Blue Paper Report on Large Model Governance (2023): From Rules to Practice

*Note: These are Jeffrey Ding's informal and unofficial translations -- all credit for the original goes to the authors and the original text linked below. These are informal translations and all credit for the original work goes to the authors. Others are welcome to share **excerpts** from these translations as long as my original translation is cited. Commenters should be aware that the Google Doc is also publicly shareable by link. These translations are part of the ChinAI newsletter - weekly-updated library of translations from Chinese thinkers on AI-related issues: https://chinai.substack.com/*

Source:
Institute of Policy and Economics, China Academy of Information and Communications Technology – 中国信通院 (CAICT) — a think tank under China's Ministry of Industry and Information Technology.
Key Laboratory of Intelligent Algorithm Security, Institute of Computing Technology, Chinese Academy of Sciences
Date: November, 2023
Original Mandarin: http://www.caict.ac.cn/kxyj/qwfb/ztbg/202311/t20231124_466440.htm

5. Exploring the main implementation tools for large model governance in China

Effective governance of artificial intelligence is inseparable from feasible governance tools and technical means. To strengthen the effective management of large models, we can start from the aspects of ex-ante algorithm filing, in-process risk assessment, and post-facto traceability detection, etc. This allows for further exploration of methods to deal with the risks of large model products such as uncontrollability and misuse, and gradually improve the large model governance system.

(1) Prior filing

As an important regulatory component of the algorithm governance system, algorithm filing is one of the ways to implement algorithm transparency requirements, aiming to protect the rights

and interests of users and maintain product safety and information security. At present, all countries have explored the registration of algorithms for large-model products and early implementation practices.

1. Current status of large model filing system

China's "Interim Measures for the Management of Generative Artificial Intelligence Services" proposes algorithm filing requirements. Article 17 states that those who provide large-scale model services with public opinion attributes or social mobilization capabilities should perform algorithm filings as well as filing procedures for changes and cancellations — in accordance with the "Internet Information Service Algorithm Recommendation Management Regulations". Large model service providers can fill in information related to subjects, algorithms, and products in the Internet Information Service Algorithm Filing System and complete the reporting procedures.

Algorithm is the basic reporting unit for algorithm filing and the bearing object of the filing number. It is mainly divided into two parts: basic attribute information and detailed attribute information. The basic attributes include algorithm name, role, application field, and other information such as an "Algorithm Safety Self-Assessment Report" and "Content to be Disclosed" form. The detailed attributes of the algorithm include algorithm data, algorithm model, algorithm strategy, risk and prevention mechanisms, etc. The person filling in the report should use the algorithm as the anchor point to fill in the report and associate it with the products and functions that use the algorithm. Take "Wenxin [Ernie] Large Model Algorithm" as an example, and "Baidu." is used as the reporting subject. Information related to its subject and algorithm should be filled in, and related to "Wenxin Yiyan [ErnieBot]" that uses "Wenxin Large Model Algorithm" (APP, website)" products. After passing the review, the registration number will be issued to the filing entity by reference to the algorithm. The entity should indicate its filing number and provide a public information link in a conspicuous location such as the website or application that provides services to the outside world.

2. Implementation of China's filing system

As of November 2023, in two batches, the Cyberspace Administration of China has issued a total of 151 deep synthetic service algorithm filing numbers, including 100 generative artificial intelligence algorithm filing registry numbers. Among these 100, there are 64 registration numbers for service providers and 36 numbers for technical support units for services. From the perspective of filing entities, entities that have been issued registration numbers cover 11 provinces, and the four provinces and cities of Beijing, Guangdong, Zhejiang, and Shanghai account for 87% of the total number of generative artificial intelligence algorithm registrations.

