人工智能的法律探究

金杜律师事务所
KING&WOOD
MALLESONS

人工智能的法律探究

序言

2022 年底，以 ChatGPT 为代表的生成式人工智能（“Generative AI”）掀起了新一轮的科技革命，不断催生新的商务业态。当下人工智能的开发已经上升到国家战略的高度，它的应用正逐渐渗透到社会生活的各种领域中，包括交通运输、金融、医疗、制造业、文娱产业等。人工智能将给人类社会带来全方位的变革。然而，由此衍生的伦理与法律问题也接踵而至。

与此同时，各国监管部门纷纷出台相关政策法规，以对人工智能进行规制，内容涵盖监管框架、透明度、可解释性、数据隐私、伦理原则、安全标准、知识产权等多个方面。无论是联合国于 2024 年 3 月通过的首个关于人工智能的全球决议，亦或是全世界 28 个国家及欧盟共同签署的《布莱切利宣言》、中国提出的《全球人工智能治理倡议》，以及欧盟的最新立法《欧盟人工智能法案》，其均在强调人工智能的治理应当重视安全风险，没有底线的运用人工智能会为人类社会带来毁灭性危机。

此前，中国生成式人工智能领域的首部专门立法——《生成式人工智能服务管理暂行办法》已于 2023 年 8 月 15 日起正式生效实施，对生成式人工智能服务提供者和使用者提出了一系列合规要求。除了立法监管外，多国法院已开始处理涉及人工智能的法律纠纷。例如，中国北京互联网法院和广州互联网法院此前已分别就首例涉及人工智能生成物（“AIGC”）的可版权性，以及生成式人工智能服务提供者的知识产权侵权案件作出判决。美国多家人工智能企业正面临数起集体诉讼纠纷，相关案件均已被美国法院受理。欧盟近期也诞生了涉 AIGC 可版权性问题的第一案。

金杜作为一家根植亚洲、服务世界的国际领先律师事务所，拥有广泛的资源和一支具备跨学科背景的专业律师团队，也是中国最早从事人工智能法律实务的律师事务所之一。金杜人工智能团队实力雄厚，在人工智能投融资法律服务、人工智能知识产权保护、人工智能数据合规、人工智能争议解决等领域具有丰富的经验。我们曾多次为人工智能领域的上下游企业及相关公司提供咨询服务、相关协议起草修改、监管执法应对、诉讼和交易服

务，并且能够根据客户的具体需求，设计和实施定制化的“一站式”人工智能问题解决方案，为客户的商业战略服务，帮助客户实现有利的商业结果。同时，金杜还与全球各地的合作伙伴建立了紧密的合作关系，能够为客户提供全球范围内人工智能方面的法律支持。在日新月异的技术发展图景中，金杜人工智能团队始终保持对前沿领域的观察研究，为人工智能从业者提供了最新的实用建议，多次受邀撰写涉及生成式人工智能、大模型及机器学习等领域的专业文章，获得了业界的高度肯定与赞扬。

本次发布的《人工智能的法律探究》，由金杜人工智能团队合力撰写，汇集了团队过去公开发表的文章，凝结了团队在人工智能领域的专业知识和丰富经验，旨在为读者提供人工智能领域法律问题的有益参考与实务参照。本期刊分为政策解读、法律问题探究、案例分析、AI 合规管理及境外的 AI 发展五大部分，涵盖了中美欧英多国人工智能立法解读、生成式人工智能训练数据合法性、AIGC 的可版权性、AIGC 知识产权侵权等热点问题，并且涉及教育、汽车、医疗、娱乐及金融等不同行业领域。除用中英双语解读国内外人工智能监管法规、剖析涉及人工智能的司法案例外，本期刊还分主题、分场景聚焦生成式人工智能服务提供者和使用者可能面临的科技伦理审查、知识产权、数据隐私及国家安全等风险和问题，并提出了针对性的防范措施。我们将以本期刊为起点，上下求索，精益求精，携手社会各界共同把握人工智能时代的新机遇。

2024 年 4 月

宋海燕

金杜国际合伙人

目录

AI 案例评析

AI 合规管理

境外的 AI 发展

AI 政策解读

The Latest Draft Measures on the Management of Generative AI

宋海燕 陈佩龄

In response to the rapid development of Generative AI services in the past few months, especially the launch of ChatGPT since last December, on April 11, 2023, Chinese Cyberspace Administration issued a draft policy Measures on the Management of Generative Artificial Intelligence Services ("**Draft AI Policy**"), soliciting feedback from the general public with respect to the regulation and management of Generative AI services.

The Draft AI policy addressed a few important topics surrounding AI and AIGC, including (1) the definition of Generative AI and to what extend this Draft AI Policy should apply; (2) whether text and data mining ("**TDM**") might constitute copyright infringement in training AI; (3) misinformation risks associated with AIGC; (4) data privacy issue; and last but not least (5) security evaluation of AI service provider and its filing procedure etc.

I. Generative AI and the Scope of this Draft AI Policy

To start with, "Generative AI" is defined in the Draft AI Policy as a technology that generates text, images, sound, video, codes and other contents based on algorithms, models, and rules.

Article 2 of the Draft AI Policy stipulates that, regardless whether a Generative AI service or product is developed or used in China, as long as it is targeting at consumers located within China, that AI service or product provider will be subject to the Draft AI Policy.

II. TDM Issue Training AI

Unlike the UK, EU, Japan or Singapore that allows text and data mining exception in training AI, the Draft AI Policy explicitly states in Article 7 that, the AI product/service provider shall be responsible for the "legality of the source of pre-training and optimization training data", thus requiring the AI product/service provider to obtain clearance from IP rights holders prior to using such copyrighted works to train AI. Given the limited few statutory categories of "fair use" exceptions in Chinese Copyright Law and the mixed results of copyright fair use cases in the recent Chinese case law, the use of copyrighted works to train AI (i.e., TDM) still faces big challenges of copyright infringement suits.

III. Misinformation

Chinese legislators are apparently well aware of the risks associated with AIGC, in terms of

potential discrimination issue, misinformation and violation of right of publicity etc. Specifically, the Draft AI Policy (Articles 4, 7 and 12) requires AI product/service providers to apply "technical means" to "avoid the generation of illegal content, false information, and discriminatory content as much as possible."

IV. Data Privacy

In response to the data privacy concerns associated with generative AI, the Draft AI Policy addresses the issue. Specifically, it emphasizes the obligation of AI product/service providers to protect users' personal information and privacy. Article 11 of the Draft AI Policy sets a limit on information retention for generative AI products/services, stipulating that "providers shall not illegally retain input information that can infer the user's identity, shall not draw portraits based on the user's input information and usage, and shall not provide the user's input information to others." Article 13 of the Draft AI Policy also requires AI product/service providers to establish a mechanism for receiving and handling user complaints, stating that "measures shall be taken to stop such generation and prevent enduring hazard" when an infringement is discovered.

V. Security Evaluation and Filing Procedure

Article 6 of the Draft AI Policy requires that any release of AIGC products must undergo a security evaluation and be filed with the competent authorities. Article 6 stipulates that "Before providing services to the public with generative artificial intelligence products, it shall report to the National Cyberspace Administration for security assessment in accordance with the Regulations on the Security Assessment of Internet Information Services with the Attributes of Public Opinion or the Ability of Social Mobilization, and go through the procedures of algorithm filing, alteration and cancellation in accordance with the Regulations on the Management of Algorithmic Recommendations in Internet Information Services."

Conclusion

It should also be noted that the Draft AI Policy is silent on the copyrightability of AIGC, i.e., whether AI generated content should be entitled to copyright protection, which is traditionally reserved for human authors. It is unclear whether Chinese Copyright Office will issue a separate policy statement on the copyrightability of AIGC, if it is not to be covered under this Draft AI Policy. In view of the recent two cases[1] on AIGC, we feel that Chinese attribute on the copyrightability of AIGC might be similar to the copyright policy statement[2] issued by the U.S. Copyright Office on March 16, 2023. This Draft AI Policy is one of the recent legislative efforts that Chinese government made to regulate the development of AI technology and services. In the Regulations on the Management of Algorithmic Recommendations in Internet Information Services and the

---

[1] See: 北京互联网法院 (2018) 京 0491 民初 239 号；北京知识产权法院 (2019) 京 73 民终 2030 号 ; 广东省深圳市南山区人民法院（2019）粤 0305 民初 14010 号。

[2] Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence.https://www.federalregister.gov/documents/2023/03/16/2023-05321/copyright-registration-guidance-works-containing-material-generated-by-artificial-intelligence. last visit：4/17/23.

Regulations on the Administration of Deep Synthesis Internet Information Services issued in 2022, Chinese government also addressed the issues of deep fake and data privacy associated with AI development.

The Measures for the Administration of Generative Artificial Intelligence Service (Draft for comments)

Article 1 For the purpose of promoting the sound development and standardized application of generative artificial intelligence, the Measures are formulated in accordance with the Cybersecurity Law of the People's Republic of China, the Data Security Law of the People's Republic of China, the Personal Information Protection Law of the People's Republic of China and other laws and administrative regulations.

Article 2 The Measures shall apply to those who develop and utilize generative artificial intelligence products to provide services to the public within the territory of the People's Republic of China.

Generative artificial intelligence as mentioned in the Measures refers to the technology of generating text, image, sound, video, code and other contents based on algorithms, models and rules.

Article 3 The State supports independent innovation, popularization and application of artificial intelligence algorithms, frameworks and other basic technologies as well as international cooperation, and encourages priority to be given to the use of secure and trustworthy software, tools, calculations and data resources.

Article 4 The provision of generative artificial intelligence products or services shall comply with the requirements of laws and regulations, observe social ethics, public order and good morals, and meet the following requirements:

(1) The content generated by generative artificial intelligence should demonstrate the socialist core values. It must not contain content such as subverting state power, overthrowing the socialist system, inciting secession of the country, undermining national unity, promoting terrorism and extremism, promoting ethnic hatred and discrimination, violence, obscene and pornography, false information, and content that may disturb economic and social order.

(2) In the process of algorithm design, training data selection, model generation and optimization, and service provision, measures should be taken to prevent discrimination based on race, ethnicity, belief, nationality, region, sex, age, occupation, etc.

(3) Respect intellectual property rights and business ethics, and shall not use algorithms, data, platforms and other advantages to carry out unfair competition.

(4) The content generated by generative artificial intelligence shall be authentic and accurate, measures should be taken to prevent the generation of false information.

(5) Respect the legitimate interests of others, prevent injury to the physical and mental health of others, prevent damage to the right of likeness, right of reputation and personal privacy, and the infringement of intellectual property rights. It is prohibited to illegally obtain, disclose or use personal information, privacy and business secrets.

Article 5 Organizations and individuals (hereinafter referred to as "**providers**") that use generative artificial intelligence products to provide services such as chat, text, image and sound generation, including supporting others to generate text, image and sound by providing programmable interfaces, shall assume the responsibilities of producers of the content generated by such products; Where personal information is involved, it shall bear the statutory responsibility of the person handling personal information and fulfill the obligation of personal information protection.

Article 6 Before providing services to the public with generative artificial intelligence products, it shall report to the National Cyberspace Administration for security assessment in accordance with the Regulations on the Security Assessment of Internet Information Services with the Attributes of Public Opinion or the Ability of Social Mobilization, and go through the filing, alteration and cancellation procedures for algorithm in accordance with the Regulations on the Management of Algorithmic Recommendations in Internet Information Services.

Article 7 Providers shall be responsible for the legality of pre-training and optimized training data sources of generative AI products.

The pre-training and optimized training data used for generative AI products shall meet the following requirements:

(1) It should comply with the requirements of the Cybersecurity Law of the People's Republic of China and other laws and regulations;

(2) The contents should not infringe intellectual property rights;

(3) If the data contains personal information, consent of the subject of personal information shall be obtained or other circumstances prescribed by laws and administrative regulations shall be complied with;

(4) The authenticity, accuracy, objectivity and diversity of data should be ensured;

(5) Other regulatory requirements of the Cyberspace Administration of China on generative artificial intelligence services.

Article 8 When manual labeling is used in the development of generative artificial intelligence products, providers shall make clear, specific and operable labeling rules in accordance with the requirements of the Measures, conduct necessary training for labeling personnel, and verify the correctness of labeling contents by sampling.

Article 9 The provision of generative artificial intelligence services shall require users to provide

real identity information in accordance with the provisions of the Cybersecurity Law of the People's Republic of China.

Article 10 Providers shall clarify and disclose the applicable groups, circumstances, and uses of its services, and take appropriate measures to prevent users from excessively relying on or indulging in the generated content.

Article 11 In the process of providing services, providers shall undertake the obligation to protect the input information and usage records of users. Providers shall not illegally retain input information that can infer the identity of the user, shall not draw portraits based on the user's input information and usage, and shall not provide the user's input information to others, except as otherwise provided for by any law or regulation.

Article 12 Providers shall not generate discriminatory content based on the user's race, nationality, sex, etc.

Article 13 Providers shall establish a mechanism for receiving and handling users' complaints and promptly handle requests for correction, deletion or shielding of their personal information; When generated text, image, sound or video is found to infringe upon the right of likeness, right of reputation, personal privacy and trade secrets of others, or fail to meet the requirements of the Measures, measures shall be taken to stop such generation and prevent enduring hazard.

Article 14 Providers shall, within the life cycle, provide safe, robust and continuous services to ensure the normal use of users.

Article 15 For the generated content found in operation or reported by users that does not meet the requirements of the Measures, in addition to taking measures such as content filtering, it shall be prevented from being regenerated by means of model optimization training within 3 months.

Article 16 Providers shall mark the generated images, videos and other contents in accordance with the Provisions on the Administration of Deep Synthesis of Internet-based Information Services.

Article 17 Providers shall, in accordance with the requirements of the National Cyberspace Administration and relevant authorities, provide necessary information that can affect the trust and choice of users, including the description of the source, scale, type and quality of pre-training and optimized training data, manual labeling rules, scale and type of manual labeling data, basic algorithm and technical system, etc.

Article 18 Providers shall guide users to have a scientific understanding and rational use of the content generated by generative artificial intelligence, not to damage the image, reputation or other legitimate rights and interests of others by using the generated content, and not to engage in commercial speculation or improper marketing.

When users find that the generated content does not meet the requirements of the Measures, they

shall have the right to report to the National Cyberspace Administration or relevant authorities.

Article 19 Providers shall suspend or terminate the service when it finds the user's violation of laws and regulations, business ethics or social morality in the process of using the generative artificial intelligence products, including engaging in network speculation, malicious posting and commenting, creating spam, programming malicious software, and implementing improper commercial marketing, etc.

Article 20 Providers who violate the provisions of the Measures shall be punished by the National Cyberspace Administration and relevant authorities in accordance with the Cybersecurity Law of the People's Republic of China, the Data Security Law of the People's Republic of China, the Personal Information Protection Law of the People's Republic of China and other laws and administrative regulations.

Where laws or administrative regulations do not provide for it, the National Cyberspace Administration and relevant authorities shall, in line with their functions and duties, issue warnings, circulate notice of criticism and order correction within limited time; Where corrections are refused or the circumstances are serious, the offender shall be ordered to suspend or terminate the use of generative artificial intelligence to provide services and be fined not less than 10,000 yuan but not more than 100,000 yuan. If the case constitutes an act violating the administration of public security, penalties for the administration of public security shall be imposed in accordance with law; If the case constitutes a crime, criminal responsibility shall be pursued in accordance with law.

Article 21 The Measures shall be implemented as of [day][month] 2023.

Thanks to CHAN PuiLing and Wang Mo for their contributions to this article. Attached is the full English translation of the Draft AI Policy prepared by CHAN PuiLing at KWM.

China's First Regulation on the Management of Generative AI

宋海燕 赵怡冰

At present, the new round of technology revolution triggered by generative artificial intelligence ("**AI**") is taking the world by storm. On 14 June 2023, the European Parliament passed the text of the Artificial Intelligence Act ("**EU AI Act**") by an overwhelming majority, which places certain obligations upon generative AI systems (e.g. ChatGPT).1 China is also highly concerned with the regulation of generative AI services. On 13 July 2023, the Cyberspace Administration of China ("**CAC**"), in conjunction with the national Development and Reform Commission and five other state agencies, jointly issued China' s first administrative regulation on the Management of Generative AI services - Interim Measures for the Administration of Generative Artificial Intelligence Service ("**Chinese AI Measures**"). The Chinese AI Measures had come into effect on 15 August 2023.

The finalized Chinese AI Measures have been substantially revised, after a 3-month public consultation period, compared with the earlier draft the Measures for the Administration of Generative Artificial Intelligence Service ("**Draft AI Policy**") issued in late April 2023 (you may refer to our previous analysis of the Draft AI Policy2). Compared with the earlier draft, the finalized Chinese AI Measures give more consideration to the current technological limitations of generative AI. On the other hand, it adopts the AI governance approach as "classified and graded supervision", thus shaping China's AI regulation as "top-level general legislation & industry regulation".

The finalized Chinese AI Measures address a few important topics surrounding generative AI services, including: (1) the scope of this regulation—to whom these AI Measures apply; (2) AI governance—which is based on classified and graded supervision; (3) responsibilities to AI service providers; and (4) protection of personal information.

I. Scope of the Chinese AI Measures

To start with, "Generative AI Technologies", in the Chinese AI Measures, refer to models and related technologies that have the ability to generate content such as text(s), picture(s), audio, and video(s).

Article 2 of the Chinese AI Measures limits the scope of this regulation to those who "provide

---

[1] See the EU Artificial Act (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021)0206 – C9-0146/2021–2021/0106(COD)): https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html, last visit: 19/07/23.

[2] https://mp.weixin.qq.com/s?__biz=MzA4NDMzNjMyNQ==&mid=2653305758&idx=2&sn=9bf2c85eb776ed23ace51e8e788c5c1c&chksm=843a5034b34dd9226a697496a72fc285ed353ab12ec1f1dab140577e727672568541dde74ffd&scene=21#wechat_redirect.

generative AI services to the public within the territory of the People's Republic of China ('**PRC**')", and excludes those (1) who develop and apply generative AI technologies and services for their own "internal use" or "research purposes"—not offered to the public within the PRC, and (2) who provide generative AI services outside the territory of the PRC. Based on the current legislative language, it seems to suggest that if Chinese providers of generative AI services incorporate overseas AI technologies (e.g. ChatGPT) into their own products and then provide such services to the public within the PRC, they are also likely to subject to the Chinese AI Measures.

II. AI Governance: The Idea of "Classified and Graded Supervision"

The Chinese AI Measures seem to follow the same governing principle as with the governance of internet platforms and data compliance in China, where Article 3 and 16 of the Chinese AI Measures propose the governance of AI based on "classified and graded supervision". However, unlike the EU AI Act, which explicitly sets out a "four-level" risk-based regulatory approach, the Chinese AI Measures have not provided specific criteria for the classification and grading of AI generative services yet.

III. Providers' Responsibilities

In terms of managing content, providers of generative AI services need to comply with the following principles:

During the data training process, providers shall only use the data and foundation models that are "legally acquired" and shall not infringe the IP rights of others (Article 7). Also, providers shall improve the "authenticity, accuracy, objectivity and diversity" of training data;

In terms of AI generated content, providers need to "manage the AI generated content" (Article 14) and take down illegal content or suspend and terminate users' services, when necessary; and

Providers are also asked to add "tags" on the AI generated content so as to distinguish such content from human-authored works.

Providers also need to establish a complaint and reporting mechanism (Article 15) to deal with the complaints from the public.

Last but not the least, when providers of generative AI services have the ability to "post/share public opinion" or "mobilize the society", they are required to undertake "security assessment" and "algorithm filing" (Article 17).

IV. Protection of Personal Information

With regard to the protection of personal information, the Chinese AI Measures prohibit the collection of "unnecessary" personal information and the "illegal" provision of users' input information to others (Article 11).

As far as the obligation of algorithm transparency is concerned, the Chinese AI Measures only

require providers to take "effective measures" to enhance the transparency of generative AI services (Article 4), thus leaving room for providers to fulfill this obligation in more flexible ways. However, compared with the EU AI Act, the Chinese AI Measures again fail to provide specific criteria as to how to judge the transparency of algorithms.

Conclusion

The Chinese AI Measures are the first attempt of Chinese government to regulate the generative AI services. Yet a few important issues remain to be clarified, such as the copyrightability of AIGC, the specific criteria for "classified and graded" AI supervision, the definition of the providers' ability to "support the public opinion or mobilize the public", and the extent of algorithms disclosure under the algorithm transparency requirements, just to name a few. We expect to see future regulatory guidance as the generative AI landscape continues to evolve.

Thanks to Sun Heyao, Shang Wendi, and Wu Zongying for their contributions to this article.

“卧看星河尽意明”——全球首部生成式人工智能法规解读

宁宣凤 吴涵 张浣然 刘畅 吴仁浩

引言

自 Open AI 研发的聊天机器人程序 ChatGPT（Chat Generative Transformer）于 2022 年 11 月 30 日发布以来，生成式人工智能（Generative AI）技术迅速商业化，引发社会各界对于其潜在商业及社会价值的讨论。同时各国监管部门对该类前沿技术经历从禁止到有限开放的过程[1]，虽然急于监管，但怠为提供具体的指引。由于对生成式人工智能底层技术缺乏深入理解，且对于其应用前景预测有限，各国尚未形成体系化的监管思路及规则。然而时不我待，在前沿技术巨大社会价值与部分不可测风险之间，亟需平衡发展与安全的监管规则作为市场主体的指路明灯，协助技术服务提供方和使用方妥善控制风险，促进发展。在该大背景下，国家互联网信息办公室（以下简称“网信办”）及时地响应市场需求，于 2023 年 4 月 11 日发布《生成式人工智能服务管理办法（征求意见稿）》（“《征求意见稿》”）。

最终，在广泛征求社会各界意见后，2023 年 7 月 13 日，网信办、国家发展和改革委员会、教育部、科学技术部、工业和信息化部、公安部、国家广播电视总局七部委联合发布《生成式人工智能服务管理暂行办法》（“《暂行办法》”），并于 2023 年 8 月 15 日正式生效，成为全球首部针对生成式人工智能的法规。值得注意的是，《暂行办法》确定了发展作为主基调，直接删除了《征求意见稿》中“不得进行用户画像”“用户实名制”等争议性规定，并修改了“利用生成式人工智能生成的内容应当真实准确，采取措施防止生成虚假信息”“保证数据的真实性、准确性”“对于运行中发现、用户举报的不符合本办法要求的生成内容，应在 3 个月内通过模型优化训练等方式防止再次生成”等争议性表述[2]。

在数字型社会迈向智能化社会的当前，新型技术将不断涌现。不可否认的是，技术壁垒造成的认知差距短时间内将不断拉大，因此如何应对新技术新应用对于社会管理的影响是全球通用的难题。国家法律法规不仅体现本国的监管智慧，反映国家社会管理的水平，

---

[1] 当地时间 2023 年 3 月 31 日，意大利个人数据保护局宣布即日起暂时禁止使用 ChatGPT。2023 年 4 月 13 日，西班牙国家数据保护局因 ChatGPT 可能违反《通用数据保护条例》而对 OpenAI 启动初步调查程序；法国国家信息自由委员会也决定对收到的有关 ChatGPT 的五起投诉展开调查。此外，欧盟数据保护委员会（EDPB）也于 2023 年 4 月 13 日宣布成立针对 ChatGPT 的专门小组。值得注意的是，随着 OpenAI 在隐私数据方面作出让步，4 月 29 日，意大利个人数据保护局撤销了对 ChatGPT 的暂时禁令。

[2] “全球首部生成式 AI 法规即将生效，多个争议条款被删改”，参见 https://mp.weixin.qq.com/s/yW9putNcX0XIEHcH8tQ3Pw。

也将被置于放大镜之下被各国审视，成为国家竞争力的重要组成部分，出现类似的布鲁塞尔效应（Brussels effect）。因此在为《暂行办法》迅速出台欢呼雀跃的同时，我们需要深刻理解其历史沿革、蕴含的监管逻辑，才能合理预测后续人工智能监管大局中的思路。因势利导，不急不躁，让中国的人工智能技术和监管规则都能屹立在国际潮头。

一、中国算法与人工智能治理的历史沿革

（一）法律规范：“法治之网”逐步完善

在人工智能治理的立法层面，我国以纲领性文件为方向，近年来在法律及标准层面以算法治理为重点构建人工智能监管的“桥头堡”。

在纲领文件层面，我国先后颁布《关于加强互联网信息服务算法综合治理的指导意见》（“《算法治理意见》”），明确了我国算法治理目标是建立治理机制健全、监管体系完善、算法生态规范的算法安全综合治理格局；《关于加强科技伦理治理的意见》提出伦理先行的治理要求；以及《新一代人工智能发展规划》《新一代人工智能伦理规范》《关于加快场景创新以人工智能高水平应用促进经济高质量发展的指导意见》等促进人工智能应用发展的文件。

在法律层面，《数据安全法》明确算法治理的原则，要求技术促进经济社会发展，增进人民福祉，符合社会公德和伦理；《个人信息保护法》从个人信息保护角度规制自动化决策；《互联网信息服务算法推荐管理规定》（“《算法推荐规定》”）重点治理以“大数据杀熟”为代表的算法歧视性决策；《互联网信息服务深度合成管理规定》（“《深度合成规定》”）秉持“辨伪求真”的愿景，以深度合成技术作为规制客体；《暂行办法》则整体聚焦生成式人工智能的规范应用与发展。在此基础上，还有诸多聚焦于电子商务、金融科技等特定行业领域的算法治理相关规定散见于《电子商务法》等法律法规中。

除立法之外，我国重视标准体系建设，利用产业标准进一步提高治理效能，例如国家标准化管理委员会等五部门印发《国家新一代人工智能标准体系建设指南》以加强人工智能领域标准化顶层设计，推动人工智能产业技术研发和标准制定；《人工智能伦理安全风险防范指引》针对人工智能可能产生的伦理安全风险问题，给出安全开展人工智能研究开发、设计制造、部署应用等相关活动的规范指引；此外，信安标委发布的 2023 年度第一批网络安全国家标准需求中，还包括拟作为《征求意见稿》配套标准的《生成式人工智能预训练和优化训练数据安全规范》《信息安全技术生成式人工智能人工标注安全规范》等标准。

（二）治理思路：风险治理

我国采用风险治理的理念早在《网络安全法》《数据安全法》规定的分类分级保护制

度中已有体现，《网络安全法》明确“国家实行网络安全等级保护制度”；《数据安全法》要求根据数据引发的外部风险，即对国家、社会的价值以及出现安全事件后造成的危害后果将数据划分为核心数据、重要数据和一般数据，并适配不同程度的保护制度。

《算法治理意见》进一步明确“坚持风险防控，推进算法分级分类安全管理”为治理的基本原则之一；《算法推荐规定》明确规定要根据算法推荐服务的舆论属性或社会动员能力、内容类别、用户规模、算法推荐技术处理敏感程度、对用户行为的干预程度等对其进行分类分级管理；《深度合成规定》和《算法推荐规定》均要求具有舆论属性或社会动员能力的服务提供者履行备案义务，并进行专门安全评估；《暂行办法》依然延续风险治理的思路，并在第十六条明确提出整体性的分类分级监管理念。

（三）治理方式：行业规制方式成效显著，顶层通用立法即将到来

《算法推荐规定》《深度合成规定》以及《暂行办法》均体现针对性立法、敏捷治理的思路——针对特定类型算法服务或应用制定规范。采取行业规制的方式使监管部门能迅速回应算法与人工智能在特定领域、行业引发的问题，实现敏捷治理、精准治理。

然而，尽管上述法规相互衔接，能够体现监管思路的延续性，但通用型人工智能具有迅速普及的特点，其在垂类领域的应用也将快速多样化，“点对点”的监管可能会“目不暇接”“疲于应对”。为体现科技立法的前瞻性，在人工智能不断发展的时代，我们需要横向综合性的人工智能立法在新技术、新形式出现时提供一般性监管要求。2023 年 6 月 6 日发布的《国务院 2023 年度立法工作计划》已明确将《人工智能法》纳入计划，自此我国算法与人工智能的治理体系兼具了精准设计、敏捷迅速与统一通用、涵盖广泛的优势，将更好地推动技术发展与风险管控双重目标实现。

二、责任主体义务梳理：压实服务提供者义务

《暂行办法》对生成式人工智能在具体场景的应用提出了基本、通用的合规要求，例如生成式人工智能提供者应当承担网络信息内容生产者责任。另一方面，《暂行办法》也对相关组织在新闻出版、影视制作等特殊领域利用生成式人工智能服务从事特定活动预留了一定空间，多领域主管部门依据职责，对本领域内的生成式人工智能服务应用开展行业垂直监管。

责任主体义务的设计通常是每部法律法规的重点，尽管《深度合成规定》中精确划分深度合成服务技术支持者、服务提供者与使用者的责任与义务，但《暂行办法》明确以服务提供者为治理抓手，围绕服务提供者进行义务设计。

《暂行办法》第二十二条的规定将“包括通过提供可编程接口等方式支持他人自行生

成文本、图像、声音等”修改为“生成式人工智能服务提供者，是指利用生成式人工智能技术提供生成式人工智能服务（包括通过提供可编程接口等方式提供生成式人工智能服务）的组织、个人”，即旨在压实生成式人工智能服务提供者义务，有助于避免相关主体怠于履行合规义务或相互推诿，使规范落到实处。

在《暂行办法》下，生成式人工智能服务提供者主要需履行算法训练相关义务、内容管理相关义务、使用者相关义务以及监管机制相关义务，具体如下：

义务类型	义务内容	义务综述
算法训练相关义务	• 算法训练数据处理合法性（第七条） • 训练数据质量要求（第七条） • 数据标注及培训要求（第八条） • 算法纠偏与报告义务（第十四条第一款）	在算法训练上，《暂行办法》设立事前、事中、事后全链条监管机制，包括： • 要求依法开展预训练、优化训练等训练； • 采取有效措施提高训练数据质量，增强训练数据的真实性、准确性、客观性、多样性； • 制定合乎要求的数据标注规则，并配备完整的监测评估和培训机制，确保标注规则有效落实； • 发现违法内容后，采取模型优化训练等算法纠偏措施，并向主管部门报告。
内容管理相关义务	• 内容生产者责任（第九条第一款） • 个人信息处理者责任（第九条第一款） • 内容标识（第十二条）	整体而言，涉及内容生成的，《暂行办法》要求提供者承担内容生产者责任，并根据《深度合成规定》进行内容标识；涉及个人信息处理的，提供者应切实履行个人信息保护义务。
使用者相关义务	• 服务协议签订（第九条第二款） • 合理使用与防沉迷机制（第十条） • 使用者输入数据与个人信息保护（第十一条） • 服务稳定性义务（第十三条） • 违法内容与使用者违法行为管理（第十四条） • 使用者投诉举报处理机制（第十五条）	考虑到提供者与使用者之间的权利义务关系，以及生成式工智能对使用者可能产生的影响，《暂行办法》要求提供者制定服务协议。 此外，要求提供者采取多种措施，保障使用者个人信息保护等各项权益，并引导使用者（尤其是未成年人用户）树立对生成式人工智能的正确观念，确保使用者合理、合法使用生成式人工智能服务，防止使用者利用生成式人工智能服务从事违法行为，生成违法内容。
监管机制相关义务	• 安全评估与算法备案义务（第十七条） • 使用者投诉举报机制（第十八条） • 监管配合及算法披露（第十九条）	在监管机制上，《暂行办法》延续《算法推荐规定》《深度合成规定》的监管手段，要求进行安全评估和算法备案，同时和其他互联网信息服务监管相同，要求建立有效投诉举报机制以应对公众监督。另外，包括算法披露义务在内的监管配合的制度有助于增强监管有效性。

值得注意的是，尽管当前生成式人工智能的服务提供者多为技术开发方，仍可以满足上述合规及技术要求。但随着市场的发展，将会有大量非技术持有者借助技术支持方对外提供生成式人工智能服务，如果缺乏技术开发和管理手段，服务提供方应当就上述义务与技术支持方达成明确且详尽的协议，以确保自身合规义务的履行。

三、《暂行办法》的主要变化梳理

（一）明确适用范围与域外效力

与《征求意见稿》相比，《暂行办法》以反向列举的方式排除了对“应用生成式人工智能技术，未向境内公众提供生成式人工智能服务”情形的适用。鉴于《征求意见稿》由网信办发布，结合网信办的职权范围（互联网信息内容管理等），其适用范围应当与《算法推荐规定》《深度合成规定》类似，即集中在互联网信息服务的场景中。

尽管《暂行办法》由七部委联合发布，理论上打破了其适用范围的禁锢，但考虑到《暂行办法》围绕服务提供者进行算法备案、安全评估以及网络信息内容治理等义务设计安排，可以合理认为第二条规定中的“境内公众”指境内不特定的公众，而仅在内部研发生成式人工智能，或公司基于内部辅助办公等目的引入并允许员工使用的行为，则并未落入《暂行办法》的适用范围。

《征求意见稿》	《暂行办法》
第二条 ~~研发、~~利用生成式人工智能产品，面向中华人民共和国境内公众提供服务的，适用本办法。 ~~本办法所称生成式人工智能，是指基于算法、模型、规则~~生成文本、图片、~~声音~~、视频、代码等内容的技术。	**第二条** 利用生成式人工智能技术向中华人民共和国境内公众提供生成文本、图片、音频、视频等内容的服务（以下称生成式人工智能服务），适用本办法。 国家对利用生成式人工智能服务从事新闻出版、影视制作、文艺创作等活动另有规定的，从其规定。 行业组织、企业、教育和科研机构、公共文化机构、有关专业机构等研发、应用生成式人工智能技术，未向境内公众提供生成式人工智能服务的，不适用本办法的规定。
	第二十条 对来源于中华人民共和国境外向境内提供生成式人工智能服务不符合法律、行政法规和本办法规定的，国家网信部门应当通知有关机构采取技术措施和其他必要措施予以处置。

但需注意的是，一方面这并不意味着对向不特定企业提供的生成式人工智能服务的排除适用；另一方面，这也不意味着研发、内部使用生成式人工智能无需受到任何约束。如前所述，考虑到可能的顶层通用立法《人工智能法》已被纳入立法计划，不排除前述行为将可能受到《人工智能法》为人工智能确立的通用基线要求约束。欧盟的《人工智能法案》同样采取类似思路，规定仅用于研发目的和非专业目的的人工智能系统，仍需要遵守透明度义务。研发、内部使用生成式人工智能仍需要加强透明性，对其应用同时也建议满足比如信息内容管理等要求。

就域外效力而言，《暂行办法》新增第二十条规定，在“向中华人民共和国境内公众提供”的基础上特别强调《暂行办法》对于境外生成式人工智能服务的域外效力。由此可见，相关企业无论是嵌入式集成，或是通过应用程序接口在产品和服务中封装等方式使用境外生成式人工智能服务，均应当充分考量该等境外服务在中国法下的合法合规性，以及因此导致的供应中断风险，乃至可能承担的连带责任等。

此外，尽管目前尚未有直接因接入境外生成式人工智能 API 的产品和服务受到处罚的案例，我们亦注意到，近日已有通过 API 方式接入 GPT 的应用程序，因“于境外 ChatGPT 有链入，未做网安备案，公司运维的两个网站未做网安备案，网络安全制度不完善，未履行网络安全义务等行为”受到公安部门行政处罚的案例。依据行政处罚决定书，处罚依据为《网络安全法》第五十九条第一款之规定，即对应的违法行为系未履行网络安全等级保护义务（第二十一条）以及未制定网络安全事件应急预案（第二十五条）之义务。就此我们理解，接入境外生成式人工智的产品和服务可能受到监管关注的可能性较高，应注意积极履行相关合规义务。

