garnering global attention in the field of AI governance. However,while embracing the transformative potential of large model technology,new challenges have arisen concerning AI safety risks. These large models are approaching or even surpassing human capabilities across various dimensions such as language proficiency,computation power,and reasoning abilities. Consequently,they bring to light multidimensional safety risks including dissemination of false content,deep fakes,manipulation of public opinion,impact on employment opportunities,and infringement upon individual rights. Nevertheless,unlike traditional technology safety risks,those stemming from novel technologies and business models present three distinct challenges that necessitate enhanced safety governance.

First,the uncertain and dynamic nature of AI safety risks are more pronounced. The regulatory objectives for content safety risks are ambiguous,and large-scale model safety risks exhibit evolutionary characteristics. This poses challenges for government departments as regulators,technological innovation and application entities as regulated parties,and the public as stakeholders to anticipate safety risks in advance. Consequently,the phenomenon of “shared ignorance” becomes more serious. Addressing this new challenge requires collaborative efforts from multilateral and multi-stakeholder entities to explore and accumulate safety knowledge and resources universally,thereby enhancing overall safety capabilities.

Second,the opposition and interplay between development and safety as fundamental governance objectives are increasingly apparent and manifested at various levels. For instance,there exist overlapping and conflicting demands between government entities acting as regulators and the enterprises being regulated. The allocation of responsibilities among different government departments also overlaps with certain boundaries. Moreover,enterprises often report one another under the pretext of safety risks,engaging in cutthroat competition. These emerging dynamics not only necessitate multi-stakeholder involvement and shared accountability but also pose challenges to traditional governance frameworks and practices.

Third,the governance mechanism for safety risks in large-scale models is becoming increasingly complex and disorderly. Existing mechanisms such as algorithm registration,red team testing,and algorithm impact assessment face a series of challenges including unclear institutional positioning,fragmented testing standards,and lack of credibility in evaluation benchmarks. These issues have resulted in the current governance mechanism not only failing to effectively address safety risks but also causing chaos such as regulatory arbitrage.

The aforementioned three new challenges in the governance of large-scale model safety are not limited to the present situation; they reflect general safety governance concerns arising from the iteration and innovation of AI technology. In recent years,Shanghai’s relevant institutions’ exploration and innovation in this field have provided valuable benchmark references while embodying the fundamental concept of “AI safety as global public goods”.

2 Shanghai's Practical Experience in Dealing with Large Model Safety Governance

In response to the challenges associated with governing the safety of large-scale models,relevant institutions in Shanghai have primarily focused on three governance innovations and have gained valuable experience throughout this process.

First,in response to the uncertainty and dynamic nature of safety risks,Shanghai regulatory authorities have transformed their traditional regulatory approach and adopted a proactive stance as learners,aggregators,and disseminators of safety governance knowledge. This shift aims to address the challenge posed by “shared ignorance”. Regulatory authorities have recognized that no single entity can possess all the knowledge required for the safe governance of large models. However,given their critical position,regulators are uniquely situated to engage with a wide range of frontline technological innovators while also wielding regulatory power and resources. Consequently,after quickly identifying security risks and corresponding countermeasures,they can disseminate localized knowledge into broader insights by conducting "one-on-one" consultations with specific enterprises and organizing public safety salons. This approach fosters the enhancement of safety governance knowledge and capabilities across the entire industry.

Second,in response to the issue of balancing development and safety goals,regulatory authorities in Shanghai have taken a proactive approach by establishing a collaborative mechanism among various departments. This mechanism allows for ample space for innovative application entities while also maintaining accountability through punitive measures during and after events. Upon recognizing that safety governance must not overlook the need for development and that safety issues often need to be addressed in the course of progress,regulatory authorities have adopted proactive exploration,dynamic adjustment,and consensus-driven approaches to establish baseline boundaries. These boundaries allow frontline innovators to engage in a wide range of innovation activities above the baseline. At the same time,based on the self-commitment of these innovators,regulators retain governance tools such as algorithm registration,warnings through interviews,and risk notifications,ensuring that they can still hold enterprises accountable in the event of serious consequences.

Third,in response to the fragmentation of safety governance mechanisms,regulatory authorities in Shanghai have effectively utilized the regulatory support capabilities of public scientific research and technology institutions. They have also actively engaged stakeholders throughout the process to ensure credibility in constructing safety evaluation databases and other repositories through a transparent approach. Simultaneously,the institutional positioning of this safety evaluation database is not merely as a prerequisite for market access but rather a platform for accumulating and dynamically evolving safety knowledge and capabilities.