From the perspective of registration algorithms, the text generation category accounts for 54% of the total number of generative artificial intelligence algorithm registrations, and the image generation, audio generation, and video generation categories account for 38% of the total number of generative artificial intelligence algorithm registrations. From the perspective of

products and services, the registered products involve the generation and synthesis of various text, pictures, voices, videos, virtual portraits, etc., and are widely used in e-commerce, finance, medicine, education and other fields. In the field of C-side (consumer-side) applications, taking "WPSAI Text Generation Algorithm-1" as an example, WPS has launched WPS AI, the first generative AI application on China's collaborative office track. Currently, WPS has newly upgraded smart documents, smart forms, and PPT demos, as well as other convenient functions. In the field of B-side applications, taking the "Damo Academy Interactive Multi-functional Synthesis Algorithm" as an example, the large model launched by DAMO Academy is used in open domain multi-modal content generation scenarios to serve the enterprise side of question and answer and consulting. Clients, through APIs, provide the functionality to generate multimodal information based on user input. Currently, large model algorithms such as "Wenxin Large Model Algorithm", "Skylark Large Model Algorithm" (Douyin's large model) and "iFlytek SparkDesk Large Model Algorithm" have been filed for registration one after another.

(2) Evaluation during the entire process

In large model governance, risk assessment is one of the key links to ensure compliance, and it is also an important means to ensure technical security and output content security. Mainstream countries, including China, have actively explored the implementation of safety risk assessments for such products.

1. The current situation of China's assessment system.

From the perspective of system design, China's large-model services need to conduct security assessments before they are launched and during operation. Article 17 of the "Interim Measures for the Management of Generative Artificial Intelligence Services" stipulates that large-scale model services with public opinion attributes or social mobilization capabilities should conduct a safety assessment in accordance with relevant national regulations. The requirement to carry out security assessments for information services with public opinion attributes or social mobilization capabilities began with the "Regulations on the Administration of Security Assessment of New Technologies and New Applications for News and Information Services" promulgated on December 1, 2017, also known as the "Double-New Assessments", and later further improved and clarified in the "Regulations for the Security Assessment of Internet Information Services Having Public Opinion Properties or Social Mobilization Capacity" published on November 30, 2018. Concerns about information content security risks are a core starting point for assessments.

At the level of technical standards, China has introduced national standards for algorithm security assessment, but it has not yet introduced independent large model assessment standards or regulations. Alongside 33 other units, the China Electronics Technology Standardization Institute and the Institute of Computing Technology of the Chinese Academy of

Sciences jointly developed and released the "Information security technology—Assessment specification for security of machine learning algorithms", including assessment preparation, assessment plan, assessment execution, assessment conclusion, assessment report, and other content guidelines. The China Artificial Intelligence Industry Development Alliance has drafted the "Large Model Training Model Technology and Application Evaluation Method" to explore specific regulatory requirements for controllability and trustworthiness. The industry is exploring and formulating evaluation standards for large models.

2. Implementation status and problems of the assessment system.

From the perspective of assessments in practice, the academic community and industry have currently released various evaluation tools and evaluation platforms. The current large model evaluation set focuses on evaluating the ability of large models, including Chinese language understanding ability, Chinese knowledge application and reasoning ability, etc. The data sources are mainly various examinations, and the data forms are mainly objective tasks such as multiple choice questions and true-or-false questions.

There have also been some attempts at large model safety assessments. For example, the Tsinghua CoAI team launched a Chinese large model safety assessment platform[1] to provide assessment services for ethical and safety issues of large language models; the China Academy of Information and Communications Technology has established a relevant public service platform to carry out a series of assessments of large-scale model safety based on specifications and evaluations of prevention capabilities; Beijing Academy of Artificial Intelligence established the FlagEval Libra large model evaluation system and open platform, which includes a three-dimensional evaluation framework of “abilities, tasks, and indicators”[2]; Ant Group released the integrated large model safety solution “Yitianjian”, which includes a large model safety detection platform “Yijian 2.0” and the large model risk defense platform “Tianjian.”