（二）聚焦内容治理：衔接《网络信息内容生态治理规定》

《暂行办法》第四条和第九条进一步明确了生成式人工智能服务提供者的网络信息内容生产者的主体身份，并有效衔接《网络信息内容生态治理规定》（“《信息内容治理规定》”）为网络信息内容生产者施加的义务。

1. 压实服务提供者的网络信息内容生产者责任

《信息内容治理规定》是网信办对中国境内的网络信息内容生态进行治理的主要规范，除了《暂行办法》第九条提及的“网络信息内容生产者”[3]，该规定还对“网络信息内容服务平台”[4] 施加了一定的合规要求。

---

[3] 网络信息内容生产者：《网络信息内容生态治理规定》第四十一条规定，网络信息内容生产者是指制作、复制、发布网络信息内容的组织或者个人。

[4] 网络信息内容服务平台：《网络信息内容生态治理规定》第四十一条规定，网络信息内容服务平台，是指提供网络信息内容传播的网络信息服务提供者。

如“明确适用范围与域外效力”一节所述，《暂行办法》的适用范围集中在互联网信息服务的场景中，提供者受到互联网信息内容治理相关规定规制的脉络较为清晰。

除可能作为网络信息内容服务平台承担《信息内容治理规定》项下义务外，和传统的网络信息内容服务平台相比，生成式人工智能服务的特殊性在于，除了本身可能提供平台服务外，生成式人工智能服务实际上参与了信息内容的生成，加之其广泛运用将在信息内容领域对使用者造成潜移默化的影响，因此，《暂行办法》特别要求服务提供者承担网络信息内容生产者的责任。

《征求意见稿》	《暂行办法》
第五条 利用生成式人工智能产品提供~~聊天和文本、图像、声音生成等~~服务的组织和个人~~（以下称“提供者”）~~，包括通过提供可编程接口等方式~~支持他人自行生成文本、图像、声音等~~，承担该产品生成内容生产者的责任；涉及个人信息的，承担个人信息处理者的法定责任，履行个人信息保护义务。 ~~第九条 提供生成式人工智能服务应当按照《中华人民共和国网络安全法》规定，要求用户提供真实身份信息。~~	**第九条** 提供者应当依法承担网络信息内容生产者责任，履行网络信息安全义务。涉及个人信息的，依法承担个人信息处理者责任，履行个人信息保护义务。 提供者应当与注册其服务的生成式人工智能服务使用者（以下称使用者）签订服务协议，明确双方权利义务。

2. 有效衔接有关违法信息和不良信息的规定

在明确服务提供者需承担网络信息内容生产者责任的基础上，《暂行办法》第四条进一步与《信息内容治理规定》第六条、第七条衔接，一方面，以结果为导向，禁止提供者生成违法信息；另一方，要求提供者应采取措施防止产生歧视信息。此外，由于使用者也构成《信息内容治理规定》规定的网络信息内容生产者，因此《暂行办法》第四条进一步补充了“使用生成式人工智能服务”的情形，使得其在信息内容治理领域的主体责任方面保持逻辑自洽。

此外，当前大家日益依赖互联网渠道获取信息资源，为了避免因 ChatGPT 等生成式人工智能已经出现的“胡编乱造”能力而误导公众，产生错误认识，《暂行办法》在要求不得生成违法信息的基础上，进一步要求提高生成内容的准确性和可靠性。

《征求意见稿》	《暂行办法》
第四条 提供生成式人工智能~~产品~~或服务应当遵守法律法规的要求，尊重社会公德、公序良俗，符合以下要求： （一）~~利用生成式人工智能生成的内容应当~~体现社会主义核心价值观，不得含有颠覆国家政权、推翻社会主义制度，煽动分裂国家、破坏国家统一，宣扬恐怖主义、极端主义，宣扬民族仇恨、民族歧视，暴力、淫秽色情信息，虚假信息，~~以及可能扰乱经济秩序和社会秩序的内容~~。 （二）在算法设计、训练数据选择、模型生成和优化、提供服务等过程中，采取措施防止出现种族、民族、信仰、国别、地域、性别、年龄、职业等歧视。 （三）尊重知识产权、商业道德，不得利用算法、数据、平台等优势实施不公平竞争。 ~~（四）利用生成式人工智能生成的内容应当真实准确，采取措施防止生成虚假信息。~~ （五）尊重他人合法利益，防止伤害他人身心健康，损害肖像权、名誉权和个人隐私，~~侵犯知识产权。禁止非法获取、披露、利用个人信息和隐私、商业秘密。~~	**第四条** 提供和使用生成式人工智能服务，应当遵守法律、行政法规，尊重社会公德和伦理道德，遵守以下规定： （一）坚持社会主义核心价值观，不得生成煽动颠覆国家政权、推翻社会主义制度，危害国家安全和利益、损害国家形象，煽动分裂国家、破坏国家统一和社会稳定，宣扬恐怖主义、极端主义，宣扬民族仇恨、民族歧视，暴力、淫秽色情，以及虚假有害信息等法律、行政法规禁止的内容； （二）在算法设计、训练数据选择、模型生成和优化、提供服务等过程中，采取有效措施防止产生民族、信仰、国别、地域、性别、年龄、职业、健康等歧视； （三）尊重知识产权、商业道德，保守商业秘密，不得利用算法、数据、平台等优势，实施垄断和不正当竞争行为； （四）尊重他人合法权益，不得危害他人身心健康，不得侵害他人肖像权、名誉权、荣誉权、隐私权和个人信息权益； （五）基于服务类型特点，采取有效措施，提升生成式人工智能服务的透明度，提高生成内容的准确性和可靠性。

值得注意的是，无论是境外还是境内（多为内测）生成式人工智能产品，均已经出现虚假信息的情况。除通过常见的事后纠偏措施以外，服务提供者如何采取有效措施，事先提供生成内容的准确性（比如专业领域的强化学习和事先纠偏等），将不仅仅是服务提供者的合规要求，也将是其服务竞争力的表现。

（三）风险进路：分级分类管理

《暂行办法》充分体现了我国对算法与人工智能治理的以风险为核心的分级分类监管机制。

1. 落实各行业领域生成式人工智能服务应用的分类分级监管机制

如前所述，我国在互联网信息内容领域已初具基于风险的分类分级制度雏形。具体而言，考虑到在互联网信息时代，信息快速传播对社会及公共利益产生的极大影响，《算法推荐规定》《深度合成规定》作为我国对互联网信息领域应用算法推荐技术、深度合成技术进行规制的主要文件，依据相关服务提供者是否具备舆论属性或社会动员能力，为相关

主体施加了不同层次的合规义务。此种基于是否具有舆论属性或社会动员能力的风险形成的分级要求也延续至了《暂行办法》。

《征求意见稿》	《暂行办法》
第六条 ~~利用生成式人工智能产品向公众提供服务前，应当按照《具有舆论属性或社会动员能力的互联网信息服务安全评估规定》向国家网信部门申报安全评估，并按照《互联网信息服务算法推荐规定》履行算法备案和变更、注销备案手续。~~	**第十七条** 提供具有舆论属性或者社会动员能力的生成式人工智能服务的，应当按照国家有关规定开展安全评估，并按照《互联网信息服务算法推荐规定》履行算法备案和变更、注销备案手续。

我们理解，《暂行办法》第十七条规定的安全评估与算法备案义务不仅延续了我国既有的算法治理手段，更体现了对生成式人工智能服务的风险分级要求。此前，在《征求意见稿》出台后，我们认为《征求意见稿》第六条对提供者履行算法备案与安全评估的义务的边界可能存在疑问。《暂行办法》对此作出了回应，延续了此前对算法推荐与深度合成服务提供者的分级，明确仅在“提供具有舆论属性或者社会动员能力”时，相关主体才需履行安全评估与算法备案义务。

值得注意的是，上述分级管理仅仅是冰山一角，《暂行办法》第十六条明确将分级分类监管的方法论延伸至生成式人工智能应用的全方位领域。也即，各行业主管部门均可以依据本行业内生成式人工智能技术本身的特点与在特定行业应用的特色进行分级分类监管，避免“一刀切”，使得各个风险级别的生成式人工智能服务一方面有能力在实践中充分发挥其作用，另一方面，其应用风险也可通过恰当、具有针对性的监管手段得到有效控制。

2. 为人工智能技术未来整体性分级分类提供先行示范

生成式人工智能技术仅是纷繁复杂的人工智能技术的一种，在人工智能治理过程中，除了规制特定人工智能技术类型、特定领域应用的精细化治理规则，规制整体人工智能技术及应用的通用立法也极具意义。

如前所述，我国此前在数据安全、个人信息保护以及互联网信息内容领域也体现了一定的风险分级分类思路，除了特定的行业内分级分类机制，整体的风险分级分类的思路在《暂行办法》中也得到了赓续。具体而言，《暂行办法》第三条提出国家对“生成式人工智能服务实行包容审慎和分级分类监管”。我们理解，这一规定不仅明确了分级分类在生成式人工智能服务这一特定技术服务领域的应用，更为重要的是，其为此后我国拟制定的通用性人工智能立法设置整体性的风险分级分类制度提供先行实践作用，为我国《人工智能法》预留了充分的发挥空间。

需要注意的是，我国虽然和欧盟类似，在人工智能治理领域均采取基于风险的分级分类制度。但是，基于各地区先天的文化及价值观差异，风险的视野可能存在差异。具体而言，欧盟《人工智能法案》在 2023 年发布的折衷草案中，明确了其在进行分级分类时，将人工智能系统对人们健康、安全、基本权利或环境的危害等因素涵盖在内。而我国未来如考虑对人工智能系统或技术进行整体规制，则可以基于人工智能规制内生的网络安全、数据安全、个人信息保护要求以及公共利益等因素为出发点，因地制宜，制定适宜我国实际需求的基于风险的分级分类制度。

（四）算法披露：与算法安全风险监测机制相适应

《暂行办法》对于算法披露义务的修订至少体现出两重深意，一是协调增强算法披露制度的体系建构，二是审慎考量算法透明与知识产权、商业秘密保护之间的平衡。

《征求意见稿》	《暂行办法》	《深度合成规定》	《算法推荐规定》
第十七条 提供者应当根据国家网信部门和有关主管部门的要求，提供~~可以影响用户信任、选择的必要信息，包括~~预训练和优化训练数据的来源、规模、类型、质量等描述，人工标注规则，~~人工标注数据的规模和类型~~，基础算法和技术体系等。	**第十九条** 有关主管部门依据职责对生成式人工智能服务开展监督检查，提供者应当依法予以配合，按要求对训练数据来源、规模、类型、标注规则、算法机制机理等予以说明，并提供必要的技术、数据等支持和协助。 参与生成式人工智能服务安全评估和监督检查的相关机构和人员对在履行职责中知悉的国家秘密、商业秘密、个人隐私和个人信息应当依法予以保密，不得泄露或者非法向他人提供。	**第二十一条** 网信部门和电信主管部门、公安部门依据职责对深度合成服务开展监督检查。深度合成服务提供者和技术支持者应当依法予以配合，并提供必要的技术、数据等支持和协助。	**第二十八条** 网信部门会同电信、公安、市场监管等有关部门对算法推荐服务依法开展安全评估和监督检查工作，对发现的问题及时提出整改意见并限期整改。 算法推荐服务提供者应当依法留存网络日志，配合网信部门和电信、公安、市场监管等有关部门开展安全评估和监督检查工作，并提供必要的技术、数据等支持和协助。 第二十九条 参与算法推荐服务安全评估和监督检查的相关机构和人员对在履行职责中知悉的个人隐私、个人信息和商业秘密应当依法予以保密，不得泄露或者非法向他人提供。

1. 协调增强算法披露制度的体系建构

与《征求意见稿》第十七条“应当根据国家网信部门和有关主管部门的要求”这一不

甚明确的适用前提相比，《暂行办法》明确了算法披露义务系服务提供者在有关主管部门依据职责开展监督检查时的配合义务。该等监督检查与我国现有算法治理框架一脉相承，属于算法安全风险常态化监测机制，是算法监管体系的重要组成部分。根据 2021 年网信办等九部委印发的《算法治理意见》，监管部门对算法的数据使用、应用场景、影响效果等开展日常监测工作，感知算法应用带来的网络传播趋势、市场规则变化、网民行为等信息，预警算法应用可能产生的不规范、不公平、不公正等隐患，发现算法应用安全问题。此外，这亦与《深度合成规定》《算法推荐规定》保持高度一致。

在此基础上，考虑到在《暂行办法》下，算法备案与安全评估义务均仅适用于“具有舆论属性或社会动员能力”的服务提供者，而前述算法披露义务适用于所有的服务提供者，可以理解其并非是向监管部门履行“额外”的算法披露义务，而是增强算法披露制度的完整性，并形成覆盖不同风险级别的生成式人工智能的监管体系。

2. 审慎考量算法透明与知识产权、商业秘密保护之间的平衡

就披露范围而言，《暂行办法》不但摒弃了“以影响用户信任、选择的必要信息”“基础算法和技术体系”等相对模糊不清的表述，代之以“算法机制机理”，还明确规定获知披露信息的相关机构和人员的保密义务。

一方面，这一修订一定程度上直接回应了实践中相关企业对于源代码等可能被纳入披露范围而导致的对知识产权、商业秘密保护等方面的隐忧，有助于保护企业的发展动力。另一方面，如前所述，考虑到该等算法披露义务与算法安全风险监测机制相适应，可以理解对于训练数据来源、规模、算法机制机理等披露要求的限度，将限于监管部门履行其防范算法滥用风险等方面的监管职责，而不至于毫无边界，甚至于算法绝对透明。这似乎旨在呼应这样一种主张：对算法的透明度要求不应高于人类——我们只需要人类决策者在问责和透明程序中提供足够的信息和理由，而不是以上帝之眼洞察他们的认知过程 [5]。

此外，对“算法机制机理”的理解在实践中已有章可循。目前，已按照《深度合成规定》《算法推荐规定》履行算法备案义务的企业，可以在互联网信息服务算法备案系统中查询其拟公示算法机制机理内容，其中包含备案算法的基本原理、运行机制、应用场景、目的意图等信息。

（五）审慎包容：责任主体义务限缩

《暂行办法》对于算法纠偏及训练数据的数据质量要求相关规定修改，适度为生成式

[5] Zerilli, J., Knott, A., Maclaurin, J., Gavaghan, C.: Transparency in algorithmic and human decision-making: is there a double standard? Philos. Technol. 32(4), 661–683 (2019).

人工智能服务提供者“松绑”，充分体现出坚持发展和安全并重，平衡企业利益、公共利益、个人信息保护等多方诉求的包容审慎监管态度。

1. 算法纠偏与报告义务

如我们在《不要温和地走进那良夜——对 < 生成式人工智能服务管理办法 > 的思考》[6]一文中所述，尽管一定程度上的算法透明度是算法规制可欲实现的目标，也常被认为是算法程序性规制强有力的监管工具，但在实践中，由于算法黑箱的存在会影响技术人员进行算法纠偏行为的有效性，若采取“防止不当内容再次生成”等基于结果的规制方案，可能致使责任主体履行义务存在较大障碍。因此，《暂行办法》第十四条将算法纠偏义务限于“及时采取停止生成、停止传输、消除等处置措施”具备更强的可操作性，亦能避免相关企业为此支出高昂但或“徒劳无功”的合规成本。

此外，《暂行办法》还要求企业保存采取处置措施的有关记录，并向有关主管部门报告。尽管具体的报告内容将有待明确，这意味着“采取有效措施”并非被置于监管真空，而是同样需经由监管部门评估、以有效固定生成式人工智能违法行为的问责点。

《征求意见稿》	《暂行办法》
第十五条 对于运行中发现、用户举报的不符合本办法要求的生成内容，除采取内容过滤等措施外，应在 3 个月内通过模型优化训练等方式防止再次生成。	**第十四条** 提供者发现违法内容的，应当及时采取停止生成、停止传输、消除等处置措施，采取模型优化训练等措施进行整改，并向有关主管部门报告。 提供者发现使用者利用生成式人工智能服务从事违法活动的，应当依法依约采取警示、限制功能、暂停或者终止向其提供服务等处置措施，保存有关记录，并向有关主管部门报告。

2. 训练数据的数据质量要求

生成式人工智能模型在预训练和调优过程中，有赖于大规模语料库等海量数据的“喂养”，然而越多的数据并不等于越好的数据。如果训练数据存在错误、噪声或偏差，模型可能会学习到不准确或不可靠的信息，从而影响模型的泛化能力。此外，如若数据集中存在虚假与歧视性内容等，模型的输出结果中也可能镜像呈现相应内容。因此，对训练数据的质量控制，实质是对确保生成式人工智能的可靠性与可信度，防止生成违法、不良信息等的源头控制。因此，无论是《暂行办法》还是欧盟《人工智能法案》，均对训练数据提出了一定的质量要求。

[6] 参见文章《不要温和地走进那良夜——对 < 生成式人工智能服务管理办法 > 的思考》，https://mp.weixin.qq.com/s/QLzHdO45FkSqeTRem0WboQ。

《征求意见稿》	《暂行办法》
第七条 提供者应当对生成式人工智能产品的预训练数据、优化训练数据来源的合法性负责。 用于生成式人工智能产品的预训练、优化训练数据，应满足以下要求： （一）符合《中华人民共和国网络安全法》等法律法规的要求； （二）不含有侵犯知识产权的内容； （三）数据包含个人信息的，应当征得个人信息主体同意或者符合法律、行政法规规定的其他情形； （四）能够保证数据的真实性、准确性、客观性、多样性； （五）国家网信部门关于生成式人工智能服务的其他监管要求。	**第七条** 生成式人工智能服务提供者（以下称提供者）应当依法开展预训练、优化训练等训练数据处理活动，遵守以下规定： （一）使用具有合法来源的数据和基础模型； （二）涉及知识产权的，不得侵害他人依法享有的知识产权； （三）涉及个人信息的，应当取得个人同意或者符合法律、行政法规规定的其他情形； （四）采取有效措施提高训练数据质量，增强训练数据的真实性、准确性、客观性、多样性； （五）《中华人民共和国网络安全法》《中华人民共和国数据安全法》《中华人民共和国个人信息保护法》等法律、行政法规的其他有关规定和有关主管部门的相关监管要求。

《暂行办法》在《征求意见稿》提出的训练数据“真实性、准确性、客观性、多样性”的基础上，限缩了生成式人工智能服务提供者对训练数据质量应尽的义务履行边界，从“能够保证”训练数据质量减轻至“提高”训练数据质量、“增强”训练数据的真实性、准确性、客观性以及多样性。

上述变化为由结果为导向减轻至以行为为导向。我们理解，这种变化在一定程度上打消了提供者对使用训练数据、扩大训练数据范围的顾虑，有助于其积极开展相关的数据训练活动，优化人工智能，从而使公众受益。毕竟，即使是大企业，也难以保证自身使用的海量训练数据均是真实、准确且客观的。因此，《暂行办法》调整了相关义务，即使客观上存在训练数据不真实、不准确、不客观的可能性，但不影响生成式人工智能服务提供者通过数据标注等手段，尽可能减少相关风险。

与真实性、准确性和客观性相较而言，训练数据的多样性更需从多维度进行分析与判断。生成式人工智能技术在训练优化时，单一范围的训练数据可能致使生成结果存在不足与遗漏。以 ChatGPT 为例，当 ChatGPT 的训练数据以英文语料资料为主时，虽然其可基于对话者的指示对英文写作进行充分的改写，但是在组织中文时，能力则稍显逊色。基于此，建议相关企业以生成式人工智能服务预期应用的具体领域为出发点，充分考虑不同行业的差异化与特殊性，选择恰当的训练数据范围。

四、对《暂行办法》的思考与展望

（一）“举重若轻”，敏捷与审慎监管相依，科学进步成为新主题

ChatGPT 等生成式人工智能在全球掀起热潮后，我国监管部门需要迅速研讨风险，在短时间内形成征求意见稿，时间紧任务重。同时为了平衡发展与安全，在广泛征求社会意见后，《暂行办法》大量删减争议条款，在鼓励新技术在合理框架内发展的主基调下，在技术涌现不到六个月时间内最终定稿，体现出我国在社会面临新技术引发的新旧风险时进行“敏捷治理”的积极态度。

在“敏捷治理”的基础上，基于生成式人工智能服务提供者的现实技术与能力困境，同时避免采取欧盟等在科技领域的强势态度从而可能致使人工智能创新受阻的负面影响，我们理解《暂行办法》在《征求意见稿》的基础上，充分体现了我国对人工智能等科技领域发展所采取的“包容审慎态度”，在促发展与稳安全之间寻找动态的平衡点。

值得一提的是，除《网络安全法》《数据安全法》《个人信息保护法》的“三驾马车”外，《暂行办法》还将《科学技术进步法》明确纳入其上位法的范畴。《科学技术法》作为我国法律层级的规定，明确提出“发挥科学技术第一生产力、创新第一动力”的目标；《暂行办法》作为其下位法，将相关机构研发生成式人工智能服务（不面向境内公众提供）的情形排除在管辖范围外，并支持鼓励行业组织等技术创新，“牢牢把握建设世界科技强国的战略目标”，“抓抢全球科技发展先机”。

（二）规范先行，可期待与《人工智能法》相衔接

如前所述，结合《国务院 2023 年度立法工作计划》，我国的《人工智能法》已经正式进入了立法进程中。《暂行办法》是我国人工智能整体监管的及时应对和有益尝试，随着技术发展和行业实践的推移，“行业规制 + 通用立法”的模式将成为人工智能治理的主流手段。可以预计在未来，人工智能立法不论是从横向层面还是从纵向层面，都会日臻全面和完善。

如前所述，我国自《数据安全法》起，即构建了一种基于风险的分类分级监管思路。《暂行办法》第三条和第十六条均明确提及我国在生成式人工智能治理方面将采取“分类分级监管”，这也为正在制定的《人工智能法》就人工智能整体规制层面进行分类分级监管提供了充足的空间。

（三）吸收国际共识，传递中国价值，分享中国方案

一方面，《暂行办法》制定、征求意见的过程中，全球范围内生成式人工智能治理的

相关讨论热度居高不下。欧盟自 2021 年发布《人工智能法案》提案后，进行多次协商与沟通，以确定恰当的监管规则。2024 年 3 月 13 日，欧洲议会以 523 票赞成、46 票反对和 49 票弃权的表决结果通过了《人工智能法案》，并预计在同年 4 月交由欧盟理事会批准，以使得《人工智能法案》最终成为具有强制效力的正式法律。欧盟及其他地区的立法活动为我国提供了参考，在制定《暂行办法》时，我国也对国际先进、有借鉴意义的经验进行吸收学习，博采众长，合理借鉴多数司法管辖区例如训练数据治理等监管要求，体现出《暂行办法》的国际先进性。

另一方面，作为世界范围内首部正式通过的生成式人工智能服务法规，《暂行办法》立足我国实践，通过第四条中要求尊重“社会公德和伦理”，确定中国生成式人工智能的伦理和道德原则。此外《暂行办法》还鼓励生成式人工智能技术的自主创新，并积极通过推进公共训练数据资源平台、推动公共数据共享开放、促进算力资源协同共享等措施，为企业就生成式人工智能算法、框架、芯片等的研发保驾护航。

最后，《暂行办法》第六条明确鼓励生成式人工智能算法、框架、芯片及配套软件平台等基础技术的自主创新，平等互利开展国际交流与合作，参与生成式人工智能相关国际规则制定，也旨在将中国价值和中国方案进一步对外推广。

（四）“条修叶贯”，企业宜加快完善内部合规制度体系

对于企业而言，不仅要着手履行《暂行办法》及现行算法与人工智能治理框架合规义务，更是要思考如何通过技术、制度和组织工具，建立一个内部逻辑自洽、行之有效的人工智能合规体系，以应对未来可能到来的各种监管举措。

整体而言，企业宜从人工智能全生命周期的角度思考如何搭建内部合规体系，包括但不限于：

1. 人工智能风险管理组织；
2. 人工智能道德伦理标准；
3. 人工智能开发管理制度；
4. 人工智能数据安全与合规管理规范；
5. 人工智能技术和服务监测制度；
6. 人工智能安全应急处理要求；
7. 科技伦理审查制度等。

同时，还需要考虑到企业需履行算法纠偏与报告等义务，企业应当建立内部记录机制，

以实现对人工智能开发和使用的信息记录，以满足算法透明性义务和及时响应监管的要求，有助于企业进行合规自证。

后记

在《征求意见稿》出台后，我们在《不要温和地走进那良夜——对 < 生成式人工智能服务管理办法 > 的思考》[7] 一文中曾呼唤，面对生成式人工智能部分不可测风险，“即使眼前幽暗丛生，我们也‘不要温和地走入那良夜’（Do not go gentle into that good night）”，应当以“以最大的热情鼓励和促进技术的发展，同时以最审慎的心态、最大的关注力观察人工智能的发展动向，秉承技术中立的价值，审慎包容但坚守底线”。在《暂行办法》“落地”后，我们欣喜地发现“发展和安全并重、促进创新和依法治理”真正成为监管原则，“包容审慎”也体现在具体的监管要求和方式中。

尽管我国成功发布全球首部生成式人工智能法规，可谓是“今宵绝胜无人共”，但在“欢欣鼓舞”的同时，我们仍不能放松警惕。不仅新技术新应用正在以史无前例的速度、强度和广度影响社会，各国未来陆续出台的人工智能监管规则也将引发多样化的思潮，产业发展和社会治理势必迎来第二波的冲击。在人工智能领域，监管部门如何开拓监管思路、抓准监管时机、更新监管工具，产业如何顺应合规发展方向、主动管理人工智能、快速创新应用将会持续影响各国的国际竞争力。结合《暂行办法》的成功经验，我们相信，产业与监管的有效沟通，相互理解与共同治理，将能协助大家携手“卧看星河尽意明”，最终走出幽暗，迎来更加璀璨的国家及人类的未来。

感谢实习生王艺捷对本文作出的贡献。

[7] 同上。

《生成式人工智能服务安全基本要求》要点解析

张逸瑞 冯宝宝 朱佳蔚 张津豪

随着“Sora”等多模态高性能生成式人工智能的相继出现，全球范围内针对生成式人工智能服务安全的监管呼声也日渐高涨。2023 年 7 月，国家互联网信息办公室、国家发展和改革委员会等七部门联合颁布《生成式人工智能服务管理暂行办法》（“《AIGC 暂行办法》”），在延续《互联网信息服务算法推荐管理规定》《互联网信息服务深度合成管理规定》所规定的算法备案的基础上，另行确立了针对生成式人工智能大模型的安全评估备案（“大模型备案”）。但是，《AIGC 暂行办法》对于大模型备案所关注的生成式人工智能服务安全的具体要求、评估参数、评估标准等并未进行细化解释。

2023 年 10 月 11 日，全国网络安全标准化技术委员会颁布了《生成式人工智能服务安全基本要求（征求意见稿）》（“征求意见稿”），并于 2024 年 3 月 1 日正式颁布了《生成式人工智能服务安全基本要求（TC260-003）》（“正式文件”）。《生成式人工智能服务安全基本要求》（“《AIGC 安全要求》”）作为国家专业标准化技术委员会发布的技术文件，在生成式人工智能服务安全的原则性要求方面提供了细化指引，为包括大模型备案在内的人工智能安全监管制度提供了评价工具，为各类生成式人工智能服务提供者开展安全评估、提高安全水平提供了参考。

本文将就《AIGC 安全要求》的正式文件相较于征求意见稿的重点修订内容进行梳理分析，并对生成式人工智能服务提供者为符合《AIGC 安全要求》的规定可以考虑设立的合规制度提出基础建议。

一、《生成式人工智能服务安全基本要求》正式文件与征求意见稿主要内容的对比与总结

（一）第 3 条（术语和定义）

1. 第 3.1 条（生成式人工智能服务 generative artificial intelligence service，“生成式人工智能服务”）

征求意见稿	正式文件
第 3.1 条（生成式人工智能服务 generative artificial intelligence service，**“生成式人工智能服务”**）：~~基于数据、算法、模型、规则，能够根据使用者提示~~生成文本、图片、音频、视频等内容的~~人工~~智能服务。	第 3.1 条（生成式人工智能服务 generative artificial intelligence service，**“生成式人工智能服务”**）：**利用生成式人工智能技术向中华人民共和国境内公众提供**生成文本、图片、音频、视频等内容的服务。

本定义照应了《AIGC 暂行办法》第二条对“生成式人工智能服务”的定义[1]。正式文件在对“生成式人工智能服务”的定义中删除了征求意见稿中“基于数据、算法、模型、规则”的表达，增加“利用生成式人工智能技术向中华人民共和国境内公众提供”，使得《AIGC 安全要求》中“生成式人工智能服务”这一概念的定义与《AIGC 暂行办法》中的定义保持统一。

根据本定义，结合《AIGC 暂行办法》第二条规定可知，目前我国针对生成式人工智能行业的监管侧重于针对面向中华人民共和国境内的公众提供生成式人工智能服务的组织或个人。对生成式人工智能技术进行单纯的内部研发和应用，不涉及向境内公众提供服务的行业组织、企业、教育和科研机构、公共文化机构等专业机构，并非本轮监管重点关注的对象。但是，该等企业仍需根据具体情况遵循《中华人民共和国数据安全法》（“《数据安全法》”）、《中华人民共和国网络安全法》（“《网络安全法》”）等法律法规，并且可以在一定程度上参考《AIGC 暂行办法》以及《AIGC 安全要求》对于生成式人工智能服务安全的要求，以应对未来的合规动态。

2. 第 3.2 条（提供者 service provider，“服务提供者”）

征求意见稿	正式文件
第 3.2 条（提供者 service provider，**“服务提供者”**）：以交互界面、可编程接口等形式~~面向我国境内公众~~提供生成式人工智能服务的组织或个人。	第 3.2 条（提供者 service provider，**“服务提供者”**）：以交互界面、可编程接口等形式提供生成式人工智能服务的组织或个人。

[1] 《AIGC 暂行办法》第二条规定，“生成式人工智能服务”是指“利用生成式人工智能技术向中华人民共和国境内公众提供生成文本、图片、音频、视频等内容的服务”。

本定义照应了《AIGC 暂行办法》第二十二条对“生成式人工智能服务提供者”的定义[2]。结合本文第 1.1 条的分析可知，正式文件对“提供者”的定义与《AIGC 暂行办法》中对“生成式人工智能服务提供者”的定义亦基本相同。实践中，API、SDK 等接口服务以及网页、移动应用、小程序等交互界面为目前 B 端及 C 端用户调用生成式人工智能的主流方式，故《AIGC 安全要求》在服务提供方式的列举中增加了“以交互界面”的方式提供生成式人工智能服务的情况，随着人工智能行业的高速发展，不排除未来可能出现新的调用方式。

3. 第 3.5 条（基础模型 foundation model）

征求意见稿	正式文件
无该条款。	第 3.5 条（基础模型 foundation model）：**在大量数据上训练的，用于普适性目标、可优化适配多种下游任务的深度神经网络模型。**

在征求意见稿及正式文件中，均在第 6 条（模型安全要求）中使用了“基础模型”这一概念，本定义是对“基础模型”这一概念的进一步解释，明确了需要经过主管部门备案后方可用于提供生成式人工智能服务的基础模型，区别于普通的计算机模型，是指具备深度神经网络结构的、经过优化和训练能够适配多种下游任务的通用大模型。

本定义在一定程度上照应了《AIGC 暂行办法》第七条对生成式人工智能服务提供者开展训练数据处理活动时使用合规基础模型的要求[3]。实践中，国内生成式人工智能服务行业对基础模型的调用通常分为三种类型：

类别	特点	示例
纯自研的基础模型	能够全面掌握基础模型背后的核心算法和运行规则，并独立负责处理数据训练、生成内容标记、模型优化等所有技术性事项	智谱 AI 发布的 GLM-4 国产全自研大模型[4]、腾讯发布的自研混元大模型[5]

[2] 《AIGC 暂行办法》第二十二条规定，“生成式人工智能服务提供者，是指利用生成式人工时能技术提供生成式人工智能服务（包括通过提供可编程接口等方式提供生成式人工智能服务）的组织、个人”。

[3] 《AIGC 暂行办法》第七条规定，“生成式人工智能服务提供者（以下称提供者）应当依法开展预训练、优化训练等训练数据处理活动，遵守以下规定：（一）使用具有合法来源的数据和基础模型；……”

[4] https://www.ceweekly.cn/company/2024/0116/434891.html.

[5] https://www.tencent.com/zh-cn/articles/2201685.html.

类别	特点	示例
半自研的基础模型	在第三方基础模型的基础上进行二次训练、精确调优，形成适配于自身产品的基础模型，其仅能独立掌握自身研发的增量部分	华东理工大学·X-D Lab（心动实验室）基于开源的通义千问开源模型开发了心理健康大模型 MindChat（漫谈）、医疗健康大模型 Sunsimiao（孙思邈）、教育 / 考试大模型 GradChat（锦鲤）等 [6]
完全调用第三方基础模型（不做任何调优）	对于基础模型本身无法进行任何干涉，亦无法参与任何实质的技术事项	百度千帆大模型平台操作台中的“模型仓库”存在多个第三方模型可以直接供企业和开发者调用、部署 [7]

在这三种情况下，相应的服务提供者在《AIGC 暂行办法》以及《AIGC 安全要求》下所需要履行的义务轻重亦有所不同，需要根据具体情况进一步分析。

（二）第 4 条（总则）

征求意见稿	正式文件
第 4 条（总则）：本文件支撑《生成式人工智能服务管理暂行办法》，提出了提供者需遵循的安全基本要求。提供者~~在向相关主管部门提出生成式人工智能服务上线的备案申请前，应~~按照本文件~~中各项~~要求~~逐条~~进行安全性评估，并将评估结果~~以及证明材料在备案时提交~~。 除本文件提出的基本要求外，提供者还应自行按照我国法律法规以及国家标准相关要求做好网络安全、数据安全、个人信息保护等方面的其他安全工作。	第 4 条（总则）：本文件支撑《生成式人工智能服务管理暂行办法》，提出了**服务**提供者需遵循的安全基本要求。**服务**提供者在**按照有关要求履行**备案**手续时**，按照本文件**第 9 章**要求进行安全评估，并**提交**评估**报告**。 除本文件提出的基本要求外，**服务**提供者应自行按照我国法律法规以及国家标准相关要求做好网络安全、数据安全、个人信息保护等方面的其他安全工作。**服务提供者应紧密注意生成式人工智能可能带来的长期风险，谨慎对待可能具备欺骗人类、自我复制、自我改造能力的人工智能，并重点关注生成式人工智能可能被用于编写恶意软件、制造生物武器或化学武器等安全风险。**

正式文件中增加了对生成式人工智能长期风险、伦理风险的警示内容，体现对目前全球范围内针对生成式人工智能对人类社会可能造成的潜在风险（包括网络安全和生物技术等领域的风险，放大虚假信息风险、伦理风险等）的高度关切。

[6] https://news.sciencenet.cn/htmlnews/2023/12/513458.shtm.
[7] https://developer.baidu.com/article/detail.html?id=1099866.