3 Conceptual Demonstration of “AI Safety as Global Public Goods”

The Shanghai practice embodies the fundamental concept of “AI safety as global public goods.” First,regulatory agencies are dedicated to acquiring,assimilating,and disseminating knowledge on safety governance based on public and scientific principles. They ensure that all stakeholders can benefit from the growth process of safety governance knowledge without excluding any particular stakeholder or impacting others. Second,the coordination between development and safety objectives adheres to the principles of responsibility and agility. This approach allows for innovation while maintaining a credible deterrent through accountability measures. Third,the construction of an open safety resource library and transformation of the positioning of the safety evaluation system adheres to principles such as the public principle,scientific principle,and responsibility principle. The aim is to incentivize sustained growth and maintenance of AI safety as public goods.

Action initiative to promote “AI Safety as Global Public Goods”

The preceding report elucidates the concept,principles,values, and cases of “AI safety as global public goods.” Building upon this foundation,we propose actionable initiatives for the sustained reform of future multilateral and multi-party entities in the field of AI safety. Additionally,we suggest establishing the “West Bund Dialogue: AI Safety and Governance Dialogue Network” to facilitate the implementation of these initiatives.

First,it is essential to promote open dialogue and guide the development and dissemination of knowledge,capabilities,and resources related to AI safety based on public principles. We should challenge the conventional notion that views AI safety solely as a means of competition and instead establish AI safety research institutions that serve the public interest. Encouraging private sector involvement in the production of AI safety public goods is also crucial. Moreover,we need to shift the government’s governance approach to become a learner,aggregator,and disseminator of AI safety public goods while deeply integrating into the AI technology industry ecosystem. At an international level,it is imperative to establish a stable mechanism for dialogue and exchange in order to foster the creation of global AI safety public goods.

Second,it is crucial to enhance scientific consensus and establish AI safety governance. This consensus needs to be rooted in the laws of AI technology and industrial development,and based on scientific principles. This can be achieved by recognizing new challenges and requirements in safety governance,thereby creating opportunities for institutional innovation. To address the needs of AI safety governance effectively,it is recommended to establish a range of public resources such as safety risk assessment databases,collections of safety risk governance cases,and funds dedicated to safety research. At a global level,encouraging multilateral and multi-stakeholder entities to strengthen their research efforts on AI safety governance will facilitate the gradual formation of scientific consensus.

Third,by adhering to the principle of responsibility as a boundary,we can establish reliable commitments while allowing ample room for AI technology and industrial innovation. This approach entails preserving and exploring various types of responsibility mechanisms to create secure boundaries. Government departments should be mandated to explore innovative accountability mechanisms that encompass procedural requirements such as transparency and openness,as well as substantive requirements like penalties and remedies. This will foster an environment that encourages institutional imagination in the realm of AI safety governance. Furthermore,frontline entities involved in technological innovation and applications should be incentivized to actively explore measures for governing AI safety in response to accountability demands. At a global level,it is advisable to encourage AI safety governance institutions to engage in peer review and policy learning processes aimed at enriching the toolbox and policy library pertaining to AI safety governance.

Fourth,by adopting agile principles as amethodology, collaborative cooperation can be promoted to effectively respond to AI safety risks within specific boundaries. This process aims to establish interdependent governance relationship among multilateral and multi-stakeholder entities. It is crucial for these multilateral and multi-stakeholder entities to maintain transparency and adaptability in the face of evolving AI safety risks. They should possess the ability to identify such risks and explore appropriate governance strategies in an uncertain and dynamically changing environment. Encouraging the establishment of collaborative communication mechanisms for addressing AI safety risks is essential for facilitating rapid dissemination,empowerment,and sharing of safety knowledge,capabilities,and resources.

Institution：

Shanghai AI Lab

Center of Industrial Development and Environmental Governance, Tsinghua University

School of International and Public Affairs, Shanghai Jiao Tong University

Adviser：

Xue Lan, Professor, Tsinghua University

Main Authors:

Wang Yingchun, Researcher, Shanghai AI Lab

Jia Kai, Associate Professor, Shanghai Jiao Tong University

Zhao Jing, Associate Professor, Tsinghua University

Chen Ling, Professor, Tsinghua University

Qin Chuanshen, Associate Professor, Tsinghua University

Yuan Yuan, Dean of Ali-Research

Fu Hongyu, Director, AI Center of Ali-Research

Liang Xingzhou, Ph.D Candidate, Shanghai AI Lab

Translator:

Jia Kai, Associate Professor, Shanghai Jiao Tong University

Zhao Jing, Associate Professor, Tsinghua University

Liu Boyan, Ph.D Candidate, Tsinghua University

Kwan Yee Ng, Concordia AI

Brian Tse, Concordia AI