In practice, there are still some problems in the socialization process for the assessment of large model technology in China. From the perspective of assessment technology, on the one hand, due to the large amount of assessment content, there is a lack of efficient and comprehensive assessment tools, systems, and other technologies; on the other hand, large model evaluation has not yet formed a unified evaluation method and system of indicators, and the consistency and objectivity of the evaluation require further demonstration. Regarding the market for such assessments, the phenomenon of leading entities such as scientific research institutions and colleges and universities operating independently is quite prominent. There are problems such as too many evaluation standards, serious problems with “swiping scores”[刷分], and large differences in evaluation results. For example, a certain large model ranks tenth on the SuperCLUE list, and then is ranked very highly in the "2023 AI Large Model Technical Capability Assessment Report" by the well-known consulting company IDC.

[1] http://115.182.62.166:18000/
[2] https://flageval.baai.ac.cn/#/home

3. Assessment and monitoring tool solutions with different paths in Europe and the United States

Excerpted lengthy summary passage about U.S. assessments of AI safety and security issues, such as the draft Algorithm Accountability Act (which sets up the Federal Trade Commission as the key body), the National Telecommunications and Information Administration's efforts to standardize assessments, and a DEFCON event that allowed for public red-teaming of LLMS.

The EU's assessment of base models differs from high-risk AI assessment requirements. First, in the compromise of the Artificial Intelligence Act passed by the EU Parliament in June 2023, basic models (including large models) were listed separately and were not regarded as high-risk artificial intelligence. Specifically, for the assessment of high-risk AI, strict access is adopted, requiring the CE mark, that is, passing the mandatory CE (Conformity With European) marking procedure, requiring high-risk artificial intelligence systems to complete market access and certification. Specific assessment and certification will be conducted by a third-party agency designated by the regulatory agency that has independence, corresponding capabilities, no conflict of interest, and meets minimum network security requirements. The current evaluation of foundation models requires self-evaluation through appropriate methods or the engagement of independent experts to conduct model evaluations, document analysis, and conduct extensive testing during conceptualization, design, and development.

(3) Post-facto traceability detection

Large model content traceability is used to solve the problem of identifying where content is coming from. One aspect is determining whether the content comes from humans or large models. Another aspect is determining which big model the content comes from. Traceability detection technology can be used to prevent the misuse of generated content, the contamination of human-authored data, and fine-grained tracing of the source of generated content. Common technical means include detection based on implicit identification, detection based on content distribution, etc. The former requires the participation of large model service providers, while the latter does not depend on large model service providers.

1. Logo traceability: Implicit logos (watermarks) for risk traceability based on content

Implicit identification mainly refers to logos added by modifying text, pictures, audio, and video content that humans cannot directly perceive. Implicit identification can be extracted from the content through technical means and can be very effective for tracing content sources. The purpose is to support the detection and traceability of content generated by large models.

Currently, implicit identification has gradually been laid out and applied in the field of large models. Domestic releases of relevant content labeling industry standards propose that when AI technology is used to generate image, audio, and video content, digital watermarks should be

added to the content. When outputting audio, video and text in file form, digital watermarks should also be included in the file metadata.

Domestically, Alibaba uses digital watermark technology to protect audio, video, text and other content generated by large models, such as adding digital watermarks to image content generated by services such as its "Tongyi Wanxiang" text-to-image generator, Taobao AI Fitting Room, Taobao Life, etc. It has also added digital watermarks that are resistant to screenshots to business services such as Tongyi Qianwen, IdeaLAB, and DingTalk Documents to resist screenshots. Digital watermarks have also been integrated with large model-related services such as DAMO Academy Digital People and Taobao's AI recommendation function.[3]