（三）第 5 条（语料安全要求）

1. 第 5.1 条（语料来源安全要求）

（1）第 5.1 条 a）（语料来源管理方面）

征求意见稿	正式文件
第 5.1 条（语料来源安全要求）a）（语料来源管理方面）：~~1）应建立语料来源黑名单，不使用黑名单来源的数据进行训练；~~2）~~应对各来源语料进行安全评估，单一~~来源语料~~内容中~~含违法不良信息超过 5% 的~~应将该来源加入黑名单~~。	第 5.1 条（语料来源安全要求）a）（语料来源管理方面）：1）**面向特定语料来源进行采集前，应对该来源语料进行安全评估，语料内容中含违法不良信息超过 5% 的，不应采集该来源语料**；2）**面向特定**语料**来源**进行**采集后，应对所采集的该**来源语料**进行核验**，含违法不良信息**情况**超过 5% 的，**不应使用该来源语料进行训练**。

正式文件中删除了“语料来源黑名单”制度，而修改为对来源语料本身进行安全评估的制度。我们理解，在实践中，同一语料来源项下可能存在大批量的语料，因其中某一批语料产生了安全问题而舍弃某一语料来源会造成较大的语料损失。在当下基础模型行业已经开始出现语料紧缺、训练数据不足的风险的情况下，“语料来源黑名单”制度并不利于优化生成式人工智能服务。正式文件进一步将来源语料安全制度细分为“采集前”和“采集后 + 训练前”两个阶段，要求服务提供者对语料进行双重安全评估，确保语料来源安全。

（2）第 5.1 条 c）（语料来源可追溯方面）

征求意见稿	正式文件
第 5.1 条（语料来源安全要求）c）（语料来源可追溯方面）：2）使用自采语料时，应具有采集记录，不应采集他人已明确声明不可采集的语料； 注 2：自采语料包括自行生产的语料以及从互联网采集的语料。 注 3：声明不可采集的方式~~包括但不限于~~robots 协议等。 3）使用商业语料时： ——应有具备法律效力的交易合同、合作协议等； ——交易方或合作方不能提供~~语料合法~~性证明材料时，不应使用该语料。	第 5.1 条（语料来源安全要求）c）（语料来源可追溯方面）：2）使用自采语料时，应具有采集记录，不应采集他人已明确不可采集的语料； 注 2：自采语料包括自行生产的语料以及从互联网采集的语料。 注 3：明确不可采集的语料，**例如已通过** robots 协议**或其他限制采集的技术手段明确表明不可采集的网页数据，或个人已拒绝授权采集的个人信息等**。 3）使用商业语料时： ——应有具备法律效力的交易合同、合作协议等； ——交易方或合作方不能提供语**料来源、质量、安全等方面的承诺以及相关**证明材料时，不应使用该语料； ——**应对交易方或合作方所提供语料、承诺、材料进行审核**。

本条系对语料来源合规性追溯的规定。语料的来源分为自采语料与商业语料。所谓自采语料，是指自行生产以及从互联网采集的语料。《AIGC 安全要求》规定服务提供者应当采取措施从源头追溯并确保自采语料合规性，包括（1）保存采集记录；（2）不采集他人已明确不可采集的语料，包括（a）通过 robots 协议等技术手段标明不可采集的网页数据，以及（b）正式文件增加的“个人已拒绝授权采集的个人信息”。我们理解，正式文件中将个人拒绝授权采集的个人信息明确列举于不得采集的语料中，系对《AIGC 暂行办法》第七条的进一步细化[8]。

关于商业语料，我们理解主要指并非由服务提供者自行采集，而是通过与第三方语料提供方进行交易获得的语料。《AIGC 安全要求》规定服务提供者应当采取措施，确保商业语料的合规性，包括（1）与语料提供方签署有效的交易文件；（2）要求语料提供方提供语料来源、质量、安全等承诺以及相关证明材料并进行审核。本条在一定程度上明确了商业语料交易所需遵循的合规性要求，但仍遗留了一些问题，有待进一步在实践中摸索出答案，例如，语料接收方对语料提供方所提供的承诺以及证明材料需尽到何种程度的审核义务，才能够被认定为适当地履行了合规要求等。

2. 第 5.2 条（语料内容安全要求）

（1）第 5.2 条 b）（知识产权方面）

征求意见稿	正式文件
第 5.2 条（语料内容安全要求）b）（知识产权方面）：2）语料用于训练前，~~知识产权相关负责人~~等应对语料中的知识产权侵权情况进行识别，提供者不应使用~~有侵权问题的~~语料进行训练： ——训练语料包含文学、艺术、科学作品的，应重点识别训练语料以及生成内容中的著作权侵权问题； ~~——对训练语料中的商业语料以及使用者输入信息，应重点识别侵犯商业秘密的问题；~~	第 5.2 条（语料内容安全要求）b）（知识产权方面）：2）语料用于训练前，应对语料中的**主要**知识产权侵权**风险**进行识别，**发现存在知识产权侵权等问题的，服务**提供者不应使用**相关**语料进行训练；**例如**，语料**中**包含文学、艺术、科学作品的，应重点识别语料以及生成内容中的著作权侵权问题；

正式文件删除了对“侵犯商业秘密”的重点识别要求，我们理解，一方面，对于泄露他人商业秘密这一语料安全风险，正式文件中已经于附录 A 中进行了明确；另一方面，商业秘密的识别是一项难度和成本较高的任务，可能会对人工智能产业的初期发展带来较大的压力。

[8] 《AIGC 暂行办法》第七条规定，“生成式人工智能服务提供者应当依法开展预训练、优化训练等训练数据处理活动，遵守以下规定：……（三）涉及个人信息的，应当取得个人同意或者符合法律、行政法规规定的其他情形；……（五）《中华人民共和国网络安全法》《中华人民共和国数据安全法》《中华人民共和国个人信息保护法》等法律、行政法规的其他有关规定和有关主管部门的相关监管要求。”

（2）第 5.2 条 c）（个人信息方面）

征求意见稿	正式文件
第 5.2 条（语料内容安全要求）c）（个人信息方面）：1）~~应使用包含个人信息的语料时，获得对应个人信息主体的授权同意，或满足其他合法使用该个人信息的条件~~； 2）~~应使用包含敏感个人信息的语料时，获得对应个人信息主体的单独授权同意，或满足其他合法使用该敏感个人信息的条件~~； ~~3）应使用包含人脸等生物特征信息的语料时，获得对应个人信息主体的书面授权同意，或满足其他合法使用该生物特征信息的条件。~~	第 5.2 条（语料内容安全要求 c）（个人信息方面）：1）**在**使用包含个人信息的语料**前，应取得对应个人同意或者符合法律、行政法规规定的其他情形；** 2）**在**使用包含敏感个人信息的语料**前，应取得对应个人单独同意或者符合法律、行政法规规定的其他情形。**

本条照应了《AIGC 暂行办法》中第七条对生成式人工智能服务提供者开展训练数据处理活动时使用个人信息的合规要求[9]。正式文件较征求意见稿而言，主要有以下几点调整：

首先，正式文件将“使用时”的表述修改为“使用前”。我们理解，一方面，该等修改符合《中华人民共和国个人信息保护法》（“《个保法》”）关于个人信息处理者在处理个人信息前的告知义务的规定[10]。而根据《个保法》第四条规定，个人信息的使用即属于个人信息的处理的一种情形[11]。因此，从遵循《个保法》规定的角度看，服务提供者应当在使用包含个人信息的语料之前，而非之时，即取得对应的个人同意或者符合法律、行政法规规定的其他情形。另一方面，服务提供者对语料的使用往往是多次、同时、大批量的使用，因此，要求其在使用的同时开始履行合规要求并不具备实操性，而是应当在使用乃至获取语料之前即获得个人同意，或者确保其符合法律法规的要求。

其次，正式文件删去了“应使用包含人脸等生物特征信息的语料时，获得对应个人信息主体的书面授权同意，或满足其他合法使用该生物特征信息的条件。”根据《个保法》第二十八条规定，敏感个人信息包含生物识别信息[12]。因此，即使正式文件中删去本条，由于上述第 2）条的“敏感个人信息”在《个保法》的定义中已经包括了生物识别信息，因此并不会减轻服务提供者处理个人信息时需要履行的义务。

第三，正式文件将“或满足其他合法使用该个人信息的条件”的表述一律调整为“应

---

[9] 《AIGC 暂行办法》第七条规定，“生成式人工智能服务提供者（以下称提供者）应当依法开展预训练、优化训练等训练数据处理活动，遵守以下规定：……（三）涉及个人信息的，应当取得个人同意或者符合法律、行政法规规定的其他情形；……。”

[10] 《中华人民共和国个人信息保护法》第十七条规定，“个人信息处理者在处理个人信息前，应当以显著方式、清晰易懂的语言真实、准确、完整地向个人告知下列事项：（一）个人信息处理者的名称或者姓名和联系方式；（二）个人信息的处理目的、处理方式，处理的个人信息种类、保存期限；（三）个人行使本法规定权利的方式和程序；（四）法律、行政法规规定应当告知的其他事项。前款规定事项发生变更的，应当将变更部分告知个人。”

[11] 《中华人民共和国个人信息保护法》第四条规定，“个人信息是以电子或者其他方式记录的与已识别或者可识别的自然人有关的各种信息，不包括匿名化处理后的信息。个人信息的处理包括个人信息的收集、存储、使用、加工、传输、提供、公开、删除等。”

[12] 《中华人民共和国个人信息保护法》第二十八条规定，“敏感个人信息是一旦泄露或者非法使用，容易导致自然人的人格尊严受到侵害或者人身、财产安全受到危害的个人信息，包括生物识别……等信息。”

取得对应个人同意或者符合法律、行政法规规定的其他情形”。我们理解，这一调整主要是为了与《AIGC 暂行办法》第七条的表述保持一致。

3. 第 5.3 条（语料标注安全要求）

（1）第 5.3 条 a）（标注人员方面）

征求意见稿	正式文件
第 5.3 条（语料标注安全要求）a）（标注人员方面）：1）应自行对标注人员进行考核，给予合格者标注资质，并有定期重新培训考核以及必要时暂停或取消标注资质的机制；	第 5.3 条（语料标注安全要求）a）（标注人员方面）：**1）应自行组织对于标注人员的安全培训，培训内容应包括标注任务规则、标注工具使用方法、标注内容质量核验方法、标注数据安全管理要求等；** 2）应自行对标注人员进行考核，给予合格者标注**上岗资格**，并有定期重新培训考核以及必要时暂停或取消标注**上岗资格**的机制，**考核内容应包括标注规则理解能力、标注工具使用能力、安全风险判定能力、数据安全管理能力等；**

本条内容照应了《AIGC 暂行办法》第八条中对生成式人工智能服务提供者在技术研发过程中进行数据标注的要求，并对该等要求进行了细化和进一步拓展[13]。正式文件相较于征求意见稿，增加了对于标注人员的安全培训以及具体培训内容的要求，并进一步明确了标注人员的考核机制。

（四）第 6 条（模型安全要求）（对服务提供者的要求）

1. 第 6 条 b）（模型生成内容安全方面）

征求意见稿	正式文件
第 6 条 b）模型生成内容安全方面： 1）在训练过程中，应将生成内容安全性作为评价生成结果优劣的主要考虑指标之一； 2）在每次对话中，应对使用者输入信息进行安全性检测，引导模型生成积极正向内容； 3）对~~提供服务过程中以及定期检测时~~发现的安全问题，应通过针对性的指令微调、强化学习等方式优化模型。 注：模型生成内容是指模型直接输出的、未经其他处理的原生内容。	第 6 条 b）模型生成内容安全方面： 1）在训练过程中，应将生成内容安全性作为评价生成结果优劣的主要考虑指标之一； 2）在每次对话中，应对使用者输入信息进行安全性检测，引导模型生成积极正向内容； 3）**应建立常态化监测测评手段，对监测测评**发现的**提供服务过程中的**安全问题，**及时处置并**通过针对性的指令微调、强化学习等方式优化模型。 注：模型生成内容是指模型直接输出的、未经其他处理的原生内容。

正式文件将征求意见稿中对模型生成内容的“定期检测”要求具象化为要求服务提供者建立常态化监测测评手段，强调监测测评是持续进行的过程，进一步确保提供服务全过

[13] 《AIGC 暂行办法》第八条的规定，“在生成式人工智能技术研发过程中进行数据标注的，提供者应当制定符合本办法要求的清晰、具体、可操作的标注规则；开展数据标注质量评估，抽样核验标注内容的准确性；对标注人员进行必要培训，提升尊法守法意识，监督指导标注人员规范开展标注工作。”

程的安全性。另外，正式文件中增加了及时处置安全问题的表述，对服务提供者提出问题处理的时效性要求，防止安全问题扩大带来进一步影响。

2. 第 6 条 c）（生成内容准确性方面） d）（生成内容可靠性方面）

征求意见稿	正式文件
第 6 条 d）生成内容准确性方面：生成内容~~应准确~~响应使用者输入意图~~，所包含~~的数据及表述~~应符合~~科学常识~~或~~主流认知~~、不含~~错误内容。 e）生成内容可靠性方面：~~服务按照使用者指令给出的回复~~，应格式框架合理~~、~~有效内容含量高~~，应能够有效帮助使用者解答问题~~。	第 6 条 c）生成内容准确性方面：**应采取技术措施提高**生成内容响应使用者输入意图**的能力，提高生成内容中**数据及表述**与**科学常识**及**主流认知**的符合程度，减少其中的**错误内容。 d）生成内容可靠性方面：**应采取技术措施提高生成内容**格式框架**的**合理**性以及**有效内容**的**含量，**提高生成内容对使用者的**帮助**作用**。

正式文件改变了征求意见稿中对生成内容的安全性的绝对要求，转变为要求服务提供者采取操作保障内容安全，即要求服务提供者采取技术措施实现生成内容准确性和可靠性的提高。本条的增删在一定程度上体现了监管部门对于生成式人工智能具有不可控性这一客观事实的理解，进而将监管的侧重点从单一的“结果安全”转变为兼顾“结果安全”与“程序安全”，降低了各类生成式人工智能服务提供者的合规压力。

（五）第 7 条（安全措施要求）（对服务提供者的要求）

1. 第 7 条 a）（模型适用人群、场合、用途方面）

征求意见稿	正式文件
第 7 条 a）模型适用人群、场合、用途方面： 1）应充分论证在服务范围内各领域应用生成式人工智能的必要性、适用性以及安全性； 2）服务用于关键信息基础设施、自动控制、医疗信息服务、心理咨询等重要场合的，应具备与风险程度以及场景相适应的保护措施； 3）服务适用未成年人的，应： ——允许监护人设定未成年人防沉迷措施~~，并通过密码保护~~； ~~——限制未成年人单日对话次数与时长，若超过使用次数或时长需输入管理密码；~~ ~~——需经过监护人确认后未成年人方可进行消费；~~ ~~——为未成年人过滤少儿不宜内容，~~展示有益身心健康的内容。 4）服务不适用未成年人的，应采取技术或管理措施防止未成年人使用。	第 7 条 a）模型适用人群、场合、用途方面： 1）应充分论证在服务范围内各领域应用生成式人工智能的必要性、适用性以及安全性； 2）服务用于关键信息基础设施，**以及如**自动控制、医疗信息服务、心理咨询、**金融信息服务**等重要场合的，应具备与风险程度以及场景相适应的保护措施； 3）服务适用未成年人的： ——应允许监护人设定未成年人防沉迷措施； ——**不应向未成年人提供与其民事行为能力不符的付费服务；** ——**应积极**展示有益**未成年人**身心健康的内容。 4）服务不适用未成年人的，应采取技术或管理措施防止未成年人使用。

一方面，正式文件增加了金融信息服务作为重要场合之一，进行明确列举，要求服务提供者将服务用于金融信息服务时，也需要具备与风险程度以及场景相适应的保护措施。金融安全是国家安全的重要组成部分，本条的修改照应了《金融信息服务管理规定》中对金融信息服务提供者应当履行主体责任、建立信息安全保障等服务规范的要求[14]，体现了我国监管部门对于生成式人工智能服务应用于金融行业所可能构成的潜在风险的特别关注。

另一方面，本条也照应了《AIGC 暂行办法》对未成年人保护的要求，并对实现该等要求所应采取的措施进行了细化[15]。正式文件在适用未成年人服务方面，删除了密码管理形式及单日对话次数与时长的限制，避免过度限制未成年人自由，体现《中华人民共和国未成年人保护法》保护未成年人隐私权的原则[16]。

此外，正式文件在规范未成年人付费服务方面，由要求未成年人的监护人确认付费服务内容改为要求服务提供者负责不向未成年人提供与其民事行为能力不符的付费服务，并要求服务提供者谨慎开放面向未成年人的付费服务。

2. 第 7 条 b）（服务透明度方面）

征求意见稿	正式文件
第 6 条 c）服务透明度方面： 1）以交互界面提供服务的，应在网站首页等显著位置向社会公开~~以下~~信息： ——服务适用的人群、场合、用途等信息； ——~~第三方~~基础模型使用情况。 2）以交互界面提供服务的，应在网站首页、服务协议等便于查看的位置向使用者公开以下信息： ——服务的局限性； ——所使用的~~模型架构、训练框架等有助于使用者了解服务机制机理~~的概要信息。 3）以可编程接口形式提供服务的，应在说明文档中公开 1）和 2）中的信息。	第 **7** 条 b）服务透明度方面： 1）以交互界面提供服务的，应在网站首页等显著位置向社会公开服务适用的人群、场合、用途等信息，**宜同时公开**基础模型使用情况； 2）以交互界面提供服务的，应在网站首页、服务协议等便于查看的位置向使用者公开以下信息： ——服务的局限性； ——所使用的模型、**算法等方面**的概要信息； ——**所采集的个人信息及其在服务中的用途**。 3）以可编程接口形式提供服务的，应在说明文档中公开 1）和 2）中的信息。

[14] 《金融信息服务管理规定》第五条规定，“金融信息服务提供者应当履行主体责任，配备与服务规模相适应的管理人员，建立信息内容审核、信息数据保存、信息安全保障、个人信息保护、知识产权保护等服务规范。”

[15] 《AIGC 暂行办法》第十条规定，“提供者应当……指导使用者科学理性认识和依法使用生成式人工智能技术，采取有效措施防范未成年人用户过度依赖或者沉迷生成式人工智能服务。”

[16] 《中华人民共和国未成年人保护法》第四条规定，“保护未成年人，应当坚持最有利于未成年人的原则。处理涉及未成年人事项，应当符合下列要求：……（三）保护未成年人隐私权和个人信息；……。”

本条照应了《AIGC 暂行办法》第十条针对生成式人工智能服务透明度的要求[17]。正式文件在公开信息方面取消强制公开基础模型使用情况，此改动平衡了提升生成式人工智能服务对用户的透明度与保护商业秘密这两项需求。另外，正式文件中还增加了要求服务提供者在显著位置公开采集个人信息及其在服务中的用途的规定，与《个保法》中个人信息处理规则保持一致[18]，体现对个人信息保护的重视。

3. 第 7 条 c）（收集使用者输入信息用于训练方面）

征求意见稿	正式文件
第 7 条 c）收集使用者输入信息用于训练方面： 1）~~应事前与使用者约定能否将使用者输入信息用于训练~~； 2）应~~设置关闭~~使用者输入信息用于训练的~~选项~~； 3）使用者从服务主界面开始到达该选项所需操作不应超过 4 次点击； 4）应将收集使用者输入的状态，以及（2）中的关闭方式显著告知使用者。	第 7 条 c）收集使用者输入信息用于训练方面： 1）应**为**使用者**提供关闭其**输入信息用于训练的**方式，例如为使用者提供选项或语音控制指令；关闭方式应便捷，例如采用选项方式时**使用者从服务主界面开始到达该选项所需操作不超过 4 次点击； 2）应将收集使用者输入的状态，以及（1）中的关闭方式显著告知使用者。

本条照应了《AIGC 暂行办法》第九条的规定[19]。正式文件中删除了服务提供者“应事前与使用者约定能否将使用者输入信息用于训练”这一要求，而仅保留了“服务提供者显著告知 + 允许使用者便捷关闭”的同意机制。针对收集输入信息用于训练事项，建议服务提供者在与使用者的用户协议中明确规定、向使用者发送站内信或通过其他方式提示使用者其输入信息将用于训练，并根据正式文件的要求对向使用者明确提供简便快捷的关闭输入信息用于训练的方式。

[17] 《AIGC 暂行办法》第十条规定，“提供者应当明确并公开其服务的适用人群、场合、用途”。
[18] 《中华人民共和国个人信息保护法》第十七条规定，“个人信息处理者在处理个人信息前，应当以显著方式、清晰易懂的语言真实、准确、完整地向个人告知下列事项：……（二）个人信息的处理目的、处理方式，处理的个人信息种类、保存期限；……。个人信息处理者通过制定个人信息处理规则的方式告知第一款规定事项的，处理规则应当公开，并且便于查阅和保存。”
[19] 《AIGC 暂行办法》第九条规定，“提供者应当与注册其服务的生成式人工智能服务使用者（以下称使用者）签订服务协议，明确双方权利义务。”

4. 第 7 条 d）（图片、视频等内容标识方面）

征求意见稿	正式文件
第 7 条 d）图片、视频等内容标识方面，应~~按 TC260-PG-20233A《网络安全标准实践指南—生成式人工智能服务内容标识方法》进行以下标识：~~ ~~1）显示区域标识；~~ ~~2）图片、视频的提示文字标识；~~ ~~3）图片、视频、音频的隐藏水印标识；~~ ~~4）文件元数据标识；~~ ~~5）特殊服务场景的标识。~~	第 7 条 d）图片、视频等内容标识方面，应**满足国家相关规定以及国家标准要求**。

本条照应了《AIGC 暂行办法》中第十二条关于生成内容标识的规定[20]。正式文件中删去了内容标识的特定依据，我们理解，服务提供者对于生成内容标识的规定应当遵循包括《互联网信息服务深度合成管理规定》、TC260-PG-20233A《网络安全标准实践指南—生成式人工智能服务内容标识方法》在内的现有以及未来可能出现的相关国家标准、行业标准的规定。

5. 第 7 条 e）（训练、推理所采用的计算系统方面）

征求意见稿	正式文件
无该条款。	**第 7 条 e）训练、推理所采用的计算系统方面：** **1）应评估系统所采用芯片、软件、工具、算力等方面的供应链安全，侧重评估供应持续性、稳定性等方面；** **2）所采用芯片宜支持基于硬件的安全启动、可信启动流程及安全性验证，保障生成式人工智能系统运行在安全可信环境中。**

本条照应了《AIGC 暂行办法》第六条的规定[21]。正式文件中增加了对训练和推理所采用的计算系统方面的安全要求，提出计算系统供应链安全评估要求，并对芯片安全程度提出支持标准，从软件硬件两方面出发保障计算系统的安全运行，也与《中华人民共和国计算机信息系统安全保护条例》中计算机信息系统的使用单位应当建立健全安全管理制度的要求相适应[22]。

[20] 《AIGC 暂行办法》第十二条规定，“提供者应当按照《互联网信息服务深度合成管理规定》对图片、视频等生成内容进行标识。”
[21] 《AIGC 暂行办法》第六条规定，“促进算力资源协同共享，提升算力资源利用效能。推动公共数据分类分级有序开放，扩展高质量的公共训练数据资源。鼓励采用安全可信的芯片、软件、工具、算力和数据资源。”
[22] 《中华人民共和国计算机信息系统安全保护条例》第十三条规定，“计算机信息系统的使用单位应当建立健全安全管理制度，负责本单位计算机信息系统的安全保护工作。”

6. 第 7 条 g）（向使用者提供服务方面）

征求意见稿	正式文件
第 7 条 f）向使用者提供服务方面： 1）对明显偏激以及明显诱导生成违法不良信息的问题，应拒绝回答；对其他问题，应均能正常回答； 2）应设置监看人员，及时根据国家政策以及第三方投诉情况提高生成内容质量，监看人员数量应与服务规模相匹配。	第 7 条 g）向使用者提供服务方面： 1）**应采取关键词、分类模型等方式对使用者输入信息进行检测，使用者连续三次或一天内累计五次输入违法不良信息或明显诱导生成违法不良信息的，应依法依约采取暂停提供服务等处置措施；** 2）对明显偏激以及明显诱导生成违法不良信息的问题，应拒绝回答；对其他问题，应均能正常回答； 3）应设置监看人员，并及时根据**监看情况**提高生成内容质量**及安全**，监看人员数量应与服务规模相匹配。 注：**监看人员的职责包括及时跟踪国家政策、收集分析第三方投诉情况等**。

本条照应了《AIGC 暂行办法》第十四条中对服务提供者就违法内容采取处置措施的要求[23]。正式文件增加了对使用者输入信息的检测及相应处置要求，加强对使用者输入信息的监管，并要求服务提供者根据监看情况而非仅根据国家政策以及第三方投诉情况来进行内容优化。总体而言，本条向服务提供者提供了生成内容监控制度的设立要点，分别包括（1）在输入阶段，检测输入信息、对多次输入违法信息的用户采取处置措施；（2）在内容生成阶段，对诱导性问题设立拒绝回答的机制；（3）设置人员监控制度，根据监控情况及时调整基础模型，提高生成内容质量及安全。

7. 第 7 条 h）（模型更新、升级方面）

征求意见稿	正式文件
第 7 条 g) 模型更新、升级方面: 1）应制定在模型更新、升级时的安全管理策略； 2）应形成管理机制，在模型重要更新、升级后，再次进行安全评估，并按规定向主管部门重新备案。	第 7 条 h）模型更新、升级方面： 1）应制定在模型更新、升级时的安全管理策略； 2）应形成管理机制，在模型重要更新、升级后，再次**自行组织**安全评估。

正式文件删除了在模型重要更新、升级后须重新备案的要求，修改为由服务提供者自行组织安全评估。这一修改与《具有舆论属性或社会动员能力的互联网信息服务安全评估

[23]《AIGC 暂行办法》第十四条规定，“提供者发现违法内容的，应当及时采取停止生成、停止传输、消除等处置措施，采取模型优化训练等措施进行整改，并向有关主管部门报告。提供者发现使用者利用生成式人工智能服务从事违法活动的，应当依法依约采取警示、限制功能、暂停或者终止向其提供服务等处置措施，保存有关记录，并向有关主管部门报告。”

规定》（“《互联网信息服务安全评估规定》”）中互联网信息服务提供者应当在使用新技术新应用导致重大变更等情形下自行开展安全评估的规定相适应[24]。根据《互联网信息服务安全评估规定》第七条规定[25]，如该等模型重要更新、升级属于使用新技术新应用导致重大变更情形，服务提供者完成该等安全评估后需要将该等安全评估报告提交至网信部门等主管部门。如服务提供者涉及提供互联网新闻信息服务的，还需根据《互联网新闻信息服务新技术新应用安全评估管理规定》规定[26]报请国家或者省、自治区、直辖市互联网信息办公室组织开展安全评估。

8. 第 7 条 i）（服务稳定、持续方面）

征求意见稿	正式文件
无该条款。	**第 7 条 i）服务稳定、持续方面：** **1）应将训练环境与推理环境隔离，避免数据泄露和不当访问；** **2）应对模型输入内容持续监测，防范恶意输入攻击，例如 DDoS、XSS、注入攻击等；** **3）应定期对所使用的开发框架、代码等进行安全审计，关注开源框架安全及漏洞相关问题，识别和修复潜在的安全漏洞；** **4）应建立数据、模型、框架、工具等的备份机制以及恢复策略，重点确保业务连续性。**

本条照应了《AIGC 暂行办法》第十三条对生成式人工智能服务稳定性与持续性的要求[27]。正式文件单独增加了本条款，既遵循了《AIGC 暂行办法》的要求，也与目前全球各主要国家和地区对人工智能服务鲁棒性（robustness）的普遍关注态度一致。正式文件中明确了保障稳定性与可持续性须注意的技术要点，包括隔离训练环境与推理环境、持续监测模型输入内容、定期安全审计、建立备份机制和恢复策略等。

---

[24] 《互联网信息服务安全评估规定》第三条规定，“互联网信息服务提供者具有下列情形之一的，应当依照本规定自行开展安全评估，并对评估结果负责：……（二）使用新技术新应用，使信息服务的功能属性、技术实现方式、基础资源配置等发生重大变更，导致舆论属性或者社会动员能力发生重大变化的；……。”

[25] 《互联网信息服务安全评估规定》第七条规定，“互联网信息服务提供者应当将安全评估报告通过全国互联网安全管理服务平台提交所在地地市级以上网信部门和公安机关。具有本规定第三条第一项、第二项情形的，互联网信息服务提供者应当在信息服务、新技术新应用上线或者功能增设前提交安全评估报告；具有本规定第三条第三、四、五项情形的，应当自相关情形发生之日起 30 个工作日内提交安全评估报告。”

[26] 《互联网新闻信息服务新技术新应用安全评估管理规定》第七条规定，“有下列情形之一的，互联网新闻信息服务提供者应当自行组织开展新技术新应用安全评估，编制书面安全评估报告，并对评估结果负责：（一）应用新技术、调整增设具有新闻舆论属性或社会动员能力的应用功能的；（二）新技术、新应用功能在用户规模、功能属性、技术实现方式、基础资源配置等方面的改变导致新闻舆论属性或社会动员能力发生重大变化的。国家互联网信息办公室适时发布新技术新应用安全评估目录，供互联网新闻信息服务提供者自行组织开展安全评估参考。”

《互联网新闻信息服务新技术新应用安全评估管理规定》第八条规定，“互联网新闻信息服务提供者按照本规定第七条自行组织开展新技术新应用安全评估，发现存在安全风险的，应当及时整改，直至消除相关安全风险。按照本规定第七条规定自行组织开展安全评估的，应当在应用新技术、调整增设应用功能前完成评估。”

《互联网新闻信息服务新技术新应用安全评估管理规定》第九条规定，“互联网新闻信息服务提供者按照本规定第八条自行组织开展新技术新应用安全评估后，应当自安全评估完成之日起 10 个工作日内报请国家或者省、自治区、直辖市互联网信息办公室组织开展安全评估。”

27 《AIGC 暂行办法》第十三条规定，“提供者应当在其服务过程中，提供安全、稳定、持续的服务，保障用户正常使用。”

（六） 第 9 条（安全评估要求，原第 8 条）

1. 第 9.1 条（评估方法，原 8.1 条）

征求意见稿	正式文件
第 ~~8.1~~ 条（评估方法） ~~对提供者的~~要求如下。 a）~~应在服务上线前以及重大变更时开展~~安全评估，评估~~可~~自行开展~~安全评估~~，也可委托第三方评估机构开展。 b）安全评估应覆盖本文件所有条款，每个条款应形成单独的评估结论，评估结论应为符合、不符合或不适用： 1）结论为符合的，应具有充分的证明材料； 2）结论为不符合的，应说明不符合的原因，采用与本文件不一致的技术或管理措施，但能达到同样安全效果的，应详细说明并提供措施有效性的证明； 3）结论为不适用的，应说明不适用理由。 c）应将本文件各条款的评估结论以及相关证明、支撑材料写入评估报告： 1）评估报告应符合~~开展评估时~~主管部门要求； 2）撰写评估报告过程中，因报告格式原因，本文件中部分条款的评估结论和相关情况无法写入评估报告正文的，应统一写入附件。 ~~d~~）自行开展安全评估的，评估报告应至少具有三名负责人共同签字： 1）单位法人； 2）整体负责安全评估工作的负责人，应为单位主要管理者或网络安全负责人； 3）安全评估工作中合法性评估部分的负责人，应为单位主要管理者或法务负责人。	第 9.1 条（评估方法） 要求如下。 a）**按照本文件自行组织的**安全评估，**可由提供方**自行开展，也可委托第三方评估机构开展。 b）安全评估应覆盖本文件**第 5 章至第 8 章**中所有条款，每个条款应形成单独的评估**结果**，评估结果应为符合、不符合或不适用： 1）结**果**为符合的，应具有充分的证明材料； 2）结**果**为不符合的，应说明不符合的原因，**有以下特殊情况的应补充说明**： **采用与本文件不一致的技术或管理措施，但能达到同样安全效果的，应详细说明并提供措施有效性的证明；** 已采取技术或管理措施但尚未满足要求的，应详细说明采取的措施和后续满足要求的计划。 3）结**果**为不适用的，应说明不适用理由。 c）应将本文件**第 5 章至第 8 章**中各条款的评估结**果**以及相关证明、支撑材料写入评估报告： 1）评估报告应符合**履行备案手续时的相关**要求； 2）撰写评估报告过程中，因报告格式原因，本文件中部分条款的评估结**果**和相关情况无法写入评估报告正文的，应统一写入附件。 **d）应在评估报告中形成整体评估结论：** **1）各条款的评估结果均为符合或不适用时，整体评估结论为全部符合要求；** **2）部分条款评估结果为不符合时，整体评估结论为部分符合要求；** **3）全部条款均为不符合时，整体评估结论为全部不符合要求；** **4）第 5 章至第 8 章中推荐性条款的评估结果不影响整体评估结论。** **e**）自行开展安全评估的，评估报告应至少具有三名负责人共同签字： 1）单位法**定代表**人； 2）整体负责安全评估工作的负责人，应为单位主要管理者或网络安全负责人； 3）安全评估工作中合法性评估部分的负责人，应为单位主要管理者或法务负责人。

正式文件中对安全评估结果为不符合的特殊情况做出了进一步规范，要求服务提供者对已经采取技术或管理措施但未符合《AIGC 安全要求》项下强制性安全措施要求的部分进行详细说明，阐明采取的措施以及后续满足要求的计划。另外，正式文件中还新增了要求安全评估报告形成整体评估结论的内容，并明确了该结论的评估标准，进一步提高了安全评估报告的完整性与科学性。

二、生成式人工智能服务安全制度的修订要点与合规制度建议

（一）修订要点

整体而言，正式文件较征求意见稿的修订大致分为三个方面：其一，将各类定义与条款表述与《个保法》、《AIGC 暂行办法》等法律规定拉齐，保证规范概念体系的一致性；其二，从可行性的角度对部分生成式人工智能服务的安全要求进行了删繁就简、灵活处理，平衡了包括用户知情权、内容安全在内的合规需求与发展需求；其三，在语料采集、语料标注、内容安全监测、服务稳定性等方面为服务提供者提供了更为明确、详细、与当前发展水平相适应的制度设立的指引。

（二）安全合规制度建议

在归纳梳理了《AIGC 安全要求》正式文件中对服务提供者的各项服务安全评估要求后，我们建议服务提供者可考虑采取下述制度，以保障自身所提供的生成式人工智能服务的安全性。

1. 设立语料来源安全管理制度

（1）设立语料采集前及采集后安全评估制度，结合《AIGC 安全要求》附录 A 中针对语料安全风险信息的分类，对风险语料设立分级分类识别、关键词识别等识别机制。

（2）设立语料来源追溯制度。将采集的语料根据《AIGC 安全要求》分为自采语料[28]与商业语料，并针对不同类型的语料，通过内部培训、协议条款约定、交易审核流程等环节建立起语料追溯制度。

2. 设立语料内容安全管理制度，包括：

（1）语料内容过滤制度。服务提供者可以通过多种机制确保语料内容的合法性和适宜性，包括但不限于：关键词过滤、分类模型、人工抽检等；

（2）知识产权管理制度。服务提供者可以确立针对生成式人工智能的知识产权管理制度，包括但不限于：设立知识产权负责人和管理策略、设立知识产权风险识别制度、建

[28] 《AIGC 安全要求》正式文件规定，对于汇聚了网络地址、数据链接等能够指向或生成其他数据的情况，如果需要使用这些被指向或生成的内容作为语料，应将其视同于自采语料。

立用户易于访问的投诉举报渠道以及知识产权侵权问题报告渠道等。

3. 建立语料标注规范，包括：

（1）标注人员管理制度，包括但不限于：安全培训制度、考核制度、职能划分等；

（2）标注规则，包括但不限于：

- 确立规则基本内容，制定包含标注目标、数据格式、方法和质量指标的标注规则；
- 确立功能性与安全性标注规则，分别对功能性和安全性标注制定规则，至少覆盖数据标注和审核环节。

（3）标注内容抽检审核制度，包括但不限于：

- 功能性标注抽检制度，对每批标注语料进行人工抽检，不准确的内容需重新标注，包含违法不良信息的批次应废弃；
- 安全性标注审核制度，确保每条安全性标注语料至少需经一名审核人员审核通过。

（4）设立数据隔离存储制度，对安全性标注数据进行隔离存储，以保障数据的安全性和完整性。

4. 设立模型生成内容安全制度，包括但不限于：

（1）安全性评价规则，在模型训练过程中，将生成内容的安全性作为评估生成结果好坏的主要指标；

（2）输入信息安全性检测制度，对用户输入进行安全性检测，引导模型生成积极正向的内容；

（3）常态化监测与优化机制，建立常态化的监测评估机制，对服务过程中检测到的安全问题进行及时处理，并通过指令微调或强化学习等技术手段优化模型。

5. 其他综合管理制度

（1）模型适用性和安全性制度，包括但不限于：

- 特殊场合的保护措施：为关键信息基础设施领域，或者如自动控制、医疗信息服务等重要场景提供服务时，结合《网络安全法》《数据安全法》《关键信息基础设施安全保护条例》《信息安全技术 关键信息基础设施安全保护要求》等规范制定与风险相适应的保护措施；

- 未成年人保护制度：结合《未成年人网络保护条例》等未成年人保护相关法律法规设立未成年人使用规则，允许监护人设定防沉迷措施，展示有益内容，防止未成年人接触不适用服务。

（2）内容监管与质量控制制度，设置关键词和分类模型监管制度以及人员监看制度：对用户输入进行监管，对违法不良信息采取处置措施，并根据监看情况提高内容质量及安全，监看人员数量应与服务规模匹配。

（3）环境隔离与安全审计制度，隔离训练与推理环境，定期进行安全审计。设立持续监测与备份恢复制度，持续监测输入内容，建立数据和模型的备份及恢复策略。

结语

《生成式人工智能服务安全基本要求》正式文件的出台，不仅为企业提供了更明确的合规指导，也体现了我国在促进技术创新的同时，对于保障网络安全、个人隐私以及社会公共利益的坚定立场。面对快速发展的生成式人工智能技术，建立和完善相应的合规制度，不仅符合当前的法规要求，更使其在不断变化的法律环境中保持灵活性和前瞻性，确保技术创新的同时，能够有效管理风险，保护用户权益。我们将持续关注，为企业保驾护航。

感谢实习生缪逸泓、张文溢、何一辰对本文作出的贡献。

《生成式人工智能服务安全基本要求》实务解析

宁宣凤 吴涵 吴舸 张浣然 刘畅

引言

自 2022 年初以来，我国陆续发布算法推荐、深度合成与生成式人工智能服务相关的规范文件，初步构建起对特定领域人工智能技术与服务的监管机制。具体至生成式人工智能服务领域，在《生成式人工智能服务管理暂行办法》（“《暂行办法》”）的监管框架下，形成了由算法备案制度和生成式人工智能（大语言模型）备案（“大模型备案”）构成的“双备案制”的实践机制。

算法备案制度最早在《互联网信息服务算法推荐管理规定》（“《算法推荐规定》”）中确立，企业可通过中央网信办的互联网信息服务算法备案系统提交算法备案申请，流程和备案内容细则均较为固定。而大模型备案自《暂行办法》施行之日起也仅有半年，还需要与服务提供者开展更多有效地沟通、互动来积攒监管经验以制定明确、具体的规则，从而指引企业履行大模型备案义务，尤其是备案所需的安全评估。

在此背景下，2023 年 10 月 11 日，全国网络安全标准化技术委员会秘书处发布《生成式人工智能服务安全基本要求（征求意见稿）》（“《征求意见稿》”），就包含语料安全、模型安全在内的生成式人工智能服务安全的基本要求广泛征求社会公众意见。2024 年 3 月 1 日，历时近半年，《生成式人工智能服务安全基本要求》（“《基本要求》”）正式发布。根据规范内容，我们理解，《基本要求》对《暂行办法》相关合规要求例如数据来源合法、内容安全等在执行规则方面的细化，并对生成式人工智能服务提供者在实践中开展安全评估提供有效的路径，不仅能推动企业提高其自身的生成式人工智能服务安全能力，还可为监管部门评价特定生成式人工智能服务的安全水平提供参考标准。

基于前述，本文尝试明晰《基本要求》的出台背景与实践定位，梳理《基本要求》所涉的各类安全要求，以便为相关企业遵循执行《基本要求》提供抓手。

一、规范背景与定位

（一）系对《暂行办法》的细化支撑，对生成式人工智能服务其他适用法律法规的增强衔接

从规范效力来看，《基本要求》属于全国信息安全标准化技术委员会编制的技术文件，是一种旨在引导、指引生成式人工智能服务安全发展的指南类文件，而不具备强制性法律效力。但若逐一比对《暂行办法》除安全评估相关要求之外的通用规定（即第 5-7 章），可以看到《基本要求》并非是空中楼阁地架设额外合规义务，而是对《暂行办法》接近于一一对应的细化、解释，以及对于《暂行办法》上位法、其他监管生成式人工智能服务的法律法规的增强衔接性规定，故可以为服务提供者有的放矢落实《暂行办法》，在现行网络空间治理法律框架下合法合规提供生成式人工智能服务提供实践指引与监管侧重参照。

《基本要求》与《暂行办法》规定的具体对应关系，可参见下表：

《基本要求》规定			《暂行办法》规定
语料安全	语料来源安全	• **语料来源管理**：不得使用含违法不良信息超过 5% 的语料 • **语料搭配**：从语言、模态、境内外来源方面提出多样性要求 • **来源可追溯**：针对开源、自采、商业语料及使用者输入信息等不同来源语料提出可追溯要求 • **国家要求阻断的信息不应作为语料**	第四条第（二）款 第七条第（一）（四）（五）款 第九条
	语料内容安全	• **采取内容过滤措施过滤违法不良信息** • **采取知识产权保护措施**：设置知识产权负责人并建立管理策略、识别语料知产侵权风险、建立投诉举报渠道等 • **采取个人信息保护措施**：取得合法性基础	第四条第（三）（四）款 第七条第（二）（三）（五）款 第九条
	语料标注安全	• **标注人员规则** • **标注规则** • **标注内容准确性要求** • **隔离存储安全性标注数据**	第八条

《基本要求》规定		《暂行办法》规定
模型安全	• **应使用第三方已备案模型提供服务** • **模型生成内容安全要求**：将内容安全嵌入训练目的、采取模型输入信息检测和常态化检测机制 • **生成内容准确性** • **生成内容可靠性**	第四条第（五）款 第九条 第十七条
安全措施	• **模型适用人群、场合、用途要求：针对向关键信息基础设施、未成年人提供服务的特殊要求** • **服务透明度要求** • **使用者输入信息用于训练要求** • **内容标识义务** • **训练、推理所采用的计算系统要求** • **接受公众或使用者投诉举报** • **向使用者提供服务** • **模型更新、升级** • **服务稳定、持续**	基本涵盖《暂行规定》第三章“服务规范”要求的安全措施，故此处对具体条款不予一一列举。

值得注意的是，除《暂行规定》及其上位法外，考虑到《基本要求》列明的参考文献还特别包括了《中华人民共和国密码法》《商用密码管理条例》以及《网络信息内容生态治理规定》等生成式人工智能服务通常受到规制的法律法规，故从《基本要求》的规定中同样可以看到对前述规范的增强衔接性规定。例如，安全措施要求之“训练、推理所采用的计算系统要求”明确提出“对系统所采用芯片宜支持基于硬件的安全启动、可信启动流程及安全性验证”，即建议企业采用可信计算芯片，并应当注意遵循密码法、商用密码相关规定。

（二）大模型备案的配套指南

另一方面，根据《基本要求》总则，除说明其旨在支撑《暂行办法》外，“服务提供者在按照有关要求履行备案手续时，按照本文件第 9 章要求进行安全评估并提交评估报告。”

结合我们的备案相关项目经验，《基本要求》所指称的备案手续即是大模型备案，从实践中大模型备案的实践情况来看，《基本要求》实质上属于大模型备案的配套指引，其第 9 章“安全评估要求”对备案所需安全评估应涵盖的要点进行逐一细化，第 8 章“其他要求”及附录 A 则是对于安全评估材料必备附件的细化要求。

总体而言，我们理解《基本要求》是《暂行办法》等规定的有益细化补充，尽管暂时不具备强制法律效力，但被法规、规章等正式法源引用或其实际内容被作为监管执法参照时，其效力也会发生转化。考虑到《基本要求》属于结合大模型备案支持工作经验形成，文本成熟度较高，不排除网信部门在未来的大模型备案与生成式人工智能行政执法活动中将其作为参照性标准，这也是我国 APP 治理等网络空间治理领域的常见实践。

二、重点合规观察

相比 2023 年 10 月份发布的《征求意见稿》的内容，《基本要求》对生成式人工智能服务在各项安全方面的要求，提出了进一步细化的指引，同时对《征求意见稿》的部分内容进行了删除。如下列明了《基本要求》提出的重点合规要点：

（一）对关键术语作出明确定义

作为与《暂行办法》中“生成式人工智能服务提供者”定义的衔接，并为了明确《基本要求》的适用对象，《基本要求》所确定的“服务提供者”为“以交互界面、可编程接口等形式提供生成式人工智能服务的组织或个人。”此前在《征求意见稿》中的前述定义还包括了“面向我国境内公众提供服务”的前提限定，《基本要求》对此范围限定进行了删除，这一修订与《暂行办法》的整体适用范围保持一致。

其次，《基本要求》针对“训练语料”“抽样合格率”“基础模型”“违法不良信息”等实践中可能存在争议的关键术语均进行界定，例如“训练语料”包括所有直接作为模型训练输入的数据，而无论某一训练阶段，包括在预训练或优化训练过程中的输入数据。抽样合格率需要结合《基本要求》附录 A 列明的 31 种安全风险样本进行占比计算。

可以想见，明确上述关键术语定义，也将有助于生成式人工智能服务提供者有效落实合规义务，并使得人工智能服务供应链上的多元主体（例如训练语料提供者、基础模型开发者等）在合作过程中进一步界分各方责任义务。

（二）明确五类安全风险

鉴于生成式人工智能本身可能无法充分理解输入与输出内容的真实内涵，且受制于模型训练数据集等技术局限性，生成式人工智能的输出内容可能存在虚假、低劣、含有偏见与歧视性质，甚至输出与人类伦理准则不相符的内容。在广泛应用下，前述违法不良信息内容更有可能对于公众的事实认知造成影响并进一步引导舆论传播。因此，对于生成式人工智能输出的内容安全治理一直是立法与监管的“安全底线”。

《基本要求》在附录 A 中列明了语料及生成内容的五类主要安全风险，其中 A.1 类（包

含违反社会主义核心价值观的内容）以及 A.2 类（包含歧视性内容）作为关键词库必须覆盖的安全风险类别，属于五类安全风险中风险等级较高的情况，这也与《网络信息内容生态治理规定》中关于违法信息种类的列举意图一脉相承。

另外三类安全风险包括 A.3 类商业违法违规风险、A.4 类侵犯他人合法权益的风险以及 A.5 类无法满足特定服务类型的安全需求。随着数字经济的飞速发展，在新经济、新业态、新模式发展中逐渐出现了利用数据与技术手段从事不正当竞争的行为，在本次《基本要求》列明的 A.3 类商业违法违规风险项下，选择纳入了“利用算法、数据、平台等优势，实施垄断和不正当竞争行为”的安全风险，与《中华人民共和国反不正当竞争法（修订草案征求意见稿）》的立法方向进行了衔接。

但是，《基本要求》附录 A 列明的五类安全风险中也存在定义模糊、在实践中可能较难理解与界定的内容，例如 A.3 类商业违法违规风险项下的“违反商业道德”风险，A.4 类侵犯他人合法权益风险项下的“危害他人身心健康”风险，以及 A.5 类无法满足特定服务类型的安全需求风险项下的“内容不可靠”“无法对使用者形成帮助”等概念。

（三）合规义务清单

《基本要求》整体从语料安全、模型安全以及生成式人工智能服务的安全措施、词库题库等维度上对服务提供者提出了一系列较为细致的合规义务，其中语料安全保障义务又分为语料的来源安全、内容安全以及标注安全要求保障义务。同时，《基本要求》明确列明服务提供者需要根据《暂行办法》开展的安全评估应当针对《基本要求》中第 5 章至第 8 章的每一条款形成单独的评估结果，也即为服务提供者列明了合规义务清单，该等内容主要包括：

1. 语料安全要求

首先，就语料来源安全而言，《基本要求》首先删除了《征求意见稿》中关于建立语料黑名单机制的要求，也即“单一来源语料内容中含违法不良信息超过 5% 的，应将该来源加入黑名单”，但这不代表《基本要求》对于语料来源包含违法不良信息的比例没有进行规定；相反，《基本要求》对于语料来源的管理要求新增提出了事前评估、事后核验的双重安全保障措施，即对面向特定语料来源采集前需要进行安全评估，同时采集后也需要进行二次核验，以便完全排除掉“含有违法不良信息超过 5% 的情况”，从源头上全面避免了不良语料进入数据训练的问题。

其次，延续《算法推荐规定》《暂行办法》等规定对于算法偏见、算法歧视等要求，

《基本要求》提出需要就语料语言及语料类型进行不同语料的多语料来源的搭配，以提高语料来源的多样性，并且可以合理搭配境内外来源的语料。

第三，随着人工智能在公众社会中的普及，其生成内容可能会被广泛传播、引用和使用，而当出现违法不良信息输出或输出内容侵犯权益的情况时，直接传播者可能并非单一责任主体，违法不良信息或侵权内容可能存在于语料本身。因此，语料的可溯源性一直是保障生成式人工智能输出内容合法、安全的必要措施，也是定位输出内容责任主体、压实信息内容安全治理责任的有效办法。《基本要求》针对使用开源语料、自采语料、商业语料三种不同情形提出了细化规定，尤其是当服务提供者使用商业语料时，除了确保语料交易的法律效力、交易方或提供方对语料安全的承诺保障外，《基本要求》明确提出服务提供者同时应当对交易方或合作方所提供的语料、承诺、材料进行合规审查。这对于依赖第三方语料库的服务提供者而言为一项新增合规义务，但在服务提供者自行审核语料安全性时，应当以何种方式或者审核结果达到何种效果时方可确认某一语料的安全性，目前尚不非常明确。服务提供者可以考虑从交易方或合作方提供的基本书面材料有效性，结合《基本要求》附录列明的语料及生成内容的主要安全风险清单等方面进行多方面的审核。

2. 语料内容安全要求

《基本要求》对于语料内容的安全要求，主要围绕搭建以知识产权保护为基础的策略与结构，包括应当专门设置语料以及生成内容的知识产权负责人，允许并支持第三方就语料使用情况以及相关知识产权情况进行查询的功能设置等规定。结合此前北京互联网法院审结的人工智能生成图片著作权纠纷案、广州互联网法院审结的生成式人工智能训练数据集侵权案，可以看出《基本要求》对于在语料内容的训练、使用以及在事后为知识产权相关权益方提供畅通的投诉与查询通道等方面，同样继承了目前的实践监管趋势，表明了重点保护知识产权的立场与态度。

第二，秉承《网络信息内容生态治理规定》的主旨，《基本要求》在内容安全治理方面提出服务提供者应当采取关键词、分类模型、人工抽检等方式，充分过滤全部语料中的违法不良信息，从源头避免违法不良信息与内容的生成。

第三，针对语料中包含个人信息及敏感个人信息的情形，《基本要求》对服务提供者仅规定了事前告知与获得个人信息主体同意的前置义务，但并没有明确个人信息主体在事后可以通过何种方式向服务提供者主张权利、提出问询或投诉，或要求服务提供者删除其个人信息。服务提供者可以结合《个人信息保护法》中关于个人对个人信息的处理知情权、决定权、有权限制或拒绝他人对其个人信息进行处理的规定，在其服务产品的显著位置向

个人信息主体说明其有权主张前述权利的途径与方式，以便权利人的权益在使用生成式人工智能服务事前、事中与事后均可得到充分的保障。

3. 语料标注安全要求

《基本要求》针对语料标注人员、标注规则、标注内容的准确性均提出了要求，例如，相对《征求意见稿》而言，新增要求服务提供者应当自行组织对标注人员的安全培训，这与目前大模型备案中的安全评估要点也进行了衔接，同时明确了培训的内容应当包括例如标注任务规则、标注内容质量核验方法等事项。在标注内容的准确性保障方面，《基本要求》对于功能性标注与安全性标注提出了不同的确认规则，针对功能性标注需要进行人工抽检、抽检对象的语料以每一批为单位。而针对安全性标注则需要进行人工审核，审核对象的语料以每一条为单位。

4. 模型安全要求

《基本要求》首先明确如服务提供者需要基于第三方基础模型提供服务，必须使用已经主管部门备案的基础模型。尽管“基于”的范畴尚不明确，结合《暂行规定》及实践中的服务上线要求，我们理解，这意味着直接接入境内外未备案的服务提供者可能无法上线生成式人工智能服务，但并未禁止基于未备案基础模型进行二次开发的服务在完成备案要求后上线。此外，《基本要求》分别针对模型生成内容的安全性、准确性、可靠性提出了具体要求，以正确引导生成式人工智能服务可以安全地为使用者输出安全、有效、符合科学常识以及主流认知的内容。

5. 安全措施要求

在上线范围方面，《基本要求》提出了根据服务应用场景不同采取分类分级的保护措施的要求，针对用于关键信息基础设施，以及如自动控制、医疗信息服务、心理咨询、金融场景等本身存相对较高风险的内容安全、数据安全、个人信息保护安全的情况，服务提供者需要就此配备与具体场景风险程度以及场景相适应的保护措施，加强对重点重要场景的安全保障。

在针对未成年人使用者的保护方面，对于面向未成年人提供生成式人工服务的提供者，《基本要求》删除了《征求意见稿》中关于限制未成年人单日对话次数与时长的前提限定，对于可能涉及消费、支付功能或场景的情况，《基本要求》同样做出了微调，将“需经过监护人确认后未成年人方可进行消费”修订为“不应向未成年人提供与其民事行为能力不符的付费服务”，这一调整与《民法典》《未成年人网络保护条例》等规定中关于未

成年人从事与其年龄、认知等民事行为能力相符的行为以及引发的责任承担的规定意图相一致，对于通过生成式人工智能提供的各项付费服务，尤其是有益于未成年人身心健康的服务和内容，不应简单地要求所有未成年人均需事前获得其监护人的确认，从使用者的角度而言，前述修订也更易于例如辅助学习类别、寓教于乐类别的生成式人工智能服务得到更加广泛地应用。

第三，《基本要求》对生成式人工智能服务的使用者及其使用规范也提出了具体要求，例如，当需要收集使用者输入的信息用于数据和模型训练时，《基本要求》要求的授权模式为“默认开启 + 显著告知单独关闭渠道”，即“Opt-out”模式。也即服务提供者只需要为使用者提供关闭其输入信息用于训练的途径与方式即可，并非由使用者主动选择开启该等功能，这对于需要实时过滤、微调、更新模型训练语料与词库题库的需求而言可能为一项更高效便捷的模式，同时也能促使生成式人工智能服务的输出内容尽可能多地采集到不同使用者的输入内容以及同一内容的不同表述样本，有益于对不同问题的需求与反馈统计等，以便及时完善模型、内容质量及其配套安全措施。

需要注意的是，虽然前述“Opt-out”模式为目前同类生成式人工智能服务的主流实践做法，但该等模式下的个人信息处理的合法性基础可能较为薄弱，结合此前 X/Twitter、Zoom 等国内外企业更新隐私政策宣布将用户数据用于训练人工智能模型存在广泛争议的情况，以及基于同意将用户数据用于模型训练后响应删除权等个人数据权利存在技术困难等实践情况，我们理解目前亦有诸多企业将前述情形采取或调整为用户 Opt-in 模式，即默认情况下不会使用用户提交的数据来训练或优化模型，除非用户主动选择共享给公司；或明确不将该等数据用于模型训练的做法。

此外，《暂行办法》第十四条提出，提供者发现使用者利用生成式人工智能服务从事违法活动的，应当依法依约采取警示、限制功能、暂停或者终止向其提供服务等处置措施，保存有关记录，并向有关主管部门报告。《基本要求》承接并量化了前述主体责任的落实方案，明确规定了对违规使用或恶意使用生成式人工智能服务的情况应当设置使用拦截功能，如果使用者连续三次或一天内累计五次输入违法不良信息或明显诱导生成违法不良信息的，应依法依约采取暂停提供服务等处置措施，这为提供者在实践中设置有关拦截机制方面提供了相对具体的合规义务落实指引。

值得注意的是，《基本要求》删除了《征求意见稿》规定的在模型重要更新、升级之后，需要再次进行安全评估并重新向主管部门备案的要求。根据目前的规定，服务提供者在模型完成重要更新和升级后，只需再次自行组织安全评估即可，这对于服务提供者的合

规成本以及提供服务的业务连续性均是利好信号。但是，前述“重要更新和升级”所指具体内容与范围、以及该事项是否会落入需要办理变更备案的情况，尚待进一步释明。

6. 关键词库与测试题库要求

建立关键词库与测试题库是保证语料安全的重要前提之一，《基本要求》结合其附录 A 列明的五类安全风险，对于关键词库与生成内容测试题库进行了对应规定，可以看出针对风险等级较高的 A.1 类与 A.2 类安全风险均规定了最低条目数量的要求。

同时，相比《征求意见稿》，《基本要求》新增了关键词库与测试题库的定期更新的推荐性条款，服务提供者可以参照该等规定，按照网络安全、信息内容安全治理以及业务实践情况与需要，及时更新关键词库与测试题库。

三、合规建议

（一）增强语料等源头治理工作

《暂行办法》第九条压实了生成式人工智能服务提供者的网络信息内容生产者责任，即落实生成内容的具体责任主体。但是，生成式人工智能的权利侵害结果形成原因可能复杂多样，仅关注传统的侵害结果维度容易忽略了人工智能技术的设计意图、训练数据、科技伦理等源头性因素对人工智能造成侵害结果的直接影响。因此，相关企业应增强人工智能的源头治理力度，将人工智能侵害治理追溯至人工智能的设计、开发等源头阶段，优化人工智能侵害的责任主体链。

具体而言，在算法设计方面，应采取有效措施防止产生民族、信仰、国别等歧视，并将伦理道德纳入技术体系，推动“科技向善”。在人工智能训练、优化、测试数据等语料治理方面，应关注语料的来源、质量、安全乃至语料搭配等方面，制定恰当的治理规则以实现语料的源头治理。例如，保留开源语料的开源许可协议、自采语料的采集记录等以确保语料来源的可追溯性，对采集后的语料进行核验以避免因使用存在大量违法不良信息的语料而致使人工智能的运行受到相应的负面影响。

（二）坚持安全底线，尤其是内容安全

《基本要求》旨在细化生成式人工智能服务在安全方面的基本要求，从而贯彻落实“发展与安全并重”的人工智能理念。这一以安全为底线护航人工智能产业发展的思路也应贯彻至生成式人工智能服务提供者等相关企业的业务实践当中。

一方面，企业宜遵循《基本要求》的颗粒度，逐步落实语料来源安全、语料内容安全、语料标注安全、模型安全与安全措施的相关要求，完善内部的安全治理体系，保障用户权

益、社会公共利益。另一方面，《基本要求》还重点对包含语料内容安全和生成内容安全的内容安全提出了关切，例如通过附表形式明确列举了现阶段识别出的主要内容安全风险，细化了内容安全目标粒度或提出具有实操性的安全举措。我们理解，在网络消息传播速度飞速的互联网信息时代，对生成式人工智能服务相关的内容安全的重视也是互联网内容治理整体框架下的“题中应有之义”。

总的来说，我们理解，企业在积极推进生成式人工智能安全治理工作时，宜重点加强内容安全治理，采取诸如人工与机器结合审核、引入或开发高质量语料库等方式保障输入与输出内容的安全。实践中，我国各行业积极建设高质量语料库，如人民网开发主流价值语料库、智源人工智能研究院联合其他单位发布开源可信中文互联网语料库。相关企业可充分利用现有优质语料资料，实现内容安全与技术发展的“同频共振”。

（三）加强合规体系建构，完备内部制度

伴随《暂行办法》《科技伦理审查办法（试行）》《网络安全标准实践指南——生成式人工智能服务内容标识方法》《基本要求》等与生成式人工智能技术及服务相关的法律法规、技术文件的出台，构建制度完备的生成式人工智能合规体系是相关企业在人工智能迅捷发展时代下的重要命题。

我们理解，完备的人工智能合规体系应以网络安全、数据安全和个人信息保护为底层架构，即应首先按照《网络安全法》《数据安全法》《个人信息保护法》及配套规则构建健全的网络安全与数据合规治理体系，例如制定并落实数据分级分类保护规则、安全事件应急响应规则。在此基础上，结合《暂行办法》等人工智能相关规定，补充完善训练数据治理、内容治理、生成内容标识、科技伦理审查等由人工智能引发的特殊合规要求。

此外，根据 2023 年度国务院立法计划，我国《人工智能法草案》已被纳入提请全国人大常委会审议的范畴当中。我们理解，随着人工智能监管顶层设计的出台、包含生成式人工智能在内的人工智能治理规范体系的不断完善，人工智能相关企业应依据最新立法与监管动态，及时更新既有合规体系。为了确保企业内部既有人工智能合规体系能够有效落地与执行，企业还宜积极引入具备人工智能技术与法律双重背景的治理人才，以畅通新兴技术发展与企业合规之间的桥梁。

（四）开展大模型备案遵照执行《基本要求》

如前所述，《基本要求》作为全国网络安全标准化技术委员会发布的技术文件，并非具有强制性法律效力的正式立法，也暂未被纳入强制性国家标准的范畴。但是，一方面，

不具有法定强制性效力不影响《基本要求》因其规范支撑作用与明晰的执行颗粒度，而可能在实践监管活动中被相关监管部门选定为评估生成式人工智能服务安全能力等方面的具有可落地性的执法标准。

另一方面，考虑到《基本要求》基本涵盖了大模型备案过程中要求生成式人工智能服务提供者开展安全评估的具体评估指标例如训练语料来源、标注人员情况等，相关服务者在实际开展大模型备案工作时，也宜遵照《基本要求》的相关合规要求自行或委托第三方机构开展安全评估，以提高备案效率。

（五）开展大模型备案前组织自测工作

自大模型备案开展以来，大模型备案的实践要求处于动态调整的过程当中，例如，在保留训练数据来源等内容的基础上，新增了企业提供大模型下载通道、实测账号与内容审核机制等要求。我们理解，这在一定程度上体现出监管部门对大模型备案的重视程度：随着实践经验的积累，适时调整备案要求从而契合当下生成式人工智能服务的发展情况，实现精准收集信息、提供更为柔性的指导之目的。

因此，在正式开展大模型备案前，为提高备案的可行性，如可行，相关企业宜自行或引入上海人工智能实验室等第三方专业机构开展大模型自测工作，遵循《基本要求》的安全评估办法及具体评估事项开展测试工作，以“查漏补缺”，优化自身安全评估流程，提高自身的生成式人工智能服务的安全水平。

AI 法律实务

AI-Generated Content and Copyright (China)

宋海燕 赵怡冰 王默

A Practice Note providing an overview of the developing governance of AI-generated content (AIGC) and the emerging copyright-related issues associated with AIGC under the laws of China (PRC). The Note considers the subsistence of copyright in AIGC, copyright ownership, and potential copyright infringement issues arising when either training a generative AI model or generating AIGC. It also introduces a few notable Chinese copyright cases involving AIGC, including the Beijing Film Law Firm v BD case, the Shenzhen Tencent v. Shanghai Yingxun case, the LI v. LIU case, and the SCLA v. AI Company case.

The meaning and scope of AI or generative AI (GAI) vary between fields and jurisdictions. The OECD's Council on Artificial Intelligence adopted a recommendation including a definition of "AI system" (Recommendation of the Council on Artificial Intelligence OECD/LEGAL/0449, adopted 22 May 2019, amended on 8 November 2023).

This Note provides an overview of the developing governance of AI-generated content (AIGC) and the emerging copyright- related issues associated with AIGC under the laws of China (PRC). The Note considers the subsistence of copyright in AIGC, copyright ownership, and potential copyright infringement issues arising when either training a GAI model or generating AIGC. It also introduces a few notable Chinese copyright cases involving AIGC, including the Beijing Film Law Firm v. BD case, the Shenzhen Tencent v. Shanghai Yingxun case, the LI v. LIU case, and the SCLA v. AI Company case (the Ultraman case).

I. AIGC and Common AIGC Applications

AIGC refers to the content, including texts, pictures, audios and videos that are generated using advanced generative AI (GAI) techniques (Article 22, Interim Measures for the Management of Generative Artificial Intelligence Services 2023 (2023 GAI Measures)).

China has witnessed the rapid developments of AIGC applications, from the first and most well-known text-generating AIGC application (that is, ChatGPT launched by OpenAI in November 2023), to the picture-generating AIGC applications (such as MidJourney and Stable Diffusion), and the audio and video generating AIGC applications (such as Imagen Video).

In China, the following AIGC applications are active in the market:

- Text-generating: ERNIE Bot developed by Baidu.
- Picture-generating: Miaoya developed by Reshuffle technology.
- Audio and video generating: Zenvideo developed by Tencent.

II. Main Legislation

Main Chinese legislation relates to AIGC and the associated copyright-related issues include:

- The Copyright Law of the PRC 2020 (2020 Copyright Law, with effect from 1 June 2021), issued by the Standing Committee of China's National People's Congress. As the third revision of Chinese Copyright Law, the revised law includes detailed provisions on:
  - the composition of a copyrightable work;
  - the ownership of a copyrightable work;
  - limitations and exceptions to exclusive rights (that is, fair use); and
  - infringement liability rules.
- The Regulations for the Implementation of Copyright Law of the PRC 2013 (2013 Copyright Implementing Regulations), issued by the State Council. The implementing regulations supplement the 2020 Copyright Law by providing practical guidelines and operational details. As the third revision, the regulations clarify:
  - the scope of copyrightable works;
  - the rights of copyright holders;
  - procedures for handling copyright infringement cases; and
  - penalties for copyright violations.
- The 2023 GAI Measures, issued by the Cyberspace Administration of China, in conjunction with the National Development and Reform Commission and five other state agencies. The measures are China's first GAI regulations that regulate GAI developers and address a few important topics surrounding GAI services, including:
  - the definition of AIGC providers, that is, organisations and individuals that provide GAI services using GAI technology, including through the provision of programmable interfaces (Article 22(2));
  - the principle of GAI regulation in China (that is, under classified and graded supervision) (Article 3); and
  - the responsibilities of AIGC providers (such as undertaking security evaluation and reporting) (Articles 7-15).
- The Regulations on the Protection of Right of Dissemination via Information Network 2013 (2013 Information Network Regulations, first issued in 2006 and revised in 2013), issued by the State

Council. The regulations are the Chinese version of the US Digital Millennium Copyright Act (DMCA) (see Practice Note, Digital Millennium Copyright Act (DMCA): Safe Harbors for Online Service Providers), providing safe harbour provisions for internet service providers (ISP).

- The Beijing High People's Court Guidelines for the Trial of Copyright Infringement Cases 2018 (2018 Beijing High Court Guidelines). Although China is not a case-law jurisdiction where precedents made by higher courts are legally binding, yet the guidelines issued from this high-profile court may be influential to other lower courts in China (see Practice Note, Case Guidance System in China). The guidelines clarify the judicial interpretations of the Beijing High People's Court on copyright-related infringement cases and reflects the court's opinion on:
  - what constitutes a copyrightable work;
  - what constitutes originality;
  - how to determine the ownership; and
  - what constitutes copyright infringement.

III. Copyrightability for AIGC Outputs

(I) Copyrightability: General Test

Under Chinese copyright law, a work that meets the following conditions may be protected by copyright:

- It is original.
- It can be represented in a certain format.
- It belongs to intellectual creations of human beings.

The 2020 Copyright Law stipulates the definition of works, which refers to ingenious intellectual achievements that are original and can be presented in a certain form in the fields of literature, art and science (Article 3). (For a non-exhaustive list of recognised types of works of authorship, see Practice Note, Copyright (China): Overview: Works Protected by Copyright.)

The 2013 Copyright Implementing Regulations also affirm the requirements of originality and human authorship for a copyrightable work. The regulations specify that the term "works" as referred to in the 2020 Copyright Law means intellectual creations with originality in the literary, artistic and scientific domain, insofar as they can be reproduced in a tangible form (Article 2).

The 2018 Beijing High Court Guidelines further clarify that works must meet the requirement of being created by human authors and the criteria of originality.

When examining whether the object of copyright claimed by the plaintiff constitutes a work, the court should generally consider the following factors:

- Whether it is a creation made by a natural person in the fields of literature, art and science.

• Whether it possesses originality.

• Whether it has a certain form of expression.

• Whether it is reproducible.

(Article 2.1, 2018 Beijing High Court Guidelines.)

In determining the originality of an expression, the court should generally consider the following factors:

• Whether the expression is independently created by the author.

• Whether the arrangement of expressions manifests the author's selection and judgment.

(Article 2.2, 2018 Beijing High People's Court Guidelines.)

(II) Copyrightability: Application of General Test to AIGC

In view of the above legislation, an AIGC output may be subject to copyright protection under the context of the 2020 Copyright Law, provided that there is sufficient "human authorship" in the AIGC outputs. However, what constitutes "human authorship" is often debatable and subject to Chinese courts' discretion.

For instance, in Beijing Film Law Firm v. BD, the Beijing Internet Court deemed that the graphics and the initial report automatically generated by the legal analysis software Wolters Kluwer (WK) lacked human authorship and thus should not be subject to copyright protection (but finding that the final article published by the plaintiff was copyrightable because the final article was substantially different from the initial auto-generated report and reflected the author's personal judgement) (see Beijing Film Law Firm v. BD). Yet, in Shenzhen Tencent v. Shanghai Yingxun, the Shenzhen Nanshan District People's Court ruled that, the subject article produced or assisted by AI-writing software reflected the writer or editor's personal judgement and arrangement and hence should be subject to copyright protection in China (see Shenzhen Tencent v. Shanghai Yingxun).

It is also noteworthy that in LI v. LIU, China's first case concerning the copyrightability of AI-generated pictures, the Beijing Internet Court ruled that, since the plaintiff had "input over 150 prompts and experimented various parameters multiple times" when using Stable Diffusion to generate the AI picture, the plaintiff has exercised "aesthetic choices and personal judgment in the entire generation process," thus finding the AI-generated picture to be a copyrightable work (see LI v. LIU).

The Beijing Internet Court's decision in LI v. LIU seems to contradict with the US decisions on the copyrightability of AIGC outputs (for example, the Zarya of the Dawn case, the A Recent Entrance to Paradise case and the Théâtre D'opéra Spatial case), in which both the US Copyright Office and US courts denied copyright protection to AIGC outputs that lack human authorship. For instance, in Théâtre D'opéra Spatial, on the same issue of "prompts" input by humans to generate AI pictures,

the Review Board of the US Copyright Office concluded that even over 600 prompts input by the applicant were not enough to constitute human authorship.

However, it is worth noting that the different outcomes made by the courts of these two jurisdictions are not because that Chinese courts believe non-humans could be authors, or that Chinese copyright law does not require "human authorship," but rather that the two courts or jurisdictions seem to have different standards and interpretations as to what constitutes human authorship.

IV. Ownership of Copyright in AIGC Outputs

(I) Copyright Ownership: Default Rules Under Chinese Law

With regard to the ownership of copyrightable works, the default rule under the 2020 Copyright Law is that a work vests its copyright ownership in its creators (Article 11), with a few exceptions such as employment work (Article 18), commissioned work (Article 19), computer software, and audio-visual work (Article 17).

The copyright owner of a work may either be a natural person (that is, ownership belongs to the author who creates the work) or a legal entity (that is, ownership belongs to the employer or commissioner). Article 9 of the 2020 Copyright Law provides:

"Copyright owners include authors, and other natural persons, legal entities and unincorporated organisations that enjoy copyright in accordance with this law."