Internationally, Meta and the French Institute for Research in Computer Science and Automation (INRIA) jointly developed Stable Signature, which can embed digital watermarks directly into images automatically generated by AI to prevent them from being used for illegal purposes. The digital watermark generated by Stable Signature is not affected by destructive operations such as cropping, compression, and color change, and can be traced back to the original source of the image. It can be applied to models such as diffusion and generative adversarial networks, such as the famous Vincentian graph software SD (Stable Diffusion) and MJ (Midjourney).[4]

Although implicit identification enables basic traceability and tracking capabilities for large model content, it still faces certain challenges. From the perspective of identifying objects, comprehensiveness cannot be guaranteed at the technical level. Specifically, existing large model identification objects are mainly pictures or videos, and it is still technically difficult to identify text content. From the perspective of cross-platform interoperability and mutual recognition, large model identification schemes have not reached a unified standard. For example, large model identifiers such as Non-Fungible Token (NFT) often use longer hash identifiers, while identifiers for user-generated content are usually randomly generated by the platform or named by users themselves. . From the perspective of identification traceability, it is also impossible to trace the content provider. Most of the existing large model identifiers embed the identifier as part of the data to generate content. The content provider can control the generation and use of these data, and the relevant responsible persons can only be traced back to the content user rather than the content provider.

2. Content traceability: risk traceability based on content distribution detection

The large model detection tool traces the source based on the generated content and does not rely on auxiliary information such as identification and generation logs. It is an important part of implementing the governance concept of "technology to govern technology."

All sectors of society are currently actively developing and exploring automated detection methods and tools. From the perspective of text detection, distinguishing text generated by humans and large language models becomes a crucial issue. Major language model

[3] https://www.163.com/dy/article/ICMA9DOS0514R9KQ.html
[4] See: https://www.sohu.com/a/727144326_121649381

manufacturers and research institutions have announced generated content recognition tools. For example, OpenAI launched an AI-generated content identifier called "Text Classifier", and Stanford University launched DetectGPT method to identify machine-generated text. The Institute of Computing Technology of the Chinese Academy of Sciences proposed a detection tool called LLMDet. Compared with existing detection methods, this tool can locate the basic model of generated text sources and demonstrates good detection performance while ensuring speed and security.[5] During the same period, research teams from Peking University and Huawei also proposed multi-scale learning solutions to improve the performance of text detectors in AI-generated corpus.[6] From the perspective of image and video detection, the need for detection and traceability of multimedia synthesis technology is becoming increasingly urgent. In November 2022, Intel launched the deep synthesis detection tool FakeCatcher, which detects AI face-changing videos by detecting blood flow. Officials say it can perform real-time detection and display results within a few milliseconds, and can detect fake videos produced by artificial intelligence algorithms and AI tampering. The video accuracy is 96%. Researchers from Harbin Institute of Technology (Shenzhen) and Nanyang Technological University proposed the task of detecting and locating multi-modal media manipulation and open sourced the multi-modal media manipulation data set. Compared with the existing single-modal deep forgery detection task, it is better at identifying whether inputted images and text are real or fake. It can also give further details on the manipulated content.[7]

Generated content detection still faces some general challenges and difficulties that need to be solved urgently. First, due to the complexity of the language itself, it is difficult to identify. For example, the OpenAI detector's success rate in detecting AI-written content is only 26%; second, the content generated by large models has many changes, high randomness, and large quantities, which makes detection tools difficult. Put forward higher requirements on timeliness and detection efficiency. Third, large model detection is now mostly carried out in the product dimension. Large model products are in full bloom, and data interoperability of multiple products has not been realized, which puts forward higher requirements for the universality of detection tools.

6. Ideas and suggestions for improving China's large model governance system

At present, AI governance has moved from conceptual discussion to the forefront of practical exploration. Facing the exponential growth of large model applications, large model governance should coordinate with multiple entities, take into account multi-dimensional goals, integrate multiple values, grasp governance priorities, innovate governance tools, strengthen global cooperation and dialogue, and promote the construction of inclusive and shared artificial intelligence governance. system.