The default rule regarding copyright ownership is that a work vests its ownership with its author, who creates the work. Article 11 of the 2020 Copyright Law provides:

"Except otherwise provided in the law, the copyright in a work belongs to its author, and that the author of a work is the natural person who created the work."

In terms of legal persons' work, Article 11 of the 2020 Copyright Law provides:

"Where a work is created according to the intention and under the supervision and responsibility of a legal entity or unincorporated organisation, the legal entity or unincorporated organisation is the author of the work."

In terms of employment work, Article 18 of the 2020 Copyright Law provides:

"A work created by a natural person when fulfilling the tasks assigned by a legal entity or unincorporated organisation is deemed to be an employment work. Unless otherwise provided in Paragraph 2 of this Article, the copyright of such a work is enjoyed by the author, but the legal entity or unincorporated organisation has a priority right to exploit the work within the scope of its professional activities."

In terms of commissioned work, Article 19 of the 2020 Copyright Law provides:

"The ownership of copyright in a commissioned work can be agreed upon in a contract between

the commissioning party and the commissioned party. In the absence of such a contract or of an explicit agreement in the contract, the copyright in such a work belongs to the commissioned party."

In terms of audio-visual work, Article 17 of the 2020 Copyright Law provides:

"Among audiovisual works, the copyright of cinematographic works and television programs is enjoyed by producers, but the scriptwriter, director, photographer, composer and lyricist enjoy the right of authorship and the right to receive compensation from the producer.

The copyright ownership of audio-visual works other than those specified in the preceding paragraph will be agreed upon by the parties; where there is no agreement or the agreement is unclear, the copyright is enjoyed by the producer, but the authors have the right of authorship and the right to receive compensation.

The authors of script, music and other works that may be used separately from the audio-visual work have the right to separately exercise their right of copyright."

In terms of computer software, Article 9 of the Regulations for the Protection of Computer Software 2013 provides:

"Except where otherwise provided in these regulations, the copyright in the software belongs to its developer.

In the absence of proof to the contrary, the natural person, legal entity or other organisation whose name appears on the software will be its developer."

(II) Copyright Ownership: Application in AIGC Outputs

Where an AIGC output may be subject to copyright protection in China, the next question is to determine who owns the copyright to the AIGC output.

In addition to the framework for copyright ownership vesting under Chinese copyright law, the terms of use agreements posted by Chinese AIGC providers may also provide a glimpse into the industry practices (see Copyright Ownership: Default Rules Under Chinese Law).

There are usually three scenarios. A summary of the example terms is set out below:

- **AIGC outputs owned by AIGC users**. [Company A]: "Unless otherwise agreed or stipulated by the applicable laws and regulations, you (user) own the intellectual property right (IPR) to the content generated [by this AIGC tool] based on the content uploaded by you that is legally owned by or authorised to you. You also agree to authorise us ([Company A]) to use the content that you upload, generate, and synthesise for the purpose of providing this service." This looks a common approach adopted by many AIGC providers.
- **AIGC outputs owned by AIGC providers**. [Company B]: "All IPRs related to this platform, including but not limited to copyrights, trademarks, patents, trade secrets, and all information and content

generated by this service (including, but not limited to, texts, images, audios, videos, graphics, interface designs, layout frameworks, relevant data or electronic documents), are protected by laws and regulations, and belong to [Company B] and its affiliated entities, who have the complete legal rights, ownership and other lawful rights. Users only have the right and permission to use this service and its related content in accordance with the terms of this agreement."

- **Silent**. [Company C]: "[Company C] is the principal entity responsible for the development and operation of this Service. [Company C] owns all the rights, within the scope allowed by the applicable laws and regulations, to all data, information, outputs, and other products generated during the development and operation of this service, except for the rights that are expressly preserved by the other rights holders under the law."

In the case of LI v. LIU, the Beijing Internet Court held that the plaintiff who used an AI tool to generate AI pictures is the author of the subject picture, as the picture was generated as a result of the plaintiff's intellectual input and reflected the plaintiff's personalised expressions (see LI v. LIU). The court also indicated that:

- An AI service itself could not be considered as an author of a copyrightable work because an AI is not a human being.
- Neither the developers or providers of AI services could be considered as the author in this case because they had no intent to create the subject picture and they had not actually participated in the subject picture creation process.
- Based on the terms of use agreement posted on the website of the AI tool, the AI developers already waived their rights, if any, in the AIGC output, that is, they "do not claim rights to the output content."

V. Copyright Infringement Associated with AIGC

To identify copyright infringement risks associated with AIGC, it is important to understand the following:

- The exclusive rights of copyright holders under Chinese copyright law, and exceptions and limitations to these exclusive rights.
- What exclusive rights may be infringed during the AIGC creation process, and whether any exceptions and limitations may apply.

(I) Exclusive Rights of Copyright-Holders

A copyright holder enjoys the following exclusive rights:

- The right of publication, that is, the right to decide whether to make a work available to the public.
- The right of authorship, that is, the right to claim authorship and to have the author's name

mentioned in connection with the work.

- The right of alteration, that is, the right to alter or authorise others to alter one's work.
- The right of integrity, that is, the right to protect one's work against distortion and mutilation.
- The right of reproduction, that is, the right to produce one or more copies of the work by means of printing, photocopying, rubbing, sound recording, video recording, duplicating, re-shooting, or digitising, and so on.
- The right of distribution, that is, the right to provide the public with original copies or reproduced copies of works by means of selling or donating.
- The right of lease.
- The right of exhibition, that is, the right to publicly display the original copies or reproduced copies of works of fine art and cinematographic works.
- The right of performance, that is, the right to publicly perform works, and to publicly transmit the performance of works by various means.
- The right of projection.
- The right of broadcasting.
- The right of dissemination via information networks.
- The right of production, that is, the right to fix works on a carrier by audiovisual production.
- The right of adaptation, that is, the right to modify a work for the purpose of creating a new work of original creation.
- The right of translation, that is, the right to transform the language of a work into another language.
- The right of compilation, that is, the right to choose or edit some works or fragments of works so as to form a new work.
- Other rights which are enjoyed by the copyright owners.

(Article 10, 2020 Copyright Law.)

Under Article 10, the first four bullets relate to the personal rights of an author (which are not assignable), the fifth to 16th bullets relate to the economic rights of an author, and the last bullet is a catch-all provision that allows a bit of flexibility to include other rights that are not expressly listed under this provision.

(II) Fair Use Exception

The 2020 Copyright Law lists out the exceptions and limitations to the exclusive rights, where no permission or compensation from the copyright owner is required to use the copyrightable work, provided that the name or designation of the author and the title of the work are mentioned, and

that the use must not impact the normal use of the work or unreasonably harm the copyright holders' lawful rights and interests.

The exceptions and limitations include:

- Use of a published work for the purposes of the user's own private study, research, or self-entertainment.
- Appropriate quotation from a published work in one's own work for the purposes of introduction of, or comment on, a work, or demonstration of a point.
- Inevitable reappearance or citation of a published work in newspapers, periodicals, radio stations, television stations, or other media for the purpose of reporting news.
- Reprinting by newspapers or periodicals or other media, or rebroadcasting by radio stations or television stations or other media, of the current event articles on the issues of politics, economy and religion, which have been published by other newspapers, periodicals, radio stations or television stations or other media, except where the copyright owner has declared that publication or broadcasting is not permitted.
- Publication in newspapers or periodicals or other media, or broadcasting by radio stations or television stations or other media, of a speech delivered at a public assembly, except where the author has declared that publication or broadcasting is not permitted.
- Translation, adaptation, compilation, and broadcasting or reproduction, in a small quality of copies, of a published work for use by teachers or scientific researchers in classroom teaching or scientific research, provided that the translation or reproduction is not published or distributed.
- Use of a published work by a state organ within the reasonable scope for the purpose of fulfilling its official duties.
- Reproduction of a work in its collections by a library, archive, memorial hall, museum, art gallery, art museum or similar institution, for the purpose of the display or preservation of a copy of the work.
- Free of charge performance of a published work, that is, with respect to the performance, neither fees are charged from the public nor the remuneration is paid to the performers, nor the performance is for profit.
- Copying, drawing, photographing, or video recording of an artistic work located or on display in a public place.
- Translation of a work published by a Chinese citizen, legal entity or unincorporated organisation, which is created in the national common language and characters, into a minority nationality language for publication and distribution within the country.
- Providing published works for dyslexics in a barrier-free way through which they can perceive.

- Other circumstances prescribed by laws and administrative regulations.

(Article 24, 2020 Copyright Law.)

The first paragraph of Article 24 sets out the three-step test for the fair use doctrine under Chinese copyright law (similar to Article 9 of the Berne Convention):

- Fair use may only apply in certain special circumstances, that is, the 12 explicit exceptions under Article 24(1) to (12) and the catch-all provision under Article 24(13) for other permissible use circumstances that may be provided under other laws.
- Fair use must not conflict with a normal exploitation of the work.
- Fair use must not unreasonably prejudice the legitimate interests of the copyright owner. For example, the copyright owner's right of signature should be protected during fair use.

(III) IP Provisions Under Chinese AIGC Legislation

The 2023 GAI Measures are the first Chinese AI regulations, which address:

- The principle of GAI governance in China.
- The responsibilities of AIGC providers, including requiring AIGC providers to respect the IP of other rights holders.

In terms of the AIGC training process, although not explicitly stipulated, the 2023 GAI Measures require all materials used for AIGC training purposes to be legally acquired, thus requiring AIGC providers to get permission from copyright holders before they use the copyrighted materials for AIGC training purposes.

AIGC providers should carry out pre-training, optimised training, and other training-data processing activities according to the law, and comply with certain obligations, including:

- Using data and underlying models from legitimate sources.
- Where IPRs are involved, not infringing on the IPR lawfully enjoyed by others.

(Article 7, 2023 GAI Measures.)

In terms of AIGC outputs, AIGC providers are also required to manage the AIGC outputs and take down illegal content or suspend and terminate users' services, when necessary. Where a AIGC provider discovers illegal content in the AIGC outputs, it should promptly:

- Take disposal measures, such as stopping generation, transmission, and eliminating the illegal content.
- Make rectification through measures such as model-based optimisation training.
- Report to the relevant competent department.

(Article 14, 2023 GAI Measures.)

(IV) Copyright Infringement Risks in AIGC Training Process

Unlike the express text and data mining (**"TDM"**) exceptions provided under the UK or the EU legislation, China does not have specific TDM exceptions under its copyright law. Also, because Chinese courts tend to apply stringent interpretations of the fair use exception (including the three-step test) in copyright infringement cases, there are serious copyright infringement risks when AIGC providers use unauthorised copyrighted works in its AIGC training process. (For more information on the TDM exceptions under the UK or EU legislation, see Practice Notes, Copyright: permitted acts: Copies of text and data analysis for non-commercial research (TDM) and Digital Copyright Directive: key provisions: Text and data mining exceptions (Articles 2 to 4).)

In November 2023, a group of artists sued a popular Chinese lifestyle social media platform "Little Red Book" (also known as, Xiaohongshu), arguing that the defendant had used their copyrighted works without authorisation when training its AI-painting model. The set of cases are currently pending before the Beijing Internet Court.

(V) Copyright Infringement Risks in AIGC Outputs

Where an AIGC output is found to be substantially similar with a prior copyrightable work, the AIGC provider may be held liable for copyright infringement.

For instance, in a 2024 Chinese court ruling involving "Ultraman," the Guangzhou Internet Court held that the AI-generated picture "Ultraman" by the Chinese defendant was substantially similar with the pictures of Ultraman owned and registered by the Japanese company Tsuburaya (which grants an exclusive licence to the plaintiff in China valid until 31 March 2024), thus finding the defendant liable for violating the plaintiff's right of reproduction and the right to prepare derivative works (that is, the right of alteration) (see SCLA v. AI Company).

The Guangzhou Internet Court also ruled that the defendant failed to exercise reasonable care under Article 4 (respecting IPRs) of the 2023 GAI Measures when providing GAI services, thus should be liable for compensating the plaintiff's losses. Specifically, the court indicated that the defendant failed to:

- Provide a complaint and reporting mechanism for rights holders (Article 15, 2023 GAI Measures).
- Include appropriate provisions in its terms of use agreement to remind AIGC users to respect the IPRs of others when generating AIGC outputs (Article 4, 2023 GAI Measures).
- Label AIGC outputs to distinguish them from human outputs (Article 12, 2023 GAI Measures).

The defendant did not itself train the AIGC data to develop the text-picture AIGC generating tool, rather it interfaced with a third party AI technology vendor to provide the AIGC services. Yet the court quoted Article 22(2) of the 2023 GAI Measures (see Main Legislation), and held that AIGC providers, by definition, also include those who provide AIGC services through interfacing with other application programs, thus finding the defendant liable for violating both the 2023 GAI

Measures and the plaintiff's exclusive rights.

VI. Notable Chinese Cases

(I) Beijing Film Law Firm v. **BD**

In April 2019, the Beijing Internet Court held that the subject article published by the plaintiff had enough human authorship and thus should be subject to copyright protection. The defendant, BD, published the plaintiff's subject article without authorisation and thus was liable for copyright infringement.

In May 2020, the Beijing Intellectual Property Court affirmed the decision made by the Beijing Internet Court. (See Beijing Film Law Firm v. BD [2018] Beijing Internet Court (Jing 0491 Min Chu No.239).)

Key Facts

The plaintiff, Beijing Film Law Firm, claimed that it organised to write an article titled "Analytical Report on the Judicial Big Data in the Film and Entertainment Industry: Film Industry in Beijing" and first published the article on its official WeChat account on 9 September 2018, thus owning the copyright to the subject article. The subject article included certain graphics and texts that were generated by the legal analysis software WK.

On 10 September 2018, the defendant BD, China's largest internet search engine, published an almost identical article on its Baijiahao account without the plaintiff's authorisation, only deleting a few sections such as the signature, introduction and search overview. The plaintiff sued the defendant for copyright infringement and petitioned for damages in the amount of RMB10,000 and a reasonable expense of RMB560.

The defendant argued that plaintiff's subject article was not copyrightable because it was automatically generated by the WK software and thus should not be subject to copyright protection in China.

Legal Issues

The first-instance court identified the legal issues involved in the case as follows:

- Whether the plaintiff has the standing to sue, as the copyright owner of a valid copyrightable work.
- Whether the defendant has constituted copyright infringement.
- Whether any defences should apply.

When evaluating the first issue (that is, whether the plaintiff has the standing to sue), the court focused its analysis on whether the subject article published by the plaintiff is a copyrightable work. Here, the court made separate analysis and determinations on the copyrightability of the following three items:

- The graphics that were generated by the WK software and included in the subject article. The court held that these graphics are not copyrightable because they were automatically generated by the WK software and did not reflect the plaintiff's original expressions.
- The report that was generated by the WK software. The court held that such a report is not copyrightable because it was also generated by the WK software rather than by human beings.
- The subject article published by the plaintiff in its final format. The court compared the subject article with the "report" generated by the WK software and found them to be "very different." Thus, the court concluded that the subject article was an original work of authorship created by the plaintiff and should be subject to copyright protection.

In terms of the second and third issues, the court found that BD's unauthorised use of the plaintiff's subject article violated the plaintiff's exclusive rights to the subject article and thus is liable for copyright infringement.

(II) Shenzhen Tencent v. **Shanghai Yingxun**

On 24 December 2019, the Shenzhen Nanshan District People's Court ruled that the plaintiff, Tencent, has exercised enough human authorship when using the AI-writing software to generate the subject article, and thus is the copyright owner of the subject article. The defendant's unauthorised use of plaintiff's subject article constituted copyright infringement.

(See Shenzhen Tencent v. Shanghai Yingxun [2019] Shenzhen Nanshan District People's Court (Yue 0305 Min Chu No.14010).)

Key Facts

On 20 August 2018, the plaintiff, Tencent, published a financial-related article on the Tencent Securities website and noted in the end that "[t]his article was automatically written by Tencent's robot Dreamwriter." The Dreamwriter software is an AI- writing assistance software developed by an affiliate of Tencent. It includes various functions from data collection and analysis, to writing, smart proofreading and automatic distribution of its reports to relevant platforms such as Tencent for publication.

The defendant, Shanghai Yingxun, without the plaintiff's permission, reprinted the article on its website. The plaintiff sued the defendant on the grounds of copyright infringement and unfair competition.

Legal Issues

In this case, the court focused on three key issues as follows:

- Whether the subject article constitutes a copyrightable written work.
- If the subject article is copyrightable, whether the plaintiff possesses the copyright of the subject article.

• Whether the defendant has constituted copyright infringement.

With regard to the first issue, when determining the "originality" of a work, the court noted that the key is to determine whether the subject article was independently created and whether it exhibits a certain degree of differentiation from existing works (creativity).

When analysing the creation process of the subject article, the court found that the main stages of generating the subject article were decided and arranged by the plaintiff, including data input and processing, setting trigger conditions, selecting article framework templates and corpus styles. The court further noted that although no natural person participated in the two minutes when the Dreamwriter software was generating the subject article, the automatic operation of the Dreamwriter software was still based on the plaintiff's decisions. As such, the court concluded that these decisions and arrangements made by the plaintiff constitute intellectual activities or creations that are directly related to the subject article, and therefore, the subject article is a copyrightable work and the plaintiff is the copyright owner of the subject article.

In its final decision, the court ruled that the defendant's unauthorised publication of the subject article on the website constituted copyright infringement.

(III) LI v. LIU

On 27 November 2023, the Beijing Internet Court ruled that the plaintiff has exercised enough human authorship when using the AI tool to generate the AI picture, by inputting over 150 prompts and experimenting various tech parameters multiple times. Therefore, the court found the AI picture is a copyrightable work and that the defendant was liable for copyright infringement.

(See LI v. LIU [2023] Beijing Internet Court (Jing 0491 Min Chu No. 11279).)

Key Facts

On 24 February 2023, the plaintiff, Mr Li, used Stable Diffusion, a text-to-picture AI service, to generate a picture titled "Spring Breeze Brings Tenderness" and published it on a social media platform, Little Red Book (also known as Xiaohongshu).

The defendant, Ms Liu, a Chinese blogger, then used the subject picture as an illustration in her article, but removed the plaintiff's user ID and the watermark of Little Red Book from the picture. The plaintiff sued the defendant for copyright infringement, including violating his right of authorship and the right of dissemination via the internet.

The defendant argued that the picture she included in her article was obtained through online searches, and she could not confirm whether the plaintiff has rights over the subject picture. The defendant also argued that she had no commercial intent when using the subject picture.

Legal Issues

In this case, the court focused on three key issues as follows:

- Whether the subject picture constitutes a copyrightable work and therefore should be subject to Chinese copyright protection.
- If the subject picture is copyrightable, whether the plaintiff is the copyright owner of the subject picture.
- Whether the defendant has constituted copyright infringement.

With regard to the first issue (that is, whether the subject picture constitutes a copyrightable work), the court ruled that the subject picture is a work of fine art, and therefore subject to Chinese copyright protection. The Beijing court started its analysis by listing out the criteria for a work to be protected under Chinese copyright law, including whether the work:

- Falls within the fields of literature, art and science.
- Possesses originality.
- Has a specific form of expression.
- Is an intellectual creation.

The court held that because the subject picture is akin to commonly seen photographs and paintings, it has satisfied the first and third criteria.

Regarding the criteria of "intellectual creation," the court held that a copyrightable work should reflect the intellectual input of human beings. The court found that the plaintiff has provided intellectual inputs, including choosing the preferred AI service provider, inputting various prompts and setting and re-setting technical parameters to produce, choose and rearrange the pictures. As such, the court determined that the subject picture is an intellectual creation.

Regarding the criteria of "originality," the court held that a copyrightable work should be independently created and reflect the author's personalised expressions. Although the plaintiff did not physically draw the specific lines, the plaintiff designed elements of the picture and continuously adjusted the picture, which reflected the plaintiff's aesthetic choices and personal judgment in the entire generation process. Therefore, the court found that the subject picture possesses originality.

With regard to the second issue (that is, whether the plaintiff is the copyright owner of the subject picture), the court ruled out the possibility that an AI service itself could be considered as an author of a copyrightable work because an AI is not a human being. The court held that neither the developers or providers of AI services could be considered as the author in this case because they had no intent to create the subject picture and they had not actually participated in the subject picture creation process. Because the subject picture was generated as a result of the plaintiff's intellectual input and reflected the plaintiff's personalised expressions, the plaintiff is the author of the subject picture.

In the end, the court ruled that the defendant was liable for infringing the plaintiff's copyright in the subject picture.

(IV) SCLA v. AI Company

On 8 February 2024, the Guangzhou Internet Court issued the first-instance judgment on China's first case concerning the infringement of AIGC outputs, finding the defendant, a text-to-image AIGC provider, liable for infringing the copyright of the famous Ultraman IP. (See SCLA v. AI Company [2024] Guangzhou Internet Court (Yue 0192 Min Chu No.113).)

Key Facts

The Japanese company, Tsuburaya Productions Co., Ltd. (Tsuburaya), is the copyright owner of the famous cartoon IP Ultraman series. Tsuburaya granted to the plaintiff, Shanghai Character License Administrative Co., Ltd. (SCLA), an exclusive licence to the Ultraman series' works in China, including the right of reproduction, the right to prepare derivative works, and also the right to enforce. Tsuburaya has also registered its Ultraman series pictures at the Chinese Copyright Office.

The defendant (an AI company) provides text-picture AI generating services through its website Tab. In late December of 2023, the plaintiff found that by inputting prompts containing or related to "Ultraman," the defendant's website could generate identical or substantially similar pictures to its Ultraman series images. The plaintiff then sued the defendant for copyright infringement.

Legal Issues

In this case, the court focused on the following two issues:

- Whether the defendant infringed on the plaintiff's copyright, that is:
  - right of reproduction;
  - right to prepare derivative works; and
  - right of dissemination via the internet.
- What civil liabilities should the defendant bear if it constitutes copyright infringement?

On determination of copyright infringement, the Guangzhou Internet Court supported the plaintiff's first two infringement claims, and explained why the third claim for network dissemination right infringement was not addressed. Specially, the court specified that:

- The AI pictures generated by the AI tool were substantially similar to the original expressions of the prior copyrightable works "Ultraman." Therefore, the court found that the defendant infringed on the plaintiff's right of reproduction.
- The subject AI-generated pictures partially kept the original expressions of the "Ultraman Tiga Hybrid Image," but also formed new features of their own, which constituted unauthorised derivative works to the prior copyrightable Ultraman works. Therefore, the court found that the defendant infringed on the plaintiff's right to prepare derivative works.

- The court has already addressed the defendant's violation of the plaintiff's right of reproduction and right of preparing derivative works and ruled both in the plaintiff's favour. Therefore, the court would not address the claim on network dissemination right infringement involving the same infringing act.

In terms of liabilities, the court ruled that the defendant was responsible for ceasing the infringing act and paying compensation, and specifically analysed the following issues:

- **The definition of AIGC providers**. Under Article 22(2) of the 2023 GAI Measures, the term "AICG provider" also includes those who provide AIGC services through interfacing with other application programs. Therefore, the court ruled that the defendant was qualified as an AIGC provider, and rejected the argument that it did not develop the Tab AIGC model itself but relied on a third party vendor to generate the AIGC services by interfacing with the third party application program.
- **The specific measures that an AIGC provider should implement to avoid liability**. The court concluded that the defendant should take extra steps to the extent that its AIGC function no longer generates any pictures substantially similar to the prior copyrightable Ultraman works when its users input any Ultraman-related prompts.
- **The plaintiff's claim on training data.** The court rejected the plaintiff's request for the defendant to delete the copyrightable Ultraman works from its training data base because the defendant itself did not train the AIGC model.
- **AIGC providers' reasonable duty of care when generating AIGC outputs**. The court ruled that the defendant failed to exercise the duty of care as required by the 2023 GAI Measures in the following aspects, thus held that the defendant was subjectively at fault and should pay compensation of RMB10,000 for its infringement:
  - establishing a complaint and reporting mechanism for rights holders (Article 15);
  - including appropriate provisions in its terms of use agreement to remind AIGC users to respect the IPRs of others when generating AIGC outputs (Article 4); and
  - labelling AIGC outputs to distinguish them from human outputs (Article 12).

The Ultraman case is China's first case addressing the infringement issues associated with AIGC outputs, and it is also the first court decision interpreting China's first AI regulations (that is, the 2023 GAI Measures).

The court also emphasised the need "not to overburden AIGC providers" and that AIGC providers should take "proactive measures to fulfil reasonable and affordable duty of care," so as to leave space for the development of generative AI industry, which is still at its early stage. At the end of the decision, the court called for the creation of a Chinese-style AI governance system that is balanced, inclusive, and compatible with both innovation and protection.

AI 原生应用相关法律问题研究之一：AI+ 教育法律问题

唐丽子 孙及 王玉山 周凡尧

百度创始人、董事长兼 CEO 李彦宏在 2023 百度世界大会上指出“AI 应用原生时代即将来临。没有构建于基础模型之上的丰富 AI 原生应用，大模型就一文不值。”

大模型与人工智能应用的关系或可借鉴移动互联时代的生态圈，即智能手机和手机操作系统是底层基础，而形形色色的移动应用是核心元素及触达用户的终端。大模型正如移动互联时代的手机操作系统，未来可能仅有少数几家大模型作为底层操作系统，统一整个市场，但基于大模型的 AI 应用必将迎来百花齐放、百舸争流的大繁荣。

在搜索、文档、教育、游戏、医疗、金融、电商、专业服务、数据查询分析、营销、政务等领域，AI 应用已经展现了广泛且良好的应用前景，未来也将出现更多行业通过 AI 改进用户体验、提高效率甚至带来新一轮产业变革浪潮。

我国针对人工智能领域已经颁布《生成式人工智能服务管理暂行办法》《互联网信息服务算法推荐管理规定》《互联网信息服务深度合成管理规定》等法规，形成了初步的监管框架。2023 年 10 月 18 日，中央网信办又发布了《全球人工智能治理倡议》，围绕人工智能发展、安全、治理三方面系统阐述了人工智能治理的中国方案。但各行业的 AI 应用，除需遵守人工智能监管要求以及行业现有的监管及合规要求外，还需要兼顾基于人工智能及行业结合过程中产生的新的监管、合规问题。

一、教育 +AI 的发展趋势

截至目前，人工智能在教育行业的应用主要有三个层次：第一个层次是将一些通用且成熟的人工智能应用作为教育工具使用，或者嵌套在教育应用中。例如基于语音识别技术、图像识别技术而开发的口语测评、拍照搜题等教育产品；第二个层次是通过综合人工智能技术对教育应用进行赋能、提升既有教育应用的效率。例如基于云计算技术或大模型开发的 AI 虚拟助教、AI 互动课程、AI 作业批改等人工智能教育产品，可以减少人力工作，提高教学者和学习者的效率；第三个层次相对较为前沿，主要趋势体现为，不

仅仅通过软件或机器替代一些简单的人类工作，而是真正由人工智能主导教学、定义教学，形成完全不同于人类教师的教育方法和知识图谱，或者是对学习者的自适应学习提供智能化的支持。

应用在教育行业的人工智能技术主要包括语音识别、图像识别、计算机视觉、增强智能、智能机器人、自然语言处理、知识图谱、决策智能、数据标注服务等。

从应用场景角度，教育 +AI 产品主要可划分为教学、学习、考试、测评、管理等不同类型：

1. 教学应用场景：企业通过研发人工智能技术相关的智能软件和硬件，为教学提供便利，如通过大数据分析及可视化、学情分析和实时交互，并结合学生学习现状和学习表现对教学内容及活动进行个性化定制，例如智能教育机器人，为中小学及普通职业学生提供课程、竞赛、实践等学习机会。

2. 学习应用场景：通过智能学习软硬件等方式对学生学习情况进行精准分析，构建基础知识图谱，形成针对不同学生不同知识点的有效学习路径，并根据学生的能力和偏好为学生规划适合的学习路径并推荐学生相关的学习资源，例如思维学习机、家庭学习智慧屏、学习单词、阅读 APP 等。

3. 考试应用场景：通过 AI 技术自动批改、口语测评、考试分析（运用计算机视觉、数据挖掘、自然语言处理等技术汇总各群体考试结果，生成考试情况的相关报告并归结错因用于辅助老师的精准教学）、个性化组卷（通过数据挖掘等方式对已有题库的数据进行分析整理、组合协助老师为不同层级的学习者制定考卷），例如通过大数据提供算法的学情分析等服务、在线教育平台提供的智能陪练服务。

4. 能力测评场景：通过采集学生完整的成长数据多维度全面且综合地评价学生的发展状况并生成综合素质评价报告、课堂评测报告、师生匹配评测，用于提升课堂教学质量、为学生匹配合适教师等，例如对学生进行全方位的成长评控体系建设系统，主要聚焦在过程性评价以及课程育人、活动育人、空间育人、协同育人等多种育人模式。

5. 规划管理场景：包括智能升学生涯规划和智能职业生涯规划，通过大数据分析、可视化、智能评测、个性化推荐等方式，从学生能力学习偏好、自身学科水平、大学报考条件限制等维度帮助学生解决选课、选科、选校等生涯规划服务；基于学生差异化的特长和个性化特征评估学生职业兴趣，智能生成职业生涯规划计划，例如 AI 高考志愿填报服务平台。

二、教育 +AI 的监管框架

随着近年来互联网和 AI 技术的迅猛发展，教育行业被 AI 赋能，通过在线教育为载体向客户提供新的教育模式，并且以 AI 技术赋能的智能教育产品也相继涌现，逐渐打造智能教育新生态。在教育 +AI 的业态下，智能教育企业不仅要遵守在线教育行业的准入要求，同时也要符合互联网经营、AI 有关的监管政策，教育 +AI 面临着新的合规挑战。以下将分别从行业准入要求、主要资质证照、提供生成式人工智能服务相关的合规义务、AI 教育辅助工具的监管概要等方面进行分析。

（一）教育 +AI 行业准入要求

我国教育体系可以分为学历教育和非学历教育，学历教育为学生提供获得教育部认证的证书或学位，包括学前教育、义务教育（小学、初中）、高中教育及高等教育，非学历教育是学历教育的补充，包括校外辅导、外语培训、职业培训等。根据上述教育体系并结合在线教育行业的准入规则，按照接受教育服务的对象和教育服务的内容，在线教育主要划分为面向学龄前儿童的在线培训、面向中小学生的学科类在线培训和非学科类在线培训（语言能力、艺术、体育、科技、研学等），以及面向成人的非学历教育在线培训。

目前国家相关政策已明确不得开展面向学龄前儿童的线上培训，面向义务教育阶段和高中学生的学科类培训也受到严格的限制，曾涉及该等业务的上市公司也陆续停止小学、初中、高中阶段学科类培训业务，如 2021 年 10 月，某知名教育上市公司发布公告，停止经营中国内地义务教育阶段学科类校外培训服务。面向中小学生的非学科类在线培训，目前各地正全面规范该等培训行为，如规定培训内容和时间、加强收费管理，在满足设置标准并完成法人登记等准入流程后方可开展；面向成人的非学历教育在线培训目前尚未形成完整的监管规定，但需要根据实际从事的培训内容并结合当地政策的设置标准申请准入。具体如下：

1. 面向学龄前儿童的在线培训

面向学龄前儿童的在线培训业务目前已无法获得准入批准。根据《关于进一步减轻义务教育阶段学生作业负担和校外培训负担的意见》（“《双减政策》”），统筹做好面向 3 至 6 岁学龄前儿童的校外培训治理工作，不得开展面向学龄前儿童的线上培训。北京市于 2022 年 4 月印发了《关于进一步做好教育移动互联网应用程序备案及管理工作的通知》，明确不再受理学前线上培训教育移动应用备案申请，已备案的予以撤销。

2. 面向中小学生的学科类在线培训

面向中小学生的学科类在线培训业务目前已无法获得准入批准。根据《双减政策》，各地不再审批新的面向义务教育阶段学生的学科类校外培训机构，现有的学科类培训机构统一登记为非营利性机构，线上学科类培训机构由备案证改为审批制，未经许可不得以任何线上方式从事有偿性学科类培训；不再审批新的面向普通高中学生的学科类校外培训机构。对面向普通高中学生的学科类培训机构的管理，参照《双减政策》有关规定执行。

虽然《双减政策》的重心为小学、初中义务教育阶段，高中阶段学科类培训曾一度被部分教育培训行业人士视作为缓冲地带，但教育部发布的 2022 年工作要点中明确表示，要指导各地落实高中阶段学科类培训严格参照义务教育阶段执行的政策要求，实践中面向高中阶段学生的在线培训原则参照义务教育标准执行。

因此，虽然《双减政策》前取得审批并已登记为非营利性机构的可以继续面向中小学生提供学科类在线培训，如北京希望在线线上学科培训学校（希望学 APP—提供小学至高中课程及服务）、北京猿辅导线上学科培训学校（猿辅导 APP—提供小学至高中课程及服务）、北京途途向上线上学科培训学校（途途课堂 APP—提供小学至高中课程及服务），但由于不再审批新的面向义务教育阶段学生的学科类校外培训机构，如企业拟进入该业务领域存在障碍。

3. 面向中小学生的非学科类在线培训

面向中小学生的非学科类在线培训，目前各地正全面规范该等培训行为，在满足设置标准并完成法人登记等准入流程后方可开展。2022 年 11 月《教育部等十三部门关于规范面向中小学生的非学科类校外培训的意见》（“《非学科类校外培训的意见》”）实施，针对面向中小学生的非学科类校外培训进行了规范，要求各地要区分体育、文化艺术、科技等类别培训机构，明确相应主管部门。省级主管部门要结合本地实际，牵头制定相应类别线上和线下培训机构的基本设置标准。线上机构还应符合网络安全有关标准。非学科类线上培训机构须依法取得省级有关主管部门的行政许可后，再依法进行法人登记，并向所在地省级电信主管部门履行互联网信息服务核准手续。

上述《非学科类校外培训的意见》实施以来，各地针对非学科类校外培训陆续制定了相关的政策文件，如广东省于 2023 年 6 月发布了《广东省教育厅关于切实解决面向中小学生的非学科类校外培训机构审批工作系列问题的通知（二）》，明确了非学科类校外培训机构的审批依据、设置标准、申报材料并启动线上非学科类培训机构办学许可证的申报审批、申报指南等。

4. 面向成人的非学历教育在线培训

目前法律监管体系中没有针对面向成人的非学历教育的定义和明确的哪些业务应落入相关法律法规调整的范围，面向成人的非学历教育在线培训目前尚未形成完整的监管规定。相较于上述面向中小学生的在线培训而言，面向成人的非学历教育在线培训监管环境整体较为宽松。2021 年 7 月，教育部印发了《关于加强社会成人教育培训管理的通知》，对成人教育培训机构的名称使用、招生管理、培训内容进行了规范，提出鼓励采用“互联网 +”的混合学习模式，搭建网络学习平台和移动学习平台，加强资源建设，提升服务和管理水平，推进人工智能在教育培训和管理等方面的全流程应用，提高教育培训的便利度和实效性。

（二）教育 +AI 业务的主要资质证照要求

1. 办学许可证

面向中小学生的学科类在线培训、面向中小学生的非学科类在线培训原则上均应取得办学许可证。《中华人民共和国民办教育促进法》第六十五条规定：“本法所称的民办学校包括依法举办的其他民办教育机构”，《中华人民共和国民办教育促进法实施条例》第十六条规定：“利用互联网技术在线实施教育活动应当符合国家互联网管理有关法律、行政法规的规定。利用互联网技术在线实施教育活动的民办学校应当取得相应的办学许可。”各地区主管部门出台的相关通知（如《广东省教育厅关于切实解决面向中小学生的非学科类校外培训机构审批工作系列问题的通知（二）》），明确了非学科类校外培训机构办学许可证的申报、审批工作和审批依据。

面向成人的非学历教育在线培训，由于实践中培训包含的内容较多（如招录考试类、资格准入类、职业技能类、企业管理类、生活技能类、兴趣爱好类等）并且线上培训内容、培训形式和法律关系呈现多样化，建议根据不同的在线培训模式和培训内容并结合所在地的最新政策进行判断。

2. 增值电信业务经营许可证

教育 +AI 业务一般涉及向网络用户有偿提供信息服务（即经营性互联网信息服务），国家对经营性互联网信息服务实行许可制度，在线教育机构通常需要就此申请取得《增值电信业务经营许可证》。

根据《互联网信息服务管理办法》，从事经营性互联网信息服务（即通过互联网向上网用户有偿提供信息或者网页制作等服务活动）的主体应当向电信管理机构或者国务院信息产业主管部门申请办理互联网信息服务增值电信业务经营许可证；从事非经营性互联网

信息服务（即通过互联网向上网用户无偿提供具有公开性、共享性信息的服务活动），应向电信管理机构或者国务院信息产业主管部门办理备案手续。

3. 网络文化经营许可证

从事经营性互联网文化活动[1]的主体，应取得文化部门颁发的《网络文化经营许可证》。根据文化和旅游部办公厅于 2019 年 5 月颁发的《关于调整 < 网络文化经营许可证 > 审批范围进一步规范审批工作的通知》，“网络表演”指以网络表演者个人现场进行的文艺表演活动等为主要内容，通过互联网、移动通讯网、移动互联网等信息网络，实时传播或者以音视频形式上载传播而形成的互联网文化产品；教育类、培训类等直播不属于网络表演，不属于互联网文化产品，因而不需要取得网络文化经营许可证。其后，北京市文化和旅游局、天津市文化和旅游局于 2020 年 3 月相继发文，明确确认教育类直播不属于网络文艺表演活动，不需要申请办理《网络文化经营许可证》。

参考近期某线上教育公司招股书披露，该公司研发并提供直播或录播形式的在线课程，并辅以在线自学资料和工具，但并未披露其获得《网络文化经营许可证》。从前述规定及案例来看，通过互联网提供在线教育（比如远程视频培训等）自身并不构成互联网文化产品，原则上无需取得相关证照，但如果涉及特定文化类业务，或教育被 AI 赋能时属于经营性互联网文化活动，应取得《网络文化经营许可证》。

4. 信息网络传播视听节目许可证

根据《互联网视听节目服务管理规定》，在中国境内向公众提供互联网（含移动互联网）视听节目服务活动[2]应当取得广播电影电视主管部颁发的信息网络传播视听节目许可证。并且，申请从事互联网视听节目服务，必须为国有独资或国有控股单位，因此非国资的民营教育机构由于不符合申请条件，通常而言并不能取得该许可证。

实践中，从事线上教育的企业是否需要取得视听许可证并不完全明确。参考近期某线上教育公司的招股书披露，由于视听节目服务的定义较为含糊，尚不确定线上培训服务是否属于该定义的范畴，以及是否需要取得视听许可证。该公司提供直播或录播形式的在线培训课程及在线自学资料，但由于缺少视听许可证，于 2019 年 12 月被相关部门处以人民币 3,000 元罚款，后通过与相关广电部门访谈确认，不会因为未取得视听许可证被要求终止视听节目服务业务。

因此，为避免相关合规风险，建议结合教育 +AI 企业实际从事的业务，就是否应取得

---

[1] 经营性互联网文化活动是指以营利为目的，通过向上网用户收费或者以电子商务、广告、赞助等方式获取利益，提供互联网文化产品及其服务的活动。其中互联网文化产品是指通过互联网生产、传播和流通的文化产品，主要包括：(1) 专门为互联网而生产的网络音乐娱乐、网络游戏、网络演出剧（节）目、网络表演、网络艺术品、网络动漫等互联网文化产品；(2) 将音乐娱乐、游戏、演出剧（节）目、表演、艺术品、动漫等文化产品以一定的技术手段制作、复制到互联网上传播的互联网文化产品。

[2] 即制作、编辑、集成并通过互联网向公众提供视音频节目，以及为他人提供上传及传播视听节目服务的活动。

该证书与相关广电部门进行确认。

5. 网络出版服务许可证

在线教育公司将编辑加工后的数字化作品向客户提供是否属于前述规定中的网络出版服务，目前缺乏明确、统一的监管尺度。部分在线教育公司在招股书中披露，通过与主管部门进行访谈，确认无需取得网络出版服务许可证即可开展在线培训服务，包括通过线上平台提供音视频课程及材料。

鉴于《网络出版服务管理规定》本身较为模糊且针对在线教育没有进一步明确的规定，如公司向消费者提供相关培训视频、学习资料，建议就是否应需取得《网络出版服务许可证》咨询地方出版主管部门的意见。

（三）提供生成式人工智能服务相关的合规义务

教育 +AI 在遵守在线教育相关要求的基础上，也需要遵守并履行人工智能领域相关的合规义务，主要包括：（1）算法备案义务；（2）安全评估义务；（3）网络安全与数据合规义务等。同时，随着生成式人工智能技术的不断发展，应用于教育行业相关 AI 应用仍在不断地创新和发展，可能不断产生新的监管、合规挑战。为应对该等问题，除遵守目前规定的具体合规要求外，教育 +AI 发展过程中也需尊重和遵守人工智能领域相关伦理治理要求及科技伦理审查要求。上述合规义务及伦理治理要求具体如下：

1. 算法备案义务

如果教育 +AI 提供的算法推荐服务、生成式人工智能服务或深度合成服务具有舆论属性或者社会动员能力的，应履行算法备案手续，在提供服务之日起十个工作日内通过互联网信息服务算法备案系统填报服务提供者的名称、服务形式、应用领域、算法类型、算法自评估报告、拟公示内容等信息，履行备案手续。算法推荐服务提供者的备案信息发生变更的，应当在变更之日起十个工作日内办理变更手续。算法推荐服务提供者终止服务的，应当在终止服务之日起二十个工作日内办理注销备案手续，并作出妥善安排。

2. 安全评估义务

如果教育 +AI 提供的算法推荐服务、生成式人工智能服务或深度合成服务具有舆论属性或者社会动员能力的，还需根据《具有舆论属性或社会动员能力的互联网信息服务安全评估规定》开展安全评估，应当对信息服务和新技术新应用的合法性，落实法律、行政法规、部门规章和标准规定的安全措施的有效性，防控安全风险的有效性等情况进行全面评估，并且应该将评估报告通过全国互联网安全管理服务平台提交所在地地市级以上网信部

门和公安机关。

3. 网络安全义务

《网络安全法》对作为网络运营者的企业提出的合规义务可以总结为两个方面：一方面，从网络运行安全的角度出发，要求网络运营者应当按照网络安全等级保护制度的要求，履行安全保护义务，保障网络免受干扰、破坏或者未经授权的访问，防止网络数据泄露或者被窃取、篡改。另一方面，从网络信息安全的角度出发，要求网络运营者应当对其收集的用户信息严格保密，并建立健全用户信息保护制度，并采取技术措施和其他必要措施，确保其收集的个人信息安全，防止信息泄露、毁损、丢失。人工智能服务涉及大量网络数据和信息收集、处理环节，服务提供者属于网络运营者，应履行上述一般性网络安全合规义务，通过制度和技术手段保障网络安全。

4. 数据安全义务

《数据安全法》从多方面规定了企业的数据安全保护义务，包括数据分类分级、安全管理制度、风险监测、风险评估等，面向消费者提供生成式人工智能服务的平台运营方作为《数据安全法》项下的数据安全合规主体，因此也应当履行《数据安全法》项下的合规义务，包括但不限于：对数据的重要程度、敏感程度等进行分级，并根据其重要程度、敏感程度的不同进行分级保护；建立健全全流程数据安全管理制度，组织开展数据安全教育培训，采取相应的技术措施和其他必要措施，保障数据安全；加强风险监测，发现数据安全缺陷、漏洞等风险时，应当立即采取补救措施等。

5. 伦理治理及审查要求

人工智能的迅猛发展积极、深刻地改变了个人生活和社会运行，但同时也带来诸多伦理安全风险，《网络安全标准实践指南——人工智能伦理安全风险防范指引》概括的人工智能伦理安全风险包括失控性风险、社会性风险、侵权性风险、歧视性风险、责任性风险。

2023 年 10 月 18 日，中央网信办发布了《全球人工智能治理倡议》，立足人工智能发展、安全、治理三大核心方面，为全球人工治理问题共提出 11 项倡议，包括以人为本、尊重主权、智能向善、平等互利、敏捷治理、制度保障、公平和非歧视原则、伦理先行、协商共治、技术治理、增强发展中国家代表性等。

2023 年 10 月，科技部等多部门联合发布《科技伦理审查办法（试行）》，该办法于 2023 年 12 月 1 日起正式实施。该办法对于涉及以人为研究参与者的科技活动，包括利用人类生物样本、个人信息数据等的科技活动，或不直接涉及人或实验动物，但可能在

生命健康、生态环境、公共秩序、可持续发展等方面带来伦理风险挑战的科技活动进行的科技伦理审查和监管做出了规定，其中明确从事人工智能科技活动的单位研究内容涉及科技伦理敏感领域的应设立科技伦理（审查）委员会，如涉及具有舆论社会动员能力和社会意识引导能力的算法模型、应用程序及系统，还需报请所在地方或相关行业主管部门组织开展专家复核。提供教育 +AI 产品和服务的企业在其产品研发、运行过程中应充分尊重和遵守前述伦理治理要求，并按要求相应开展内外部伦理审查。

（四）AI 教育辅助工具的监管要求

随着教育 +AI 应用领域的进一步扩展，除上述通过线上提供智能教育外，也涌现出一批教育辅助工具类产品，如智能教育机器人、智能学习机、智能管理系统等。根据《双减政策》，线上培训机构不得提供和传播“拍照搜题”等惰化学生思维能力、影响学生独立思考、违背教育教学规律的不良学习方法，因此，AI 教育辅助产品不仅需要关注 AI 提供的便利以及对效率的提升，还需要关注产品是否符合国家教育政策。

智能教育辅助工具基于互联网连接硬件和软件平台，运用人工智能和大数据等信息技术，以满足不同教育场景的使用。因此，如果向网络用户有偿提供信息服务（即经营性互联网信息服务），可能需要就此申请取得《增值电信业务经营许可证》，产品研发和使用过程中数据来源合规、生成内容合规、知识产权保护、个人信息保护等问题也需要予以重视，具体参见下文第 3 部分“教育 +AI 的知识产权问题”和第 4 部分“教育 +AI 的个人信息保护及未成年人保护问题”。

三、教育 +AI 的知识产权问题

基于教育自身的特性，作为知识、技能、价值观传播和传递的方式，教育本身与知识产权有着天然的密不可分的关系，而 AI 技术与教育行业的结合，形成在线教育等新的教育模式以及应用 AI 技术的智能教育产品，在知识产权方面也为教育行业带来新的挑战。一方面，教育 +AI 的结合提升了教育信息化程度，助推了教育模式的革新，促进在线教育等模式的蓬勃发展，丰富了教育活动形成的作品的形式，需进一步厘清相关作品权属问题；另一方面，AI 技术在教育中的广泛应用也使得知识产权侵权的方式更为复杂多样，对企业规避知识产权侵权风险的要求不断提升。

（一）教育 +AI 形成作品的权属问题

教育 +AI 的结合丰富了教育活动形成的作品的形式，比如在教学活动中形成的在线授课内容、教学课件、授课视频、教材或试题库、在线教育软件等，这些内容如构成文学、

艺术和科学领域内具有独创性并能以某种有形形式复制的智力成果，均应属于著作权法保护的作品的范畴。

根据《著作权法》的一般规定，作品的著作权应归属于作者。所以，老师口述的在线授课内容、自行创作的教学课件等，如无特殊情况，著作权应归老师所有[3]。但是由于教育 +AI 形成的相关作品大部分需依托特定的教育机构或其他教育服务企业，作者可能为该等组织的员工或合作方，因此相关作品的著作权权属需进一步考虑职务作品、委托作品等情形，不能一概而论。

就职务作品而言，根据《著作权法》第十八条，如果作者（老师或其他教学、教研人员）是作为教育机构或相关企业的员工为完成教学、教研等工作任务而创作相关作品（如制作课件、编写教材等），则应属于职务作品，职务作品的权属存在两种情况：

1. 一般职务作品的著作权仍由作者享有，但作者单位有权在其业务范围内优先使用。作品完成两年内，未经单位同意，作者不得许可第三人以与单位使用的相同方式使用该作品。

2. 满足特定条件的职务作品，著作权可以由单位享有。比如员工利用教育机构的物质技术条件、并由机构承担责任的课件、教学及教辅材料、在线教育软件等；或者法律、行政法规规定，或者合同约定著作权由机构享有的相关职务作品。对于著作权由单位享有的职务作品，作为作者的老师或员工对作品仍享有署名权，且教育机构可以给予老师奖励。

就委托作品而言，著作权的归属可由委托人和受托人通过合同约定，合同未作明确约定或者没有订立合同的，著作权属于受托人。如果教学活动形成作品的作者系教育机构或相关企业员工以外的第三方，则该等作品可能构成机构委托其创作的作品，作品权属应通过委托合同、合作合同等进行约定，如无明确约定则著作权归第三方所有，教育机构作为委托方有权在约定的使用范围内享有使用作品或在委托创作目的的范围内使用作品。

在根据著作权法的要求不断厘清作品权属的同时，结合教育 +AI 领域的作品表现形式、传播方式的特点，相关企业或者作者为保护其对作品的著作权，可以采用必要的技术手段保护著作权，如在其作品中自动添加权利人及作者的水印、LOGO，采取技术手段限制非法复制等。

（二）AI 技术的应用提高了规避知识产权侵权风险的要求

AI 技术在教育中的广泛应用使得教育 +AI 领域知识产权侵权的方式更为复杂多样，

[3] 2023 年 4 月，北京互联网法院发布的数字教育著作权纠纷典型案例吉某诉北京某教育公司侵害作品信息网络传播权纠纷案中明确教师授课所产生的口述作品著作权一般归属于教师个人。原文见《4·26 特辑丨北京互联网法院数字教育著作权纠纷典型案例》，载微信公众号“北京互联网法院”， https://mp.weixin.qq.com/s/7oDyb2MpfuFzNy_aaXz4kA，2023 年 4 月 18 日发布。

一方面，教育 +AI 涉及大量学习内容、素材的使用，本身存在较强的知识产权合规需求，另一方面，新技术、新应用带来的新型侵权行为不断出现，例如点读笔产品借助点读笔上的摄像头可同步读出教材内容，又如 AI 早教机器人通过内置教材文件定向链接方式在线提供涉案教材的点读播放服务等。此外，AI 技术尤其是 AI 大模型的使用使得上述侵权风险被显著放大，AI 大模型的训练涉及到大量预训练、优化训练数据，数据中存在的内容不合规及侵权风险也会进一步传导到教育 +AI 应用层面，且可能进一步导致相关生成内容的侵权风险。新技术的诞生丰富了学习的方式，拓展了新的学习场景。但在这些新方式、新场景下，教育 +AI 应用及服务提供者更需关注其应用与服务的内容不应侵犯第三方的知识产权，或成为侵犯第三方知识产权的工具。

为了鼓励创作，促进文化科学的不断繁荣，著作权法在保护权利人的同时也规定了特定情形下可以不经权利人许可、不向其支付报酬的合理使用制度，就教育领域而言，可能涉及的合理使用的情形包括“为学校课堂教学或者科学研究，翻译、改编、汇编、播放或者少量复制已经发表的作品，供教学或者科研人员使用，但不得出版发行”。虽然著作权法规定了可以合理使用的例外情形，但由于这一制度本身系对著作权人权利的限制，因此除满足特定情形外还应在必要限度内使用，即不得影响作品的正常使用，也不得不合理地损害著作权人的合法权益；如果使用超过必要限度，导致影响原作品的使用或著作权人的合法权益，则仍应被认定为侵权。在教育 +AI 应用与服务中，伴随着新技术的发展，作品可能被迅速复制、在更广泛的范围内传播并被进一步演绎，这可能影响著作权人相关权能的实现；同时，商业化运用本身也存在被认定为超出合理使用范围的可能，导致相关作品的使用仍然构成侵权行为。

随着 AI 技术的进一步发展，教育 +AI 的应用和服务的外延仍将不断扩展，相关的知识产权侵权形式也可能不断增加，相关企业在业务发展的同时也需不断提升识别和规避知识产权侵权风险的能力，促进业务合规开展。

四、教育 +AI 的个人信息保护及未成年人保护问题

教育 +AI 主要依托平台、系统或通过软件 + 硬件等形式为学习者提供教育服务或教育辅助服务，为实现服务目的，需要收集学习者与服务相关的信息，其中包含学习者大量的个人信息。同时考虑受教育群体的年龄分布，教育 +AI 产品与服务的面向的群体中，大部分为不满十八周岁的未成年人，且其中很多为未满十四周岁的儿童。因此，教育 +AI 产品和服务的设计和开发需充分考虑个人信息保护及未成年人保护问题。

近年来，与未成年人个人信息保护相关的法律法规不断制定和实施，包括《中华人民

共和国个人信息保护法》《中华人民共和国未成年人保护法》《儿童个人信息网络保护规定》等。根据上述法律法规及个人信息保护相关的标准，就未成年人个人信息保护，现行规定根据未成年人的年龄制定了不同的保护要求，就收集年满十四周岁未成年人的个人信息前，应征得未成年人或其监护人的明示同意；而处理不满十四周岁未成年人个人信息的，应当取得未成年人的父母或者其他监护人的同意。

考虑到未成年人的心智发育程度，个人信息保护相关规定对于处理不满十四周岁未成年人个人信息设置了更为严苛的标准，比如将未满十四周岁未成年人个人信息直接规定为敏感个人信息，要求只有在具有特定的目的和充分的必要性，并采取严格保护措施的情形下才可以处理；单独设置儿童个人信息处理规则、用户协议和专人负责儿童个人信息网络保护；需要设置严格的内部访问权限等。

2023 年 10 月 16 日，国务院公开发布了《未成年人网络保护条例》，该条例作为我国第一部专门性的未成年人网络保护综合立法自 2024 年 1 月 1 日起施行。该条例对包含个人信息保护在内的未成年人网络保护各个方面进行了系统性的规定。

就个人信息保护方面，在现有未成年人个人信息保护相关规定的基础上，《未成年人网络保护条例》进一步明确了对于未成年人个人信息处理，处理者对内部工作人员应当以最小授权为原则，并要求个人信息处理者每年自行或聘请第三方对处理的未成年人个人信息进行合规审计，并将审计情况及时报告网信等部门。

就未成年人网络信息内容方面，《未成年人网络保护条例》明确要求不得制作、复制、发布、传播危害未成年人身心健康内容的网络信息，不得向未成年人发送、推送或者诱骗、强迫未成年人接触含有危害或者可能影响未成年人身心健康内容的网络信息等。教育服务与产品，尤其是面向未成年人的教育服务对其价值观的树立具有深远影响，因此相关服务和产品的内容更需严格审核把关。如近期某知名上市科技企业的智能教育产品中出现了违背主流价值观的内容，被家长投诉并对外发布，对其产品声誉及企业价值造成不利影响。

对于在线教育网络产品和服务相关的网络信息内容，《未成年人网络保护条例》进一步要求运营者根据不同年龄阶段未成年人的身心发展特点和认知能力提供相应的产品和服务，且需注意不得在首页首屏、弹窗、热搜等处于产品或者服务醒目位置、易引起用户关注的重点环节呈现可能影响未成年人身心健康的信息，不得通过自动化决策方式向未成年人进行商业营销等。

AI 原生应用相关法律问题研究之二：AI+ 医疗

唐丽子 孙及 周凡尧

微软研究院负责人彼得·李和其他两位合著者在其《超越想象的 GPT 医疗》一书中以丰富的案例结合深度的思考，探讨了以 GPT-4 为代表的大语言模型在医疗领域应用的诸多可能性，展现了未来医疗领域可能出现的人机结合的新型协作关系。医疗领域被认为是人工智能最值得应用的领域之一，随着人工智能技术的逐步发展，AI 在院内诊断与治疗等医疗服务、互联网医疗及健康管理、新药研发等领域为医疗行业赋能，显著提升相关医疗健康服务的水平和效率。

一、医疗 +AI 发展态势

（一）院内医疗服务

在院内医疗服务方面，按照应用场景分类，医疗 +AI 的主要应用包括 AI 医学影像、AI 医疗机器人、临床决策支持系统（CDSS）、智慧病案等。

1. AI 医学影像

医学影像是指针对人体或人体某部分，以非侵入方式取得内部组织影像的技术与处理过程。人工智能应用于医学影像，主要是利用相对成熟的图像识别算法，通过深度学习，实现机器对医学影像的分析判断，是协助医生完成诊断、治疗工作的一种辅助工具，帮助医生更快地获取影像信息，进行病灶筛查、靶区勾画、三维程序、图像分析、定量分析等。

医学影像市场需求大，在临床上，超过 70% 的诊断都依赖于医学影像，但受限于我国医疗资源分布不均，部分地区医生配备及经验不足，对医学影像分析的效率和准确性有待提高。AI 技术应用于医学影像能够显著提升医学影像阅片速度，提高诊断效率，减少错诊误诊等情况。

AI 医学影像是 AI+ 医疗发展较早的领域且相对成熟，目前市场上已有多款 AI 医学影像产品取得药监局颁发的第三类医疗器械产品注册证。

2. AI 医疗机器人

AI 医疗机器人是 AI+ 医疗的又一重要领域。AI 医疗机器人包括手术机器人、康复机器人，手术机器人基于立体视觉技术可以进行检测跟踪，在术前为医生提供个性化手术方案建议，在术中可以自主规划运动路径及范围，实现精准定位与控制，提升手术效率及精准度；康复机器人可以实现便携式穿戴，促进患者主动参与术后康复活动并客观评价康复训练的强度、时间、效果等，使得康复治疗更加系统化、规范化，满足患者术后长期康复治疗的需求。

3. 临床决策支持系统（CDSS）

临床决策支持系统（clinical decision support system，CDSS）指针对半结构化或非结构化医学问题，通过人机交互方式改善和提高决策效率的系统。CDSS 通过先进的分析能力和临床数据，促进医院、医生和患者之间的协调配合，提升临床诊疗的质量和效率，降低医院成本。人工智能技术的应用，特别是生成式人工智能技术应用于 CDSS 使系统具备强大的机器学习能力，在不断的训练和人机交互过程中总结和更新相关知识，并通过提取疾病关键信息，对疾病进行推理和判断，模拟人类专家的决策过程，辅助临床医师解决复杂医疗问题、提供诊断及个体化治疗决策，随着系统准确性的不断提高，CDSS 可以成为医生临床决策的有力助手，有效降低医生误诊率，显著提升医疗资源相对匮乏地区的整体医疗水平。

4. 智慧病案

病案即归档形成的完整病历，是医务人员在对患者进行问诊、检查、诊断、治疗、护理等医疗活动中形成的归档文字、图表、影像等材料，并进行综合、分析、整理后而书写成文的记录。智慧病案利用信息化及智能化技术，对医疗机构病历 / 病案数据进行处理，改变传统病案管理占用大量空间、资源、人力等问题，帮助病案工作者实现数据整合、数据质控、病案入库归档、病案数据应用服务的统一集成。

智慧病案管理中，病案质量的高低直接影响后续病案数据的使用，利用智慧化病案质控系统可以提高信息标准化程度，比如基于 AI 的知识图谱、自然语言处理等能力将患者口述病史整理为标准化的病历信息，可以有效提升电子病历准确率并提高医生效率，同时从源头加强病案质量管控，促进病案质控向智慧化发展。

（二）互联网医疗及健康服务

在互联网医疗及健康服务方面，医疗 +AI 可以有效提高相关服务的效率和体验，促进

产业链的延伸发展，成为互联网医疗及健康管理发展的重要助推力。

1. AI+ 互联网医疗

互联网医疗服务，包括在线问诊、诊后康复、慢病管理、互联网医院、医药电商等；与之相关的互联网医疗产业链还包括医疗信息化服务、互联网商保等。AI+ 互联网医疗可以有效提升互联网医疗服务体验，提供更为高效和个性化的服务，比如 AI 智能导诊基于其自然语言处理技术可以根据患者描述的症状和病史进行更为精准的分诊和导诊，找到患者所需要的科室和具有相应专长的医生；在线问诊过程中，相比于市面上大多数机器客服，AI 可以更为自然地与患者沟通，迅速了解患者病情并回应其问题，同时，基于患者提供的症状及病史，结合大数据分析，可以辅助执业医生为其提供针对特定患者的个性化临床诊断建议；在用药方面，AI 也可以通过对患者的病症、药物过敏史等信息进行深度分析，定制最适合患者的药物品种、剂量及用药方案，协助医生形成最佳治疗方案。AI+ 互联网医疗带来的效率提升可以促进互联网医疗服务的普及，使得医疗服务通过互联网等远程形式触达更广泛的群体。

在诊疗康复等服务基础上，互联网医疗基于其互联网基因，更容易与相关产业链互通互联，互相促进，比如互联网医疗可以延伸至日常健康管理，将短期医疗与长期健康管理相结合，综合分析用户的个体健康情况，为其提供更有针对性的医疗健康服务；又如互联网医疗与医疗商业保险，特别是互联网商保相结合，有助于提升保险服务效率、优化保险理赔流程等。AI 技术的加成将不断放大产业链延伸的促进作用，提升医疗健康综合服务水平，随着 AI 技术的进一步发展与大规模落地使用，AI+ 互联网医疗的想象空间十分广阔。

2. AI+ 健康管理

随着全民健康管理意识逐渐提高，除涉及医疗机构的相关诊疗及康复需求外，院外日常健康管理、长期慢性病管理等健康管理方面的需求日益增加，而互联网医疗的发展延伸以及相关便携式智能硬件设备为 AI+ 健康管理提供了广阔的发展空间。具体而言，AI+ 健康管理可以通过收集用户的健康数据、生活习惯及医疗记录，为用户提供个性化的健康管理建议和疾病预测。AI+ 健康管理的应用范围较为广泛，包括慢性病管理、运动管理、营养管理、睡眠监测等。

（三）新药研发

在新药研发方面，AI 技术将机器学习、深度学习、自然语言处理等技术应用于新药研发的各个环节，有助于缩短新药研发周期，降低研发成本。比如在药物发现阶段，利用

AI 机器学习、深度学习、大数据等技术，可以有效发现与疾病相关的基因、蛋白质和代谢途径，从而识别潜在的靶点，并进行大规模的药物筛选，发现具有潜在活性的药物分子，缩小候选药物的范围；在临床前开发阶段，AI 可以利用其深度学习及计算能力优化预测化合物成药性及配置药物晶型的过程；在临床开发阶段，AI 可以基于过去临床案例的成功和失败经验协助完善临床试验设计，同时基于海量患者数据精准匹配相应患者；在商业化阶段，AI 可以继续收集和分析新药使用的真实数据，对其安全性、有效性进行持续的分析反馈。

目前 AI 制药领域主要有三种商业模式，一种是为客户提供 AI 辅助药物开发平台的 SaaS 供应商，一种是为药物研发公司提供外包服务的 AI-CRO 模式，还有一种与 Biotech 模式类似，自行进行新药研发。

二、医疗 +AI 的主要法律问题

随着医疗 AI 行业的快速发展，相应的产业及监管政策也在不断出台。AI 赋能医疗相关业务打造的智慧医疗新业态既涉及传统医疗领域如医疗器械、药品研发等方面的监管，又基于 AI 技术带来的人机交互新范式与医疗健康等服务新模式，形成医疗 AI 软件驱动硬件、医疗 +AI 辅助临床决策、医疗 AI 提升互联网医疗效率与覆盖率等，相应也带来了数据安全、科技伦理等方面的问题。本文将从各类医疗 +AI 涉及的合规监管要求、医疗 +AI 的数据安全与个人信息安全、医疗 +AI 的伦理治理问题等方面展开。

（一）医疗 +AI 涉及的合规监管要求

如本文第一部分“医疗 +AI 的发展趋势”所述，AI 赋能医疗领域形成的应用具有丰富的场景，其适用的合规监管要求亦各有不同。在新药研发等领域，AI 技术的应用主要起到提升效率、降低成本的作用，在商业模式及合规监管方面与传统药物研发暂不具有实质性差异。在院内医疗、互联网医疗及健康服务等方面，医疗 +AI 形成的新的合规监管问题主要包括医疗 AI 软件注册问题，互联网医院准入及资质要求，医疗 AI 涉及的诊疗服务合规问题，医疗 AI 外资准入要求等，相关问题具体分析如下：

1. 医疗 AI 软件注册

《医疗器械监督管理条例（2021 修订）》作为医疗器械监管的基本规则，明确了医疗器械，是指直接或者间接用于人体的仪器、设备、器具、体外诊断试剂及校准物、材料以及其他类似或者相关的物品，包括所需要的计算机软件。在此基础上，国家药品监督管理局发布的《人工智能医用软件产品分类界定指导原则》及国家药品监督管理局医疗器械

技术审评中心《人工智能医疗器械注册审查指导原则》为人工智能医疗器械提供通用指导原则。

根据《人工智能医用软件产品分类界定指导原则》，人工智能医用软件是指基于医疗器械数据，采用人工智能技术实现其医疗用途的独立软件。为进一步界定人工智能医用软件是否属于医疗器械，《人工智能医用软件产品分类界定指导原则》从软件的处理对象、核心功能及用途角度作进一步明确：（1）若软件产品的处理对象为医疗器械数据，且核心功能是对医疗器械数据的处理、测量、模型计算、分析等，并用于医疗用途的，作为医疗器械管理；（2）若软件产品的处理对象为非医疗器械数据（如患者主诉等信息、检验检查报告结论），或其核心功能不是对医疗器械数据进行处理、测量、模型计算、分析，或不用于医疗用途的，不作为医疗器械管理。

基于上述定义，判断医疗 AI 软件是否需要或可以注册为医疗器械时考虑该等软件是否具备如下特点：第一，软件具备一个或多个医疗目的 / 用途，如疾病的诊断、预防、监护、治疗或者缓解，损伤的诊断、监护、治疗、缓解或者功能补偿，生理结构或者生理过程的检验、替代、调节或者支持，生命的支持或者维持，妊娠控制，通过对来自人体的样本进行检查，为医疗或者诊断目的提供信息等；第二，软件的处理对象一般为医疗器械数据，且核心功能为对医疗器械数据的处理、测量、模型计算、分析等；前述“医疗器械数据”通常指医疗器械产生的用于医疗用途的客观数据，特殊情形下可包含通用设备产生的用于医疗用途的客观数据。

实践中，我们理解大部分应用场景下的医疗 AI 软件具有人工智能医用软件的特点，或其部分功能具备上述特点，因此建议医疗 AI 企业在开发软件的同时对其涉及的各项功能及应用场景进行全面审视，涉及医疗器械监管的应及时取得相应的注册证。

2. 互联网医疗的准入及资质要求

根据《医疗机构管理条例（2022 修订）》《互联网诊疗管理办法（试行）》《互联网医院管理办法（试行）》等法律规范的规定及主管部门的相关说明[1]，医疗机构通过互联网为患者提供诊疗服务包括两种情形：一是实体医疗机构使用本机构注册的医务人员，利用互联网技术直接为患者提供部分常见病、慢性病复诊和家庭医生的签约服务；二是互联网医院包括作为实体医疗机构第二名称的互联网医院，以及依托实体医疗机构独立设置的互联网医院。互联网医院可以使用在本机构和其他医疗机构注册的医师开展互联网诊疗活动，可以为患者提供部分常见病、慢性病复诊和家庭医生的签约服务。

上述两种模式中，第一种互联网诊疗服务由实体医疗机构提供，其准入程序相对较为

[1] https://www.gov.cn/xinwen/2018-09/14/content_5322040.htm。

简单，即按照《互联网诊疗管理办法（试行）》相关规定就其开展互联网诊疗活动事宜在申办医疗机构时或取得执业证后向其执业登记机关进行申请[2]；第二种互联网医院服务的提供方为互联网医院，目前互联网医院的类型及准入程序相对较为复杂，具体如下：

互联网医院目前包括两种类型：由第三方机构独立设置的互联网医院（需以合作形式依托实体医疗机构），以及作为实体医疗机构第二名称的互联网医院，后者还分为实体医疗机构与第三方机构合作建立的第二名称互联网医院，及实体医疗机构独立设置的第二名称互联网医院。根据《互联网医院管理办法（试行）》的规定，第三方申请设置独立的互联网医院，应当向其依托的实体医疗机构执业登记机关提出设置申请，并提交设置申请书、设置可行性研究报告、所依托实体医疗机构的地址及申请方与实体医疗机构签署的合作建立互联网医院的协议书，经批准后申请执业登记；实体医疗机构申请设置作为第二名称的互联网医院，应当向其执业登记机关提出增加互联网医院作为第二名称的申请，经批准后进行变更登记。除上述设立流程和要求外，互联网医院命名还需符合对应的命名要求[3]，在诊疗科目、科室设置、人员、房屋和设备设施、规章制度等方面满足规范要求。上述不同类型的互联网医院对应的法律责任承担主体亦有不同，由第三方独立设置的取得《医疗机构执业许可证》的互联网医院，独立作为法律责任主体；实体医疗机构以互联网医院作为第二名称时，实体医疗机构为法律责任主体；互联网医院合作各方按照合作协议书承担相应法律责任。

除上述准入程序外，互联网诊疗服务及互联网医院依托互联网技术，还可能涉及互联网相关的资质要求：

（1）增值电信业务资质：互联网诊疗服务及互联网医院服务一般涉及向网络用户有偿提供信息服务（即经营性互联网信息服务），国家对经营性互联网信息服务实行许可制度，相应通常需要就此申请取得《增值电信业务经营许可证》；此外，互联网诊疗服务及互联网医院服务还可能涉及向患者出售药品，即涉及增值电信业务中的在线数据处理与交易处理业务，相应的医疗机构亦需取得该类业务所需的《增值电信业务经营许可证》。

（2）互联网药品信息服务资格证书：互联网诊疗服务及互联网医院服务提供在线问诊、在线开具处方等过程中可能涉及向患者提供用药建议或与药品信息相关的咨询服务，上述服务涉及《互联网药品信息服务管理办法》通过互联网向上网用户提供药品（含医疗器械）信息的服务活动。根据该办法，无论上述互联网药品信息服务活动是否为经营性活动，相应的网站主办单位均需申请《互联网药品信息服务资格证书》，并在网站主页显著

---

[2] 《互联网诊疗管理办法（试行）》第六条、第七条、第八条、第九条。

[3] 《互联网医院管理办法（试行）》第十二条 互联网医院的命名应当符合有关规定，并满足以下要求：（一）实体医疗机构独立申请互联网医院作为第二名称，应当包括“本机构名称＋互联网医院”；（二）实体医疗机构与第三方机构合作申请互联网医院作为第二名称，应当包括“本机构名称＋合作方识别名称＋互联网医院”；（三）独立设置的互联网医院，名称应当包括“申请设置方识别名称＋互联网医院”。

位置标注《互联网药品信息服务资格证书》编号。

（3）信息安全等级保护备案：《互联网诊疗管理办法（试行）》《互联网医院管理办法（试行）》要求开展互联网诊疗的信息系统、互联网医院信息系统按照国家有关规定实施第三级信息安全等级保护[4]。根据《信息安全等级保护管理办法》，第二级以上信息系统应到公安主管部门办理信息系统安全保护等级备案手续。