[5] https://arxiv.org/abs/2305.15004.
[6] https://arxiv.org/abs/2305.18149
[7] https://arxiv.org/abs/2304.02556

(1) Establish a concept of agile AI governance that promotes innovation

Innovation is the first driving force for development. The concept of agile governance should be explored, a flexible and comprehensive institutional framework should be established, and a positive interaction between high-quality development of artificial intelligence and high-level security should be promoted.

The first is to balance innovative development and risk management. Achieving a balance of multiple goals through agile governance does not simply emphasize risk control, nor does it pursue efficiency one-sidedly. Encourage the innovative application of large model technology in various industries and fields, support relevant institutions to collaborate in technological innovation, data resource construction, transformation and application, risk prevention, etc., and encourage independent innovation in basic technologies.

The second is to continue to strengthen the cross-departmental coordination mechanism. Strengthening cross-department collaboration is an inevitable choice for current large-scale model supervision. We should support and improve cross-department and cross-regional policy coordination, law enforcement joint response and collaboration mechanism construction, and strive to solve problems such as overlapping department functions and non-sharing of regulatory information. This will promote the institutionalization and normalization of collaborative supervision.

The third is to establish and improve a diverse and agile interaction mechanism. Create a social co-governance model with government leadership, enterprise autonomy, industry self-discipline, and societal supervision. The government guides enterprises to find problems, improve designs, reduce risks, and coordinate to solve the corresponding difficulties of pilot enterprises. Enterprises should regularly submit periodic risk assessment reports, establish and improve internal monitoring and early warning mechanisms, report incidents in a timely manner after major risks occur, enhance the prevention awareness of the whole society, and encourage citizen supervision [公民监督].

(2) Focus on the application of artificial intelligence scenarios and refine the system plan

The first is to advance the legislative process such as the Artificial Intelligence Law and clarify institutional norms from the dimensions of industrial development, ethical guidance, and bottom line red lines. Add the "text and data mining" exception clause to the "fair use" situation in China's Copyright Law to respond positively to the issue of the use of artificial intelligence works. Promote and improve data sharing and circulation specifications, and establish and improve the implementation details of personal information protection in large model scenarios.

The second is to use a regulatory sandbox pilot to understand the characteristics of the risks of scenario applications. It is recommended to select fields with mature applications of large models such as media, education, and medical care to carry out AI governance pilot work. Enterprises are encouraged to actively trial governance plans and tools, identify risk issues in main scenarios and key links, and improve the entire process and all-factor institutional supply system for large model technology applications.

The third is to refine the rule scheme for typical risks in key scenarios. Explore differentiated governance measures based on large model deployment methods, application scenarios, etc. The government takes the lead and entrusts third-party organizations to establish authoritative implementation details or standards, establish detailed rules around large model technical capabilities, training data, data annotation and other links and fields, and clarify evaluation standards and processes.

The fourth is to establish and improve the hierarchical classification list of large models. Based on the sandbox experience, comprehensively investigate and evaluate the risk level of large models, refine the list of large model classifications, and clearly refine the operational standards for classification through the issuance of normative documents, industry standards, etc., and make dynamic adjustments according to the actual situation.

(3) Innovate artificial intelligence governance tools based on current governance practices

The governance of large models requires not only improving governance concepts and rules, but also optimizing governance methods and capabilities, and further updating and enriching the governance toolbox.

The first is to optimize the regulatory system tools to promote full-process supervision before, during and after the release of the model. Clarify the evaluation effectiveness from risk levels, new technology and new application categories, improve multi-dimensional evaluation indicators such as robustness, security, privacy, fairness, etc., and coordinate evaluation systems such as information content risks, personal information protection, security, and copyright protection. Publish specific evaluation guidelines such as data review database and data annotation specifications.