除前述准入及资质要求外，对于互联网医院提供诊疗服务的医师，《互联网医院管理办法（试行）》也规定了相应的资质和经验要求，包括：依法取得相应执业资质并在依托的实体医疗机构或其他医疗机构注册，具有 3 年以上独立临床工作经验。且互联网医院提供服务的医师，应当确保完成主要执业机构规定的诊疗工作。

3. 医疗 AI 诊疗服务合规

医疗 AI 提供诊疗服务涉及线上线下等不同场景，如 AI+ 医学影像实现机器对医学影像的分析判断，AI 医疗机器人辅助手术，通过临床决策支持系统辅助医生进行临床诊断以及通过 AI+ 互联网医疗提供在线问诊等服务。AI 技术目前仍在发展过程中，与技术相伴而生的安全风险、伦理道德风险等目前尚无有效的控制措施，目前在诊疗服务中，监管层面对于医疗 AI 的使用仍持有相对谨慎的态度，比如应用 AI 手术机器人辅助实施手术目前仍属于限制类技术，需遵守相应的临床应用规范，又如互联网医疗在线问诊、处方开具等服务中严格限制人工智能的应用。相关要求具体如下：

就AI手术机器人的使用，根据国家卫健委发布的《国家限制类技术目录（2022年版）》，人工智能辅助治疗技术，即应用机器人手术系统辅助实施手术的技术属于限制类技术。《医疗技术临床应用管理办法》要求对限制类技术实施备案管理，医疗机构拟开展限制类技术临床应用的，应当按照相关医疗技术临床应用管理规范进行自我评估，符合条件的可以开展临床应用，并于开展首例临床应用之日起 15 个工作日内，向核发其《医疗机构执业许可证》的卫生行政部门备案。《国家限制类技术临床应用管理规范（2022 年版）》中对于使用人工智能辅助治疗技术规定了医疗机构层面和人员层面的资质要求，一方面要求医疗机构执业资质及技术能力、相关科室的经验与能力、手术室要求、辅助科室及设备要求等，另一方面要求开展人工智能辅助治疗技术的团队具备至少 4 名（心脏大血管外科至少 6 名）经专业培训并考核合格的、具备人工智能辅助治疗技术临床应用能力的医师、护士和（或）技师，其中医师需具有 10 年以上临床经验及副主任医师以上专业技术职务资格。此外，临床应用管理规范还规定了实施人工智能辅助治疗技术相关的技术管理、人员培训要求等。

[4] 《互联网诊疗管理办法（试行）》第十三条，《互联网医院管理办法（试行）》第十五条。

在 AI+ 互联网诊疗方面，虽然 AI 赋能可以提高智能导诊、智能分诊效率，但就具体的诊断而言，目前从监管角度对 AI 的使用仍进行严格限制。《互联网诊疗监管细则（试行）》中明确规定，人工智能软件等不得冒用、替代医师本人提供诊疗服务；处方应由接诊医师本人开具，严禁使用人工智能等自动生成处方。部分地方已出台的地方性互联网诊疗监管规定[5]中也包含上述要求。《处方管理办法》中也明确规定，经注册的执业医师在执业地点取得相应的处方权，医疗机构使用未取得处方权的人员、被取消处方权的医师开具处方的，可能被处以责令限期改正、罚款、吊销执业许可等处罚，医师未取得处方权或者被取消处方权后开具药品处方的，可能被处以暂停职业活动、吊销职业资格等处罚。如果互联网诊疗服务中违反上述规定使用人工智能技术造成医疗事故，则根据《医疗事故处理条例》规定，由于医疗机构及其医务人员使用产品的过程中，违反医疗卫生管理法律、行政法规、部门规章和诊疗护理规范、常规，过失给患者造成人身损害的事故，医院应当向患者承担民事赔偿责任。情节严重的情况下，负有责任的医院和医务人员可能承担行政责任（如限期停业整顿、吊销执业许可等），甚至可能承担刑事责任（医疗事故罪）。

4. 医疗 AI 外资准入要求

医疗 +AI 涉及的应用领域较多，不同应用领域的外商投资要求也有所差异，总体而言我国目前对于医疗 AI 硬件领域的外商投资持鼓励态度，但对互联网医疗服务、涉及遗传信息相关的领域外商投资进行了限制或禁止。医疗 AI 相关的外资准入要求简要梳理如下：

根据国家发改委和商务部发布的现行有效的《鼓励外商投资产业目录（2022 年版）》与医疗 AI 相关的鼓励外商投资产业集中在医疗 AI 设备制造领域，包括：医用成像设备（高场强超导型磁共振成像设备、X 线计算机断层成像设备、数字化彩色超声诊断设备等）、医疗影像智能辅助诊断系统及关键部件的制造；人工智能辅助医疗设备制造；高端放射治疗设备制造；高端手术器械、理疗康复设备、可穿戴智能化健康装备制造；微创手术医疗设备开发、生产：3D 成像、电子显微系统、手术机器人、机械臂、助听器及人工耳蜗等；智慧健康养老产品的研发、制造（老年用品及辅助产品制造，老年医疗器械和康复辅具制造，老年人智能与穿戴设备制造等）等。

根据国家发改委和商务部发布的现行有效的《外商投资准入特别管理措施（负面清单）（2021 年版）》，医疗机构属于限制外商投资的领域；人体干细胞、基因诊断与治疗技术开发和应用属于禁止外商投资的领域。根据《中外合资、合作医疗机构管理暂行办法》，中方在中外合资、合作医疗机构中所占的股权比例或权益不得低于 30%。此外，如前所述，从事互联网医疗服务还可能涉及提供经营性互联网信息服务，《外商投资准入特别管理措

[5] 如山东省卫健委发布的 2023 年 3 月 2 日起实施的《山东省互联网诊疗管理实施办法》、安徽省卫健委发布的 2022 年 9 月 30 日期实施的《安徽省互联网诊疗监管实施办法（试行）》等。

施（负面清单）（2021 年版）》要求从事该等业务的外资股比不超过 50%。AI+ 互联网医疗、AI 制药等领域企业在融资发展过程中，需关注上述外商投资限制。

除上述医疗 +AI 涉及的特殊合规监管要求外，医疗 +AI 也需要遵守并履行人工智能领域相关的一般合规义务，如果医疗 +AI 提供的算法推荐服务、生成式人工智能服务或深度合成服务具有舆论属性或者社会动员能力，应履行算法备案义务、安全评估义务等，具体可参见本书《AI 原生应用相关法律问题研究之（一）AI+ 教育法律问题》一文中关于提供生成式人工智能服务相关的合规义务的相关介绍。

（二）医疗 +AI 的数据安全与个人信息保护

数据是 AI 发展的基础之一，也是医疗行业实践和发展的必要条件。医疗 +AI 的发展得益于 AI 的机器学习、深度学习、自然语言处理、海量数据处理等能力，与之相伴而来的，则是医疗数据安全保护的需求。医疗数据安全涉及大量的患者、受试者个人信息安全以及公共卫生数据等重要数据安全，近年来网络安全、数据安全、个人信息保护相关法律法规不断完善，医疗 +AI 的发展应重视其中涉及的数据安全及个人信息保护问题，规避相关的合规风险。

1. 医疗 +AI 涉及的数据类型

医疗数据，是指个人在接受健康医学诊疗过程中产生的数据，相较于其他行业，医疗行业的数据来源和呈现方式更为多样和丰富，比如患者用药、检验结果、病历、检查报告、医学影像、视频、文献等。结合《信息安全技术——健康医疗数据安全指南》相关规定及法律行业实践，医疗数据大致包括：个人健康信息、人口健康信息、遗传资源信息、病历（电子病历）信息、个人健康医疗数据、医疗健康大数据、医疗支付记录、卫生资源数据、公共卫生数据等。

《网络安全法》《数据安全法》《个人信息保护法》构成现阶段数据安全与个人信息保护的基本规范，其中对于不同类型的数据 / 个人信息规定了相应的保护要求，基于上述法律的视角，医疗 +AI 涉及的数据类型包括：

（1）个人信息：根据《个人信息保护法》的定义[6]，医疗 +AI 涉及的个人信息包括医疗机构在提供医疗服务过程中收集的患者姓名、年龄、性别以及互联网医疗服务或其他医疗 +AI 软件服务中收集的用户手机号等。

（2）敏感个人信息：根据《个人信息保护法》的定义[7]，医疗 +AI 涉及的个人基因、

[6] 《个人信息保护法》第四条 个人信息是以电子或者其他方式记录的与已识别或者可识别的自然人有关的各种信息，不包括匿名化处理后的信息。

[7] 《个人信息保护法》第二十八条 敏感个人信息是一旦泄露或者非法使用，容易导致自然人的人格尊严受到侵害或者人身、财产安全受到危害的个人信息，包括生物识别、宗教信仰、特定身份、医疗健康、金融账户、行踪轨迹等信息，以及不满十四周岁未成年人的个人信息。

指纹、虹膜、声纹、掌纹、面部识别特征等生物识别信息及个人健康状况等显然都属于敏感个人信息。

（3）重要数据及核心数据：根据《数据安全法》，一旦遭到篡改、破坏、泄露或者非法获取、非法利用，对国家安全、公共利益或者个人、组织合法权益造成直接危害的数据构成重要数据，而关系国家安全、国民经济命脉、重要民生、重大公共利益等数据属于国家核心数据。参考正在审查中的最新版《信息安全技术——重要数据识别指南》，反映群体健康生理状况、族群特征、遗传信息等的基础数据，如人口普查资料、人类遗传资源信息、基因测序原始数据属于重要数据。

2. 医疗 +AI 数据收集处理合规要点

结合《网络安全法》《数据安全法》《个人信息保护法》及相关法律法规对数据和个人信息收集处理的一般性要求，以及《医疗机构病历管理规定（2013 年版）》《电子病历应用管理规范（试行）》《人口健康信息管理办法（试行）》《人类遗传资源管理条例》《人类遗传资源管理条例实施细则》等行业领域内特定类型数据管理规范，从医疗 +AI 开发实践角度，需要重点关注的数据收集处理合规要点如下：

根据《个人信息保护法》的要求，就一般个人信息的收集处理，收集方应充分履行告知同意义务，个人信息处理行为需取得个人信息主体的同意；就敏感个人信息的收集和处理，需满足特定目的和充分必要性的要求，同时须取得个人的单独同意并采取更为严格的保护措施，如将个人身份信息与生物识别信息分别存储等，此外，敏感个人信息处理者还应当进行事前的个人信息保护影响评估。针对患者病历等数据，《医疗机构病历管理规定（2013 年版）》规定医疗机构及其医务人员应当严格保护患者隐私，禁止以非医疗、教学、研究目的泄露患者的病历资料；《电子病历应用管理规范（试行）》要求电子病历系统应当为操作人员提供专有的身份标识和识别手段，并设置相应权限。

根据《数据安全法》的规定，对于重要数据处理者，在一般的数据保护要求基础上，还需符合额外的规定，包括重要数据的处理者应当明确数据安全负责人和管理机构，落实数据安全保护责任；重要数据处理者应当对数据处理活动定期开展风险评估，并将风险评估报送至有关主管部门，重要数据出境还应当向相关网信部门申报数据出境安全评估等。《人口健康信息管理办法（试行）》要求人口健康信息实行分级存储，不得将人口健康信息在境外的服务器中存储，不得托管、租赁在境外的服务器。

除上述一般要求外，针对基因等人类遗传资源相关的数据，其采集、保藏、买卖、利用和对外提供需遵守《人类遗传资源管理条例》《人类遗传资源管理条例实施细则》的特

殊规定，如《人类遗传资源管理条例》明确要求外方单位（包括境外组织及境外组织、个人设立或者实际控制的机构[8]）不得在我国境内采集、保藏我国人类遗传资源，不得向境外提供我国人类遗传资源；采集、保藏、利用、对外提供我国人类遗传资源应进行伦理审查；采集特定类型遗传资源需经国务院科学技术行政部门批准等。

（三）医疗 +AI 的伦理治理问题

医疗 +AI 应用中的伦理问题，既涉及 AI 技术带来的常规伦理问题，又涉及与医疗相关的特殊伦理问题。医疗 +AI 应用中基于人工智能的技术能力和应用场景，结合人机关系互动的强度，AI 在医疗卫生领域的作用呈现以下三个递进层次：作为工具的人工智能旨在提高效率，作为合作者的人工智能用于辅助决策，作为支持者的人工智能替代决策。随着 AI 参与程度的加深，人类在医疗卫生领域的主体性存在不断被削弱的风险，AI 深度学习的算法黑箱导致无法对其输出决策的逻辑和原理进行判断，其决策存在潜在安全风险，此外，医疗 AI 的应用还会带来医患关系异化、医疗权责划分、隐私保护等诸多方面的挑战。

全球围绕医疗 +AI 的伦理问题已经展开广泛的讨论，世界卫生组织在其发布的《医疗卫生中人工智能的伦理治理》指南中提出了医疗 +AI 的伦理治理的六项原则，包括保障人类自主权，增进人类福祉和保护安全及公共利益，确保透明度、可解释性和可理解性，发展责任和问责制，确保包容性和公平性，促进响应性和可持续性。

就 AI 伦理治理问题，2017 年，国务院印发《新一代人工智能发展规划》并在该规划中提出了制定促进人工智能发展的法律法规和伦理规范的要求，2021 年修订的《中华人民共和国科学技术进步法》规定国家建立科技伦理委员会，完善科技伦理制度规范，加强科技伦理教育和研究，健全审查、评估、监管体系。2022 年 3 月，中共中央办公厅、国务院办公厅发布了《关于加强科技伦理治理的意见》，提出了科技伦理治理原则以及基本要求，其中提出科技伦理的五项原则，包括：增进人类福祉、尊重生命权利、坚持公平公正、合理控制风险和保持公开透明。

人工智能、生命科学和医学均属于科技活动的核心领域，相较近年来飞速发展而被纳入伦理审查范畴的人工智能而言，生命科学和医学相关的伦理审查要求和实践更为普遍，《人类遗传资源管理条例》《药品管理法》《医师法》等生命科学和医学相关的法规中，对于开展遗传资源采集、利用等以及进行药物、医疗器械临床试验、医疗技术临床应用等规定了事前伦理审查的制度。随着 AI 与医疗的深入结合，医疗 +AI 将显著拓展医疗健康

[8] 《人类遗传资源管理条例实施细则》第十二条 本实施细则第十一条所称境外组织、个人设立或者实际控制的机构，包括下列情形：（一）境外组织、个人持有或者间接持有机构百分之五十以上的股份、股权、表决权、财产份额或者其他类似权益；（二）境外组织、个人持有或者间接持有机构的股份、股权、表决权、财产份额或者其他类似权益不足百分之五十，但其所享有的表决权或者其他权益足以对机构的决策、管理等行为进行支配或者施加重大影响；（三）境外组织、个人通过投资关系、协议或者其他安排，足以对机构的决策、管理等行为进行支配或者施加重大影响；（四）法律、行政法规、规章规定的其他情形。

领域适用科技伦理审查的范围，无论是传统医疗企业入局医疗 +AI 业务，还是互联网等技术企业进军医疗 +AI 市场，都需关注和遵守这一领域的科技伦理审查要求，在医疗 +AI 应用开发过程中尊重和遵守相关的伦理治理要求。

就科技伦理审查的相关具体要求，2023 年 10 月，科技部等多部门联合发布《科技伦理审查办法（试行）》，对于科技伦理审查的基本程序、标准、条件等提出统一要求。该办法对于涉及以人为研究参与者的科技活动，包括利用人类生物样本、个人信息数据等的科技活动，或不直接涉及人或实验动物，但可能在生命健康、生态环境、公共秩序、可持续发展等方面带来伦理风险挑战的科技活动进行的科技伦理审查和监管做出了规定，其中明确从事生命科学、医学、人工智能等科技活动的单位，研究内容涉及科技伦理敏感领域的，应按要求设立科技伦理（审查）委员会，并在设立科技伦理（审查）委员会后 30 日内，通过国家科技伦理管理信息登记平台进行登记。科技伦理审查的一般由科技伦理（审查）委员会进行，特别的，医疗健康领域涉及对人类生命健康、价值理念、生态环境等具有重大影响的新物种合成研究，改变人类生殖细胞、受精卵和着床前胚胎细胞核遗传物质或遗传规律的基础研究，侵入式脑机接口用于神经、精神类疾病治疗的临床研究，对人类主观行为、心理情绪和生命健康等具有较强影响的人机融合系统的研发等活动，还需报请所在地方或相关行业主管部门组织开展专家复核（相关活动由国家实行行政审批等监管措施且将符合伦理要求作为审批条件、监管内容的除外）。

根据《科技伦理审查办法（试行）》，科技伦理审查重点审查科技活动是否符合科技伦理原则，参与科技活动的科技人员资质、研究基础及设施条件等是否符合相关要求。涉及以人为研究参与者的科技活动，重点审查所制定的招募方案公平合理，生物样本的收集、储存、使用及处置合法合规，个人隐私数据、生物特征信息等信息处理符合个人信息保护的有关规定，对研究参与者的补偿、损伤治疗或赔偿等合法权益的保障方案合理，对脆弱人群给予特殊保护；所提供的知情同意书内容完整、风险告知客观充分、表述清晰易懂，获取个人知情同意的方式和过程合规恰当等。

AI 原生应用相关法律问题研究之三：AI+ 游戏法律问题

唐丽子 孙及 贾潇寒

2023 年年初，微软在北京举办了“GDC 2023 中国行—予力游戏赋能开发”大会，公布了多项 AI 技术在游戏开发领域的技术性进展和落地应用场景，从美术资源制作到游戏体验优化再到运营和营销，AI 技术能够在游戏的各个领域赋能，帮助游戏厂商降低开发成本，大幅提高游戏生产的品质和效率，也正因为如此，集合文本、图像、音频、视频等要素于一体的游戏产业，势必将成为 AI 技术落地的最佳渠道之一。本篇将着眼于 AI 与游戏行业的结合，探讨相关法律问题。

一、游戏 +AI 的发展趋势和应用场景

大型游戏的开发过程通常包含文本、图像、音效、音乐、3D 模型、动画、电影、代码等极为丰富的环节，涵盖了娱乐及媒体行业所有的内容形式。传统游戏的开发过程耗时漫长且流程复杂，开发单位以年为计投入大量人力、物力、财力，行业内存在着“质量、速度、成本”中只能有两个的不可能三角，但生成式 AI 能够有效解决生产力的问题。以销量超 4600 万份的《荒野大镖客：救赎 2》为例，游戏拥有超过 28 平方英里接近真实景象的地图和 1000 个 NPC，即使专职开发人员超 1200 人，也用了 8 年才完成，开发成本近 3 亿美元，而专职开发人员少于《荒野大镖客：救赎 2》的《星际公民》，已开发了 10 年之久，至今尚未正式发售。而随着 AI 技术的逐渐成熟，AI 可以优化从游戏策划到剧情、音频、图像、动画制作再到宣发等游戏制作全流程，提高开发人员创造效率，减少研发周期和人员规模。例如，Ghostwriter 可帮助研发人员设计游戏剧情和对话内容；Stable Diffusion 可快速创建成场景、道具、武器等游戏资产。目前，在 B 站上全程使用 AI 技术生成一款互动游戏，整个制作过程最快仅需 6 小时，与传统游戏的制作周期相比，实可谓是一项大革新。

在游戏行业领域内，AI 技术的应用场景非常丰富：

1. 绘画设计。AI 辅助绘画目前可在角色、装饰、场景、3D 影像生成、画风模仿等方面落地应用，研发人员只需输入关键词和素材，即可生成相应图像，帮助美术人员快速完成绘画设计，提升工作效率。随着 Stable Diffusion 等工具突破生成精度等问题，手握

强大 AI 工具，美术工作人员可达成“一人成军”的效果。例如，仿真游戏《微软模拟飞行 2020》与 Blackshark.ai 合作，借助 AI 和云计算，通过 2D 卫星图像生成世界各地约 15 亿座 3D 建筑物，并保证数据实时交互，突破了人工制作的桎梏。

2. 代码编写。大型游戏通常具有很高的技术经验门槛，需要编写复杂算法以支撑游戏运行。AI 模型可作为代码生成工具，根据用户的自然语言指令生成相应的代码程序。AI 可精确处理原本耗时的撰写代码、优化代码等基础工作，将研发人员从繁复的编写代码工作中解放出来，为其完成更多创造性工作节约时间和精力的同时，大幅降低游戏制作门槛，推动更多新游戏面世。

3. 内容生成。AI 可辅助设计游戏中的剧情及对话，根据用户要求生成不同场景、背景、角色的个性化内容，玩家和玩家、玩家和 NPC 的交互不再受限于固定设置，而是以更高的自由度参与游戏世界，游戏内容将极度个性化。

4. NPC 互动。AI 技术可进一步提高 NPC 真实性、合理性，提高其多轮对话能力，弱化其和真人玩家的边界感，为玩家提供千人千面的游戏反馈，提升玩家的操作空间和体验感。目前已经有不少公司，比如国外的 inwold.ai，中国的启元世界、超参数，都在尝试做 AI NPC，希望 NPC 能够在一个开放世界大环境里，基于自己的人设、和其他人的关系等，和外界自主涌现交互，而不是被简单规则约束的玩偶。

可以说，面对越来越高的开发成本，游戏行业降本增效以提高盈利能力、行业竞争力的强烈需求，共同推动了 AI 技术与游戏行业的深度结合。

二、游戏 +AI 的合规要点问题

从现阶段 AI 技术的发展程度看，AI 赋能游戏行业目前主要可能涉及到以下合规要点问题：

（一）数据来源的合法性问题

根据《生成式人工智能服务管理暂行办法》第七条，生成式人工智能服务提供者在开展预训练、优化训练等训练数据处理活动，应当确保使用具有合法来源的数据和基础模型。

如前文介绍，游戏角色形象、场景设置、对话台词、背景音效等游戏组成元素均可以通过 AI 技术辅助搭建，这个过程需要输入大量素材作为训练数据。因此，对于游戏开发企业来说，首先要关注的就是使用的数据是否具有合法来源。目前游戏厂商的数据来源主要包括自行采集、向第三方采购数据和通过公开渠道爬取数据等途径。其中，自行采集数据只要事先向用户做好告知义务并取得用户的授权同意，一般情况下不存在太大的合规问

题。可能存在问题的主要是向第三方采购数据和通过公开渠道爬取数据。

向第三方采购数据的，若第三方提供的数据权属存在瑕疵，或者未经数据所有者的授权提供，可能导致游戏开发企业使用第三方数据的合法性基础存在瑕疵，影响游戏业务的持续运营，甚至引发对游戏开发企业的索赔。

通过公开渠道（例如搜索引擎、社交平台、电商平台、其他游戏平台网站等）爬取数据是目前企业采集公开数据的常用技术手段之一，但爬取过程需要遵循合规路径，否则可能涉嫌非法爬虫行为。目前，司法实务界主要从以下四个因素来判断爬虫行为是否合法：

- 一是数据是否属于开放数据。数据是否公开不是合法性判断的标准，是否为开放数据才是，公开数据不必然等同于开放数据；
- 二是取得数据的手段是否合法。爬虫采用的技术是否突破数据访问控制，是否突破网站的 Robots 协议[1]。在司法实践中，Robots 协议规则是判断抓取信息一方行为正当性的关键因素，Robots 协议约定不能爬取的范围就是爬虫的红线；
- 三是使用目的是否合法。如果爬虫的目的是实质性替代被爬者提供的部分产品内容或服务，则会被认为目的不合法；
- 四是是否造成损害。爬虫是否实质上妨碍被爬者的正常经营，是否不合理增加运营成本，是否破坏系统正常运行。

因此，企业如果违反被爬网站的 Robots 协议，爬取了被爬网站授权范围以外的数据，用于实质性替代被爬对象的产品或服务，可能被认定为非法爬取行为，除了可能会涉及不正当竞争行为[2]的行政责任，还可能构成侵犯权利人合法权益，涉及到民事侵权责任。此外，若爬取方明知没有授权而故意避开或强行突破网站的反爬虫技术设置进行爬取，属于“未经授权”访问或获取数据，根据我国刑法规定，突破技术屏障入侵他人计算机系统、获取系统内的数据，可能构成“非法侵入计算机信息系统罪”“非法获取计算机信息系统数据罪”“破坏计算机信息系统罪”等刑事责任：

- 《数据安全法》第三十二条：“任何组织、个人收集数据，应当采取合法、正当的方式，不得窃取或者以其他非法方式获取数据。法律、行政法规对收集、使用数据的目的、范围有规定的，应当在法律、行政法规规定的目的和范围内收集、使用数据。”

---

[1] Robots 协议也称爬虫协议、爬虫规则，Robots 协议主要是限制网络爬取数据的行为。被爬取数据方将写有可爬取信息范围的 Robots 协议文件放到该网站，仅允许数据爬取方在协议范围内爬取数据。当一个爬虫程序访问一个站点时，它会首先检查该站点根目录下是否存在 robots.txt，如果存在，爬虫程序就会按照该文件中的内容来确定访问的范围；如果该文件不存在，所有的爬虫程序将能够访问网站上所有没有被口令保护的页面。

[2] 《反不正当竞争法》第二条第二款，本法所称的不正当竞争行为，是指经营者在生产经营活动中，违反本法规定，扰乱市场竞争秩序，损害其他经营者或者消费者的合法权益的行为。

- 《数据安全法》第五十一条：“窃取或者以其他非法方式获取数据，开展数据处理活动排除、限制竞争，或者损害个人、组织合法权益的，依照有关法律、行政法规的规定处罚。”
- 《数据安全法》第五十二条：“违反本法规定，给他人造成损害的，依法承担民事责任。违反本法规定，构成违反治安管理行为的，依法给予治安管理处罚；构成犯罪的，依法追究刑事责任。”

此外，若被认定为非法爬虫行为，则有关著作权保护的“避风港”原则[3]可能不再奏效。以 D 公司诉 A 公司侵犯著作权案为例，A 公司违反 Robots 协议爬取 D 公司网上的商户基本信息及点评内容，将其刊登在自身运营的 A 公司网站上，使得用户无需到达 D 公司网站即可获得商户基本信息及点评内容。法院认为，A 公司的行为超出了“合理使用”的范畴，已构成对 D 公司的市场化替代，其行为违反了《著作权法》规定，不具有合法性，对于 A 公司依据“避风港原则”提出的抗辩理由“已在接到 D 公司的通知后对相关侵权内容进行了删除”最终不予认可。

（二）游戏素材的版权侵权问题

根据《生成式人工智能服务管理暂行办法》的相关规定，提供和使用生成式人工智能服务，应当尊重知识产权、商业道德。生成式人工智能服务提供者开展预训练、优化训练等训练数据处理活动，涉及知识产权的，不得侵害他人依法享有的知识产权。

AI 游戏开发涉及的训练素材通常会包含大量他人享有著作权的文字、绘画、音乐等内容，若未获得著作权人的授权使用，可能导致著作权侵权行为。例如：如果游戏开发者利用 AI 合成技术将某受著作权保护的音乐、图像合成在游戏中，可能会涉及对原有作品的复制；如果开发者将原有作品作为参考图添加进图像编码器，按照 AI 指令进行处理并最终生成目标图像，可能会涉及到对原有作品的改编；如果 AI 生成的游戏产品与原有作品构成了实质性相似，构成侵犯原有作品作者的著作权，将导致游戏开发者面临高额索赔。

2023 年，美国加州的三名漫画家对 Stability AI 等三家 AIGC 公司发起诉讼，指控 Stability AI 使用 Stable Diffusion 模型开发的付费 AI 图像生成工具构成版权侵权。此外，据外媒报道，2023 年 7 月，一些使用 AI 技术制作的游戏在 Steam 平台上被禁止发行，原因是开发者没有获得所有必要的版权。对此，Steam 平台回应称，“由于 AI 的合法著作权及拥有权存在灰色地带，除非开发者能够确认拥有 AI 训练资料库的使用权及拥有权，否则无法发布这些用 AI 生成的游戏。”由此可见，游戏素材的版权侵权问题如无法获得

[3] 《信息网络传播权保护条例》第二十三条 网络服务提供者为服务对象提供搜索或者链接服务，在接到权利人的通知书后，根据本条例规定断开与侵权的作品、表演、录音录像制品的链接的，不承担赔偿责任；但是，明知或者应知所链接的作品、表演、录音录像制品侵权的，应当承担共同侵权责任。

有效解决，势必将导致 AI 技术赋能游戏产业的实效大打折扣。

（三）游戏内容存在违法违规信息

- 《生成式人工智能服务管理暂行办法》第四条规定：“提供和使用生成式人工智能服务，应当遵守法律、行政法规，尊重社会公德和伦理道德，遵守以下规定：（一）坚持社会主义核心价值观，不得生成煽动颠覆国家政权、推翻社会主义制度，危害国家安全和利益、损害国家形象，煽动分裂国家、破坏国家统一和社会稳定，宣扬恐怖主义、极端主义，宣扬民族仇恨、民族歧视，暴力、淫秽色情，以及虚假有害信息等法律、行政法规禁止的内容；（二）在算法设计、训练数据选择、模型生成和优化、提供服务等过程中，采取有效措施防止产生民族、信仰、国别、地域、性别、年龄、职业、健康等歧视；（三）尊重知识产权、商业道德，保守商业秘密，不得利用算法、数据、平台等优势，实施垄断和不正当竞争行为；（四）尊重他人合法权益，不得危害他人身心健康，不得侵害他人肖像权、名誉权、荣誉权、隐私权和个人信息权益；（五）基于服务类型特点，采取有效措施，提升生成式人工智能服务的透明度，提高生成内容的准确性和可靠性。”
- 《网络信息内容生态治理规定》第六条规定：“网络信息内容生产者不得制作、复制、发布含有下列内容的违法信息：（一）反对宪法所确定的基本原则的；（二）危害国家安全，泄露国家秘密，颠覆国家政权，破坏国家统一的；（三）损害国家荣誉和利益的；（四）歪曲、丑化、亵渎、否定英雄烈士事迹和精神，以侮辱、诽谤或者其他方式侵害英雄烈士的姓名、肖像、名誉、荣誉的；（五）宣扬恐怖主义、极端主义或者煽动实施恐怖活动、极端主义活动的；（六）煽动民族仇恨、民族歧视，破坏民族团结的；（七）破坏国家宗教政策，宣扬邪教和封建迷信的；（八）散布谣言，扰乱经济秩序和社会秩序的；（九）散布淫秽、色情、赌博、暴力、凶杀、恐怖或者教唆犯罪的；（十）侮辱或者诽谤他人，侵害他人名誉、隐私和其他合法权益的；（十一）法律、行政法规禁止的其他内容。”
- 《网络安全法》第十三条规定：“国家支持研究开发有利于未成年人健康成长的网络产品和服务，依法惩治利用网络从事危害未成年人身心健康的活动，为未成年人提供安全、健康的网络环境。”

AI 模型经过训练后，对一些概念具备了比较稳定的“认知”，围绕相关概念的生成内容通常表现出惊人的一致性。一旦 AI 模型的训练数据里包含违规违法内容或偏见、歧

视等有害信息，而算法又未能对其进行阻拦或纠正，则 AI 生成的游戏内容里可能也会包含违规违法及有害内容，这将导致游戏产品不能取得版号上架运营，或者出现被下架停运的后果。

（四）游戏玩家的个人信息保护问题

目前游戏公司发力 to C 业务的一大方向是利用 AI 技术学习玩家行为，为玩家提供定制化个性游戏服务。例如游戏厂商通过收集玩家的游戏时长、过往游戏经验及习惯等行为数据，分析总结玩家偏好，为其针对性地推荐个性化游戏产品，比如向热衷于交友的玩家推荐社交属性较强的游戏、向热衷于建设制造的玩家推荐偏生活类的游戏。这一过程会涉及到对用户个人信息的采集和自动化决策处理[4]，需要严格遵守个人信息保护的有关规定：

- 《网络安全法》第四十一条规定：“网络运营者收集、使用个人信息，应当遵循合法、正当、必要的原则，公开收集、使用规则，明示收集、使用信息的目的、方式和范围，并经被收集者同意。网络运营者不得收集与其提供的服务无关的个人信息，不得违反法律、行政法规的规定和双方的约定收集、使用个人信息，并应当依照法律、行政法规的规定和与用户的约定，处理其保存的个人信息。”
- 《个人信息保护法》第二十四条第二款：“通过自动化决策方式向个人进行信息推送、商业营销，应当同时提供不针对其个人特征的选项，或者向个人提供便捷的拒绝方式。”

此外，《生成式人工智能服务管理暂行办法》第十条[5]要求生成式人工智能服务提供者应当采取有效措施防范未成年人用户过度依赖或者沉迷生成式人工智能服务。目前，在未成年人防沉迷保护领域，已有不少游戏厂商通过收集玩家行为数据（例如收集玩家在终端设备操作游戏时形成的点击压力和半径、加速度方向、重力方向以及玩家在游戏中通过在线语音方式互动产生的语音数据等），利用 AI 技术进行用户画像，精准识别未成年玩家，从而实施严格的防沉迷措施。上述收集动作因涉及到未成年人的信息，因此除了需满足个人信息保护的一般性要求外，还应当符合《个人信息保护法》的以下特别规定：

- 《个人信息保护法》第三十一条：“个人信息处理者处理不满十四周岁未成年人个人信息的，应当取得未成年人的父母或者其他监护人的同意。”

再比如，目前市面上很多款游戏中都有捏脸系统，玩家可以上传一张自然人照片，利

[4] 根据《个人信息保护法》第七十三条规定，自动化决策，是指通过计算机程序自动分析、评估个人的行为习惯、兴趣爱好或者经济、健康、信用状况等，并进行决策的活动。

[5] 《生成式人工智能服务管理暂行办法》第十条 提供者应当明确并公开其服务的适用人群、场合、用途，指导使用者科学理性认识和依法使用生成式人工智能技术，采取有效措施防范未成年人用户过度依赖或者沉迷生成式人工智能服务。

用 AI 技术识图捏脸，创建其在游戏里的“分身”角色。这一过程涉及到个人敏感信息[6]的使用和处理，存在更大的安全风险。若玩家上传的是自己的照片，则该信息一旦泄露并被非法使用，将导致用户人格尊严受到侵害或者人身财产安全受到危害。若玩家上传的是他人的照片，则可能存在侵犯他人肖像权和隐私权等问题。因此，对于收集和处理个人敏感信息的，需要遵守更为严格的规定：

- 《个人信息保护法》第二十八条第二款：“只有在具有特定的目的和充分的必要性，并采取严格保护措施的情形下，个人信息处理者方可处理敏感个人信息。”
- 《个人信息保护法》第二十九条：“处理敏感个人信息应当取得个人的单独同意；法律、行政法规规定处理敏感个人信息应当取得书面同意的，从其规定。”
- 《个人信息保护法》第三十条：“个人信息处理者处理敏感个人信息的，除本法第十七条第一款规定的事项外，还应当向个人告知处理敏感个人信息的必要性以及对个人权益的影响；依照本法规定可以不向个人告知的除外。”

三、合规策略和建议

针对上述合规风险，建议游戏开发企业重视并执行以下合规策略：

1、训练数据的获取和采集应当遵循合规方式。通过公开渠道爬取数据的，应当严格遵守 Robots 协议。向第三方采购数据的，应当对数据来源的合法性进行必要审查，包括事先核查供应商资质，对其业务范围、履约能力、数据与网络安全体系建设情况等进行核查，核查第三方提供的数据权属是否清晰、完整，是否拥有数据所有者的合法授权。此外，游戏开发企业在与供应商订立采购合同时，可以在条款中要求供应商对数据来源的合法合规性作出陈述保证，并约定违约赔偿机制。

2、利用素材生成游戏内容前，对于已知的受版权保护内容，提前取得著作权人的合法授权。应当与著作权人订立许可使用合同，约定训练素材许可使用的范围、期间、权利种类，尤其是素材是否可以被直接用于游戏内容开发，以避免游戏开发完成公布后触发版权侵权风险。