The second is to strengthen the allocation of resources such as large model supervision platforms and technical tools. Build a national-level large-scale model testing and verification platform to provide services such as model testing and verification, supply and demand docking, implement testing capabilities such as model confrontation security, backdoor security, and explainability, and promote the development and sharing of technology such as hardening tools. Build official large model training and testing data sets to reduce the cost of obtaining

high-quality data. Enhance the dynamic perception, scientific early warning, trace traceability, investigation and evidence collection capabilities of large model risks, and improve the level of professionalism, precision, and intelligence in governance.

The third is to introduce societal forces to improve the service level of large model evaluation. Strengthen the strength of third-party evaluation institutions in the field of artificial intelligence, clarify qualification qualification requirements such as personnel professional capabilities, technical tool reserves, and resource platform construction, and conduct regular annual qualification reviews to jointly build a service network with complementary advantages and coordinated development, and actively promote mutual recognition and interoperability of international testing.

(4) Encourage enterprises to actively manage and control risks to promote platform compliance

Corporate compliance is a self-governance method for enterprises to operate in accordance with laws and regulations and prevent and control compliance risks. They should implement the multi-entity governance concepts and build a new pattern of large-model platform governance with the help of societal forces.

First, establish and improve the internal compliance organizational structure and working mechanism. The platform should build an integrated compliance management organizational structure with clear responsibilities, clear levels, efficient collaboration, and tight control. It should accurately allocate compliance management functions based on the requirements of hierarchical management and comprehensive coverage. Establish a coordinated compliance management working mechanism to ensure that all compliance departments can effectively carry out their work.

The second is to optimize the platform’s internal governance system to deal with internal and external risks. Establish daily risk monitoring mechanisms, violation reporting mechanisms and compliance reporting mechanisms, etc., unblock user feedback channels, optimize human supervision, user complaint reporting and remediation procedures, etc. Strengthen the platform's ability to add logos and identify logos, establish unified identification standards, and the government will promote third-party platforms to develop free logo tools. Optimize the platform content management strategy, such as identifying illegal data on the input side, prompting users and rejecting this generation request, and further punishing user accounts that still frequently input illegal data after prompts.

The third is to establish a regulatory compliance evaluation system to implement platform compliance. Compliance evaluation should focus on whether formal elements such as the platform’s compliance organization system, compliance obligation system, and risk monitoring system are perfect, and strengthen the evaluation of the construction of compliance culture. Establish a compliance liability reduction and exemption mechanism, and provide preferential

incentive policies in national project declarations, government public service resource procurement, etc. for those that have actively explored and achieved obvious results in large-scale model governance.

(5) Promote the construction of a global cooperative AI governance system

AI governance is related to the destiny of all humanity and is a common issue faced by countries around the world. Active participation in and promotion of international cooperative governance will help create a new win-win situation in this field. The first is to promote an inclusive and open global dialogue on artificial intelligence. Establish a truly broadly representative global AI governance dialogue mechanism to build consensus around common risks. It is recommended to establish an intergovernmental consultation and assessment agency to conduct exchanges on major issues such as the potential impact of artificial intelligence on the economy and society, risk assessment, and governance frameworks.

The second is to help late-developing countries better acquire and utilize artificial intelligence technology, products and services. AI has become the basis for production in the digital age. Developing countries generally lack key elements such as digital infrastructure, innovation environment, and technical talents. The development of artificial intelligence industries and applications is restricted, and the digital divide is further widened. It is recommended to design reasonable financing, assistance and capacity-building mechanisms around AI to promote fair access and safe use of AI technology.

The third is to promote international cooperation in testing and evaluation of artificial intelligence. The International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC) have carried out standard research around key terms, etc., but it will be difficult for them to respond to the international community's urgent demand for artificial intelligence safety in the short term. It is recommended to actively promote cooperation in artificial intelligence research, widely bring together artificial intelligence experts from various countries, and jointly explore testing and evaluation methods on the basis of respecting the cultural diversity, political security and other demands of all parties, and assist late-developing countries to jointly reduce the risks of large-scale model technology.