3、通过人工和技术手段加强对游戏内容侵权和违法违规信息的审查力度。除了对 AI 训练时输入的素材进行筛选外，游戏开发完成后也需要对游戏内容输出的合规性进行多轮测试，使其遵守相关内容要求。目前已有游戏厂商利用 AI 图像识别、文本分类、语音识别等技术开展游戏环境监测和净化行动，防范玩家上传不合规的图片或信息，因此，游戏

[6] 《个人信息保护法》第二十八条　敏感个人信息是一旦泄露或者非法使用，容易导致自然人的人格尊严受到侵害或者人身、财产安全受到危害的个人信息，包括生物识别、宗教信仰、特定身份、医疗健康、金融账户、行踪轨迹等信息，以及不满十四周岁未成年人的个人信息。只有在具有特定的目的和充分的必要性，并采取严格保护措施的情形下，个人信息处理者方可处理敏感个人信息。

企业可以借助 AI 技术的强大计算能力，通过不断优化词库、数据库等方式及其他技术手段，在海量的 AI 生成内容中快速识别和发现涉嫌侵权和违规的信息，当然，这需要游戏企业加大对算法和训练数据标注技术的研发投入。不过，现阶段还是应当把 AI 定义为“辅助工具”，人工介入进行游戏内容的过滤和筛选也是相当有必要的。

4、涉及对游戏玩家个人信息的使用和处理的，应当严格执行《个人信息保护法》的规定。包括，在获取玩家个人信息前通过游戏平台《用户协议》《隐私政策》以及弹窗等形式向玩家明示，充分告知玩家处理其个人信息的范围、处理方式等事项，取得玩家的授权和同意；对于通过自动化决策方式向玩家进行个性化游戏推荐的，应赋予玩家拒绝及关闭使用这一技术的权限；此外，就人脸等个人敏感信息的收集和处理，需满足特定目的和充分必要性的要求，取得个人单独同意并采取更为严格的保护措施，例如将个人身份信息与生物识别信息分别存储等；针对涉及未满十四周岁未成年人个人信息的使用和处理的，还需额外取得其父母或监护人的同意。

AI 原生应用相关法律问题研究之四：AI+ 虚拟数字人的法律问题

唐丽子 孙及 贾潇寒

近年来，随着 AI 技术快速发展，虚拟数字人行业也进入了新的发展阶段。AI 技术可覆盖虚拟数字人的建模、视频生成、驱动等全流程，一方面使虚拟数字人的制作成本降低、制作周期缩短，另一方面，多模态 AI 技术使得虚拟数字人的交互能力更上一个台阶。可以说，在 AI 技术的赋能下，虚拟数字人同时拥有“好看的皮囊”和“有趣的灵魂”将不再是设想。本文将介绍 AI 技术在虚拟数字人行业的应用及背后的法律合规问题。

一、AI+ 虚拟数字人的应用场景和发展趋势

相较于传统的真人驱动型虚拟数字人（指基于真人的动作、表情等通过动作捕捉设备驱动的虚拟数字人），AI 虚拟数字人是采用人工智能技术和仿真技术驱动生成的数字化虚拟人物，能够通过深度学习模型、神经网络渲染、自然语言处理等技术，具备灵敏且稳定的感知、表达甚至学习技能等自动交互能力，其生成过程依靠大量的数据采集与处理，并可通过深度学习等技术实现自主学习和调整。

目前，AI 虚拟数字人在多行业、多场景均有落地应用，在具体应用场景上大致可分为两类：服务型虚拟数字人和身份型虚拟数字人。服务型虚拟数字人定位于功能性，主要用于替代真人服务，代表性应用包括数字员工、数字客服、虚拟关怀师、虚拟陪伴助手等；身份型虚拟数字人定位于身份属性，主要用于娱乐或社交活动，代表性应用包括虚拟偶像、虚拟主播、虚拟数字分身等。

AI 虚拟数字人作为数字化服务的一种新形态，在全球经济下行压力加大的形势下，能够为企业提供更高效、便捷和个性化的客户服务，助力企业降本增效、推动产业数智化发展。根据 iiMedia Research（艾媒咨询）的一项数据显示，2022 年中国虚拟数字人核心市场规模为 120.8 亿元，同比增长 94.2%，预计 2025 年将达 480.6 亿元，在传统行业数字化转型及降本增效的需求推动下，中国 AI 虚拟数字人业务需求将进一步释放，预计市场规模将持续增长。

二、AI+ 虚拟数字人需关注的法律合规要点问题

在不同的应用场景下，AI 虚拟数字人将牵涉到复杂多样的法律问题，本文将择以下几个方面重点讨论可能涉及的合规问题：

（一）AI 虚拟数字人的侵权风险

该问题主要包含人格权侵权和著作权侵权两个方面：

1. 人格权侵权问题

以真人原型生成的 AI 虚拟数字人（例如以某央视主持人为原型的 AI 虚拟主持人“小小撒”，因演唱周杰伦歌曲大放异彩的龚俊数字人等）容易引发人格权领域的侵权风险。根据《民法典》，人格权是指民事主体享有的生命权、身体权、健康权、姓名权、名称权、肖像权、名誉权、荣誉权、隐私权等权利[1]。

以真人原型生成的虚拟数字人，其外形能够被识别对应到真人原型身上，若虚拟数字人采用了真人原型的姓名、艺名或昵称，或者使用了真人原型独具特色的音色进行口语表达或唱歌表演，在未获真人原型充分授权的情况下，可能涉嫌肖像权、姓名权以及自然人声音权益等一般人格权益的侵权，引发索赔问题。此前已有多个电商直播间被爆出使用明星 AI 换脸直播，涉嫌侵权。此外，以真人原型生成的虚拟数字人若在与观众互动过程中存在言行不当，也将导致其真人原型的社会评价降低，进而导致侵害真人原型的名誉权[2]。

以某公众人物 H 与某人工智能科技有限公司（“ A 公司”）网络侵权责任纠纷一案为例。A 公司运营一款陪伴智能 AI 聊天软件，用户可发送文字、图片等进行用户与智能 AI 机器人的人机互动，同时该软件为用户提供角色调教功能，用户可自行创设或添加“AI 陪伴者”。A 公司未经 H 授权，将用户制作上传的以 H 姓名、肖像为标识的虚拟形象推送至“AI 陪伴者”，并通过算法应用，将该角色开放给众多用户，允许用户调教该“AI 陪伴者”。北京互联网法院经审理认为，虽然涉嫌侵权的具体图文由用户上传，但 A 公司的产品设计和对算法的应用实际上鼓励、组织了用户的上传行为，直接决定了软件核心功能的实现，该公司不再只是中立的技术服务提供者，而应作为内容服务提供者承担侵权责任。A 公司未经同意使用 H 的姓名、肖像，设定涉及 H 人格自由和人格尊严的调教功能，构成了对 H 姓名权、肖像权、一般人格权的侵害，判令被告承担赔礼道歉、赔偿原告经济损失及精神损害抚慰金的民事责任。

2. 著作权侵权问题

除了人格权侵权问题外，AI 虚拟数字人的外观形象设计、进行的直播和歌舞表演以及通过生成式人工智能技术“创作”出的文本、绘画和音乐，均可能涉及到大量文本、绘

---

[1] 《民法典》第 1018 条规定：“自然人享有肖像权，有权依法制作、使用、公开或者许可他人使用自己的肖像。肖像是通过影像、雕塑、绘画等方式在一定载体上所反映的特定自然人可以被识别的外部形象。”第 1012 条规定：“自然人享有姓名权，有权依法决定、使用、变更或者许可他人使用自己的姓名，但是不得违背公序良俗。”第 1023 条第 2 款规定：“对自然人声音的保护，参照适用肖像权保护的有关规定。”

[2] 根据《民法典》第 1024 条规定：“民事主体享有名誉权。任何组织或者个人不得以侮辱、诽谤等方式侵害他人的名誉权。名誉是对民事主体的品德、声望、才能、信用等的社会评价。”

画和音乐等素材的使用，其中隐藏着著作权侵权风险。例如 2022 年某卫视跨年晚会上出现的邓丽君数字人，若未获得相关歌曲版权授权就进行表演，可能存在著作权侵权问题。

针对以上人格权和著作权侵权问题，运营方若计划采购 AI 虚拟数字人服务，需要考虑：AI 技术方所提供的训练数据来源是哪里？相关的人格权、著作权的授权链条是否清晰、完整？是否需要单独取得真人原型和著作权人的二次授权？要尽可能核实素材库内容本身的合规性，避免与 AI 技术方一起成为侵权行为的共同责任主体。

（二）AI 虚拟主播的直播言行失范问题

利用 AI 虚拟主播开展网络直播活动是目前 AI 技术赋能虚拟数字人行业的一大应用场景，在此类活动中需要注意， AI 虚拟主播虽然只是虚拟数字人，也不能在直播活动中肆无忌惮、随心所欲。国外曾有某虚拟偶像团体成员因言行不当问题被封禁，导致运营方遭受了巨额经济损失。因此，AI 虚拟主播直播活动的参与方，包括运营方、AI 技术提供方、直播平台运营方等，都要充分重视对 AI 虚拟主播言行内容的审核和把控，避免直播“翻车”事故并引发连带责任。

2022 年 6 月，国家广播电视总局、文化和旅游部联合发布《网络主播行为规范》，对网络主播提供表演及视听服务、参与网络营销活动等从业行为作出了规范性要求[3]，其

---

[3] 《网络主播行为规范》第十四条 网络主播在提供网络表演及视听节目服务过程中不得出现下列行为：
1. 发布违反宪法所确定的基本原则及违反国家法律法规的内容；
2. 发布颠覆国家政权，危害国家统一、主权和领土完整，危害国家安全，泄露国家秘密，损害国家尊严、荣誉和利益的内容；
3. 发布削弱、歪曲、否定中国共产党的领导、社会主义制度和改革开放的内容；
4. 发布诋毁民族优秀文化传统，煽动民族仇恨、民族歧视，歪曲民族历史或者民族历史人物，伤害民族感情、破坏民族团结，或者侵害民族风俗、习惯的内容；
5. 违反国家宗教政策，在非宗教场所开展宗教活动，宣扬宗教极端主义、邪教等内容；
6. 恶搞、诋毁、歪曲或者以不当方式展现中华优秀传统文化、革命文化、社会主义先进文化；
7. 恶搞、歪曲、丑化、亵渎、否定英雄烈士和模范人物的事迹和精神；
8. 使用换脸等深度伪造技术对党和国家领导人、英雄烈士、党史、历史等进行伪造、篡改；
9. 损害人民军队、警察、法官等特定职业、群体的公众形象；
10. 宣扬基于种族、国籍、地域、性别、职业、身心缺陷等理由的歧视；
11. 宣扬淫秽、赌博、吸毒，渲染暴力、血腥、恐怖、传销、诈骗，教唆犯罪或者传授犯罪方法，暴露侦查手段，展示枪支、管制刀具；
12. 编造、故意传播虚假恐怖信息、虚假险情、疫情、灾情、警情，扰乱社会治安和公共秩序，破坏社会稳定；
13. 展现过度的惊悚恐怖、生理痛苦、精神歇斯底里，造成强烈感官、精神刺激并可致人身心不适的画面、台词、音乐及音效等；
14. 侮辱、诽谤他人或者散布他人隐私，侵害他人合法权益；
15. 未经授权使用他人拥有著作权的作品；
16. 对社会热点和敏感问题进行炒作或者蓄意制造舆论“热点”；
17. 炒作绯闻、丑闻、劣迹，传播格调低下的内容，宣扬违背社会主义核心价值观、违反公序良俗的内容；
18. 服饰妆容、语言行为、直播间布景等展现带有性暗示、性挑逗的内容；
19. 介绍或者展示自杀、自残、暴力血腥、高危动作和其他易引发未成年人模仿的危险行为，表现吸烟、酗酒等诱导未成年人不良嗜好的内容；
20. 利用未成年人或未成年人角色进行非广告类的商业宣传、表演或作为噱头获取商业或不正当利益，指引错误价值观、人生观和道德观的内容；
21. 宣扬封建迷信文化习俗和思想、违反科学常识等内容；
22. 破坏生态环境，展示虐待动物，捕杀、食用国家保护类动物等内容；
23. 铺张浪费粮食，展示假吃、催吐、暴饮暴食等，或其他易造成不良饮食消费、食物浪费示范的内容；
24. 引导用户低俗互动，组织煽动粉丝互撕谩骂、拉踩引战、造谣攻击，实施网络暴力；
25. 营销假冒伪劣、侵犯知识产权或不符合保障人身、财产安全要求的商品，虚构或者篡改交易、关注度、浏览量、点赞量等数据流量造假；
26. 夸张宣传误导消费者，通过虚假承诺诱骗消费者，使用绝对化用语，未经许可直播销售专营、专卖物品等违反广告相关法律法规的；
27. 通过“弹幕”、直播间名称、公告、语音等传播虚假、骚扰广告；
28. 通过有组织炒作、雇佣水军刷礼物、宣传“刷礼物抽奖”等手段，暗示、诱惑、鼓励用户大额“打赏”，引诱未成年用户“打赏”或以虚假身份信息“打赏”；
29. 在涉及国家安全、公共安全，影响社会正常生产、生活秩序，影响他人正常生活、侵犯他人隐私等场所和其他法律法规禁止的场所拍摄或播出；
30. 展示或炒作大量奢侈品、珠宝、纸币等资产，展示无节制奢靡生活，贬低低收入群体的炫富行为；
31. 法律法规禁止的以及其他对网络表演、网络视听生态造成不良影响的行为。

中要求网络主播应该树立良好形象，营造积极向上、健康有序、和谐清朗的网络空间，不得发布违法违规内容，宣扬淫秽、暴力、恐怖、教唆犯罪、炒作丑闻等违反公序良俗的内容等。与此同时，该规范还规定“利用人工智能技术合成的虚拟主播及内容，参照本行为规范。”明确将 AI 虚拟主播的直播行为纳入到监管范畴中。

针对 AI 虚拟主播开展直播带货活动的，还需要遵守《电子商务法》《广告法》等电商领域的相关法规。以《广告法》为例，AI 虚拟主播的直播过程需要遵守包括但不限于：不得使用“国家级”“最高级”“最佳”“唯一”“独一无二”“全网首发”“最好”“销量第一”“最先进”“首选”等极限用语；推荐产品时，不得贬低其他商品或者服务；不得使用虚假或引人误解的言语进行产品推荐。从内容角度而言，尤其注意不得出现不当政治言论，不得宣扬基于种族、国籍、地域、性别、职业、身心缺陷等理由的歧视；不得利用未成年人角色进行非广告类的商业宣传、表演或作为噱头获取商业或不正当利益，指引错误价值观、人生观和道德观的内容；不得展示或炒作大量奢侈品、珠宝、纸币等资产，展示无节制奢靡生活，贬低低收入群体的炫富行为等。

客观来说，不管是对运营方、AI 技术提供方，或者是直播平台运营方，对 AI 虚拟主播的直播言行做合规审查，其难度相较真人主播来说要更大一些，主要原因在于，AI 虚拟主播与直播观众互动时所产生的内容，除了提前输入的脚本外，还包含了以生成式人工智能技术为基础的 AI 产物，受限于 AI 技术所依托的海量数据和生成结果的“黑盒效应”，品牌方很难全面审查数据内容的合规性。其次，基于 AI 虚拟主播直播时与观众实时互动的特点，实际上不存在对虚拟主播输出内容进行审查的空档时间。因此，相较于真人主播而言，AI 虚拟主播的直播内容不可控风险更高，直播“翻车”风险更大，这也导致了目前部分直播平台为了规避 AI 驱动的不稳定性，限制或完全禁止“AI 驱动型”虚拟主播进行直播，例如：抖音平台在《抖音关于人工智能生成内容的平台规范暨行业倡议》中阐明其对于“AI 驱动型”虚拟主播的态度：使用已注册的虚拟人形象进行直播时，必须由真人驱动进行实时互动，不允许完全由人工智能驱动进行互动。

（三）AI 虚拟数字人的商业代言风险

通过运营 AI 虚拟数字人孵化粉丝群体，进而开展商业品牌代言活动，目前是虚拟数字人的一大应用场景，典型案例比如 2021 年 5 月横空出世并一夜爆火的 AI 虚拟数字人 AYAYI，凭借着又美又飒的形象人设，在亮相一个月后就有商业邀约找上门来。目前，AYAYI 已经与一众美妆、汽车、时尚等品牌达成合作，并作为某日资美妆品牌官宣的数字代言人，为其旗下新品进行宣传。

利用 AI 虚拟数字人开展商业代言活动，运营方需要谨慎对待选品问题，比如由虚拟数字人代言美妆产品可能会引发一定争议，虚拟数字人本身没有实体，却要强调产品上脸服帖、滋润不拔干等功效，容易让消费者产生虚假宣传的感觉，引发品牌方的处罚风险。

此外，由于 AI 虚拟数字人并不满足我国《广告法》项下的广告代言人定义[4]，以真人原型生成的 AI 虚拟数字人在直播带货过程中，若消费者基于明星艺人等公众人物的影响力和信任在直播间消费购买商品的，可能被认定为构成明星本人的广告代言行为[5]。但如前所述，AI 虚拟数字人在直播过程中输出的内容可能是基于 AI 技术实时生成或由品牌方或运营方事先输入脚本的内容，并非完全是明星艺人本人的真实意思表示，若出现代言风险，此时在合同层面根据广告内容生成方式和来源明确约定最终的责任承担主体就会显得格外重要。

（四）AI 虚拟数字员工场景下需考虑从业资质等问题

将 AI 虚拟数字人作为员工投放到金融、房地产、教育等行业也是目前的一大热潮。2021 年，某房地产企业的首位虚拟数字员工“崔筱盼”甚至还因为催办单据核销率高而获得集团总部年度优秀新人奖，一时间引发热议。

在此类应用场景下，企业要谨慎考虑虚拟数字员工的投放场景，例如若将虚拟数字人作为理财顾问、保险咨询师、私人教练等角色面向客户，可能要考虑是否具备从业资质的问题。此外，还需关注其提供的服务可能属于《互联网信息服务算法推荐管理规定》第二条规定的利用个性化推送类、排序精选类、检索过滤类等算法技术向用户提供信息的应用算法推荐技术，服务提供者应承担信息安全管理、用户权益保护等责任。

（五）直播平台的安全评估和显著标识义务

根据《网络直播营销管理办法（试行）》第十三条，“直播营销平台应当加强新技术新应用新功能上线和使用管理，对利用人工智能、数字视觉、虚拟现实、语音合成等技术展示的虚拟形象从事网络直播营销的，应当按照有关规定进行安全评估，并以显著方式予以标识。”因此，对于平台内存在 AI 虚拟主播直播活动的，直播平台运营方需要注意履行安全评估和显著标识义务。

如前所述，虚拟数字人直播带货活动可能引发明星本人的代言风险，因此从运营方和艺人经纪方的角度来说，对 AI 虚拟数字人的直播带货活动进行显著标识也是运营方和艺人方面区分真人带货，规避广告代言风险的有效措施之一。此外，在技术成熟的情况下，

---

[4]《广告法》第二条第五款规定，“本法所称广告代言人，是指广告主以外的，在广告中以自己的名义或者形象对商品、服务作推荐、证明的自然人、法人或者其他组织。”

[5]《市场监管总局、中央网信办、文化和旅游部、广电总局、银保监会、证监会、国家电影局关于进一步规范明星广告代言活动的指导意见》规定，明星艺人等在商业广告中通过形象展示、语言、文字、动作等对商品或者服务进行推荐或者证明，应当依法认定为广告代言行为。

真人原型与 AI 虚拟数字人之间的“真假难辨”，可能会涉及到《反不正当竞争法》项下的经营者混淆[6]行为，此时若采用显著方式予以标注，也有助于相关方避免不正当竞争风险。

三、合规建议

针对上述问题，我们提示“AI+ 虚拟数字人”的相关参与方考虑以下合规建议：

1. 核查训练数据来源，取得完整授权，避免侵权风险。运营方在决定使用 AI 虚拟数字人服务前，应当向 AI 技术提供方了解虚拟数字人的创作背景情况，核查 AI 技术方的模型数据来源，以及相关授权链是否清晰、完整。若存在真人原型，是否取得真人原型的肖像权授权；对于已知的受著作权保护内容，需要与著作权人签订著作权授权使用协议，获得著作权授权。

2. 做好对 AI 虚拟数字人的言行内容的审查。AI 技术提供方在为客户打造 AI 虚拟数字人时，除了侧重虚拟数字人的商业营销属性外，也要坚持科技伦理先行，剔除训练数据中的违法违规内容及歧视、偏见等不良信息，强化模型技术处理的精细程度，避免 AI 虚拟数字人在直播互动过程中生成输出不当内容；运营方在直播前要做好对直播脚本、直播间背景、直播音乐、广告内容等的事先审核，此外，在 AI 生成技术目前尚不完全稳定的情况下，可考虑配备真人助播，实时监控跟进 AI 虚拟主播与观众的互动过程，及时发现问题并反馈技术方进行模型调整与优化；直播平台运营方需加强巡视力度，及时发现 AI 虚拟主播直播过程出现的违规问题并第一时间进行纠正和处理。

3. 相关方的业务合作合同里需要约定清楚侵权责任划分权责机制。运营方在与 AI 技术提供方的业务合作合同中应关注对于交付的 AI 虚拟数字人服务的侵权责任划分、训练数据合法合规性、授权来源合法性等内容是否进行了合理约定。对 AI 虚拟数字人与用户互动过程中基于 AI 技术不成熟原因生成的不当内容，可参考类似于明星代言合同中对明星不当言行的约束条款，约定由技术提供方承担一定赔偿责任。此外，对于 MCN 机构而言，如需对知名艺人形象进行虚拟数字人开发并对外授权的，建议更新相关经纪协议内容，避免超出原有的经纪范围；对于艺人而言，需要特别考虑以本人形象开发虚拟数字人进行直播带货的模式下，艺人可能被认定为广告代言人，如相关广告被认定为虚假广告的，艺人可能会被要求承担连带责任，因此，建议艺人与合作方明确限定本人形象的使用范围，设定负面清单并建立单次使用特别同意制度。

4. 谨慎考虑数字员工的投放场景和资质要求，依法履行合规义务。为了避免触发从

---

[6] 《反不正当竞争法》第六条规定，“经营者不得实施下列混淆行为，引人误认为是他人商品或者与他人存在特定联系：……（二）擅自使用他人有一定影响的企业名称（包括简称、字号等）、社会组织名称（包括简称等）、姓名（包括笔名、艺名、译名等）；（四）其他足以引人误认为是他人商品或者与他人存在特定联系的混淆行为。”

业资质问题以及因 AI 技术不稳定造成输出内容“翻车”问题，企业应谨慎考虑在创造性、核心性岗位上投放数字员工，暂时将数字员工的使用场景集中在类似客服等简单、重复性工种上，从运营安全性角度来说更为稳妥一些。此外，运营方还需注意落实《互联网信息服务算法推荐管理规定》等 AI 监管领域法律法规所要求的信息安全管理、用户权益保护等合规责任。

5. 直播平台应做好安全评估和显著标识工作。除了直播平台外，品牌方和艺人经纪方也需关注直播过程是否已做了显著标识，以尽可能规避艺人广告代言风险和不正当竞争风险。

相由 AI 生：浅谈深度伪造（Deepfake）与个人形象权

宋海燕 林德鑫

2023 年 2 月，一段由 AI 生成的知名播客主持人乔·罗根（Joe Rogan）推销男性补品的深度伪造（deepfake）广告在某社交平台上疯传[1]。这段广告使公众难以分辨名人形象的真伪，最终被平台删除。未经授权的深度伪造视频开辟了一个新的法律灰色地带——名人难以保护其肖像权，其对个人形象商业化的控制也受到影响。随着 AI 技术的革新，生成人脸的技术难度和成本大幅降低，侵害他人肖像权的行为层出不穷，而法律法规滞后于技术变革，个人形象权的保护面临着法律上的挑战。

在人工智能时代，娱乐产业对于个人形象权的保护也存在着模糊地带。2023 年 1 月，全球第一部运用深度伪造技术实现 AI 合成名人面孔的节目——《深度伪造邻居之战》（Deep Fake Neighbor Wars）在英国电视公司 ITV 的流媒体平台上线[2]。在这部没有明星实际出演的喜剧中，一群在替身演员基础上由 AI 合成的“名人”成为了邻居，在一起插科打诨。这类 AI 合成媒体（synthetic media）对名人肖像的使用给法律带来了新问题：面孔被 AI 合成的名人是否有权控制其个人形象并由此获得报酬？名人能否主张肖像权被侵害？节目制片方是否需要承担侵权责任？政府对于这类新兴内容是否有监管要求？

本文将通过对个人形象权的比较法研究，分析美国、中国关于深度伪造（deepfake）的法律规制、司法实践，以探讨人工智能技术与个人形象权的边界。

一、深度伪造（Deepfake）概念

Deepfake（深度伪造）一词由“deep learning”（深度学习）和“fake”（伪造）组合而成。深度伪造作为一种人工智能人像生成技术，主要包含以下内容：

（1）面部替换：将人像覆盖至既有图片或视频上，实现面部合成以假乱真。

（2）面部重演：操纵视频中人像的面部特征，扭曲一个人的表情，包括口型、眉毛、眼睛的运动和头部的倾斜。

（3）人脸生成：基于训练数据生成全新的人脸图像。

---

[1] https://mashable.com/article/joe-rogan-tiktok-deepfake-ad, last visited on May 4, 2023。
[2] https://www.itv.com/presscentre/press-releases/deep-fake-neighbour-wars, last visited on April 28, 2023。

Deepfake 一词起源于 2017 年。一名叫做“deepfakes”的 Reddit 平台用户在该平台发布 AI 合成的名人换脸视频。此后这种技术被称为 deepfake。深度伪造主要基于生成对抗网络（Generative Adversarial Network，GAN），这是一种人工智能深度学习模型，其工作原理是由两个神经网络相互对抗：第一个用于生成图像，第二个用于判别该输出是否真实。该模型经过交替优化训练后，生成的深度伪造内容可以与真人形象无异，足以骗过人的眼睛。

二、美国法上的个人形象权与涉及 deepfake 的法律

个人形象权（the right of publicity）是指个人有权控制代表其个体特征且具有商业价值的形象要素，并由此获得报酬的权利。其中，代表个体特征形象的要素（persona）包括：个人姓名、声音、照片、形象、签名以及其他可区分个体特征的内容等[3]。未经授权擅自商业性使用个人的名称、肖像、声音或其他具有个人特征的要素，可能构成对个人形象权的侵犯[4]。

（一）美国对个人形象权的法律规定

美国法以各州立法及判例法的形式形成了个人形象权的法律保护体系。大部分州制定了关于个人形象权的法律法规。

《加利福尼亚州民法》第 3344 章及第 3344.1 章节（California Civil Code Section 3344 and Section 3344.1，也被称为“加利福尼亚州个人形象权法（California Rights of Publicity Statute）”）将个人形象权视为一种财产权利，即将当事人的个人特征形象（包括肖像、姓名、声音、签名及相似特征）进行商业性使用并由此获得报酬的权利。由于未经同意的换脸色情视频泛滥，许多名人的形象出现于深度伪造视频中，被用于牟利目的。2019 年，加利福尼亚州颁布了打击“深度伪造”的立法（Assembly Bill 602, AB602, California Civil Code Section 1708.86.），旨在打击在加利福尼亚州未经授权的 deepfake 色情视频（使用数字或电子技术描绘个人的露骨材料），以救济非自愿色情制品的受害者，使得其有权寻求禁令救济并得到补偿性和惩罚性损害赔偿及律师费[5]。

纽约州《纽约公民权利法案》（New York Civil Rights Law）第 50 条及第 51 条规定了个人形象权——未经当事人事先书面同意，禁止他人为广告、商业用途使用其姓名、肖像、照片或声音[6]。2020 年修订时新增了针对 deepfake 视频的条款。其中，第 52-c 条规定，若个人形象未经同意被数字化（digitization）合成于色情内容中，受害者有权

---

[3] 宋海燕：《娱乐法》（第二版），商务印书馆 2018 年版，第 209 页。
[4] 宋海燕：《娱乐法》（第二版），商务印书馆 2018 年版，第 183 页。
[5] AB602, California Civil Code Section 1708.86. https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=201920200AB602, last visited on April 28, 2023.
[6] New York Civil Rights Law Section 50 and Section 51.

因 deepfake 视频的传播而获得“非法传播或发布对个人的色情描绘（sexually explicit depiction）的私人诉讼权”，且有权获得禁令救济、补偿性和惩罚性损害赔偿以及律师费[7]。

同时，个人形象未经授权被深度伪造技术用于商业用途的问题也引起了美国联邦监管部门的关注。2023 年 3 月 20 日，美国联邦贸易委员会（FTC）发布了一篇博客文章。该文章指出：鉴于名人深度伪造视频愈发频繁，FTC 建议开发或提供生成式人工智能产品的公司以及广告商考虑 AI 合成媒体带来的欺诈风险。若上述公司通过深度伪造误导消费者，其可能会面临 FTC 的执法行动[8]。

为保护个人形象权，避免深度伪造技术的不良影响，各大社交平台也先后修改其平台规则及用户政策，有条件地规制 deepfake 内容[9]。例如，2023 年 3 月 21 日，TikTok 在更新的平台社区准则中规定，禁止针对个人形象的 deepfake 视频，公众人物除外；同时，更新后的平台准则还禁止未经公众人物同意的产品代言或其他违反平台政策的涉公众人物 deepfake 合成视频。新准则同时规定，所有展示真实场景的 deepfake 视频都必须明确披露其为合成（synthetic）、虚假（fake or not real）或受更改（altered）[10]。

（二）美国涉 deepfake 的相关案例

1. 凯兰德·杨诉 NeoCortext 公司案（Young v. NeoCortext, Inc.）[11]

2023 年 4 月 3 日，美国真人秀节目《老大哥（Big Brother）》出身的名人凯兰德·杨（Kyland Young）向美国加利福尼亚中区联邦地区法院（C.D.Cal）起诉 NeoCortext 公司，主张该公司的换脸应用程序 Reface 违反了加利福尼亚州个人形象权法。被告所开发的 Reface 应用是一款深度伪造（deepfake）应用程序，其运作模式为：（1）在该应用的免费版本中，用户能够从“预设图库（Pre-set catalogue）”中选择希望换脸的名人。Reface 会扫描用户上传的本人图片或视频并以此为基础生成与预设图库中选中的明星、运动员等名人交换面孔后的新的图片或视频。免费版 Reface 生成的换脸新图片会有明显的、不可移除的“Reface 应用制作”水印。（2）免费版 Reface 有选项可引导用户跳转至收费的专业版（Pro Version）Reface，按月收取订阅费或者一次性的终生订阅费。用户付费之后使用 Reface 生成的换脸新图片或视频不再显示免费版中的水印。原告杨称被

---

[7] https://www.dwt.com/blogs/media-law-monitor/2021/05/new-york-right-of-publicity-pornographic-deepfakes, last visited on April 28, 2023。

[8] https://www.ftc.gov/business-guidance/blog/2023/03/chatbots-deepfakes-voice-clones-ai-deception-sale, last visited on April 28, 2023。

[9] 2020 年 1 月，Facebook 宣布禁止 deepfake 内容，涉及讽刺和模仿的内容除外；2020 年 2 月，Twitter 发布政策，将删除具有欺骗性及影响公共安全的 deepfake 内容。

[10] https://www.theverge.com/2023/3/21/23648099/tiktok-content-moderation-rules-deepfakes-ai, last visited on April 28, 2023。

[11] Young v. NeoCortext, Inc. § 2:23-CV-02496, https://www.classaction.org/news/big-brother-star-cries-foul-over-alleged-reface-app-publicity-rights-violations, last visited on April 28, 2023。

告 NeoCortext 公司在未经其同意的情况下“商业利用”其肖像、图像、姓名和声音，通过以下两种方式推广 Reface 应用程序的付费订阅：（1）允许用户付费移除合成图片上的水印；以及（2）用水印作为免费广告以吸引新用户。除了禁止 NeoCortext 公司将他和其他集体成员的“姓名、声音、签名、照片或肖像用于商业目的”的禁止令救济外，杨还寻求金钱赔偿（即“每个集体成员的实际损害赔偿或每次违规 750 美元”加上合理的诉讼费用和律师费），以及对拟议的集体诉讼的认证。此案尚在审理中。

2. 汉密尔顿诉斯贝特案（Hamilton v. Speight）[12]

美国法院在保护个人形象权时，还需要平衡其与美国宪法第一修正案规定的言论自由权之间的关系。2019 年至 2020 年，美国宾夕法尼亚州东区联邦法院以及联邦第三巡回上诉法院在汉密尔顿诉斯贝特案中均通过“转换性使用（transformative use）”的判断平衡了个人形象权与言论自由权之间的关系。此案是一起与深度伪造技术类似的涉及人工智能合成游戏人物的诉讼。本案中，原告兰伍德·汉密尔顿（Lenwood Hamilton）是一名前职业摔角手和橄榄球运动员、艺人以及励志演说家，其起诉涉案视频游戏的设计者莱斯特·斯贝特（Lester Speight）及涉案游戏的制作方、出品方（统称“被告”）使用其肖像作为游戏中的一个角色，故侵犯了其个人形象权。被告则辩称，其作品应享有第一修正案关于言论自由的保护。宾夕法尼亚州东区联邦法院认为本案中被告的言论自由权胜过原告汉密尔顿的个人形象权，因为涉案游戏中的角色是对原告创造并表演的“硬石汉密尔顿（Hard Rock Hamilton）”这个角色的转换性使用，并于 2019 年作出有利于被告的简易判决（summary judgment）[13]。原告汉密尔顿对该判决不服，上诉至美国联邦第三巡回上诉法院。

美国联邦第三巡回上诉法院认为，尽管被告视频游戏中的角色形象与原告具有相似的面部特征、肤色、发型、身材、声音、运动能力和服装等元素，但也存在其他显著的差异，原告最多只是该游戏角色形象的合成原材料之一；并且被告的游戏角色有独特的设计和故事（例如涉案游戏角色是在一个虚拟星球上参加一场虚拟的战争，而不是像原告一样在地球上从事职业摔角运动），“以至于它主要已成为被告自己的表达（own expression）”。因此法院认为被告对原告肖像的使用构成“转换性使用”，由此于 2020 年驳回了原告的请求，判定被告不侵权[14]。汉密尔顿对该判决不服，诉至美国最高法院，于 2021 年被驳回其对美国联邦第三巡回上诉法院的调卷令（writ of certiorari）申请（即

[12] Hamilton v. Speight, 827 Fed. Appx. 238 (3d Cir. 2020).
[13] Hamilton v. Speight, 413 F. Supp. 3d 423 (E.D. Pa. 2019).
[14] https://news.bloomberglaw.com/us-law-week/reputation-management-and-the-growing-threat-of-deepfakes, last visited on April 28, 2023。

美国最高法院拒绝重审该案件）[15]。

三、中国法涉及深度伪造的侵权风险与监管要求

（一）中国关于“深度伪造”的相关法律法规

中国针对深度伪造的法律规定主要涉及肖像权、互联网法规以及个人信息等领域。首先，《民法典》第 1019 条规定：“任何组织或者个人不得以丑化、污损，或者利用信息技术手段伪造等方式侵害他人的肖像权。未经肖像权人同意，不得制作、使用、公开肖像权人的肖像，但是法律另有规定的除外。”根据该规定，即使没有营利目的和主观恶意，未经本人同意的深度伪造行为仍可能侵害肖像权。

同时，深度伪造技术需要获取大量人脸图像中的特征数据，而人脸信息属于法律予以特殊保护的生物识别信息，使用人脸信息要获得肖像权人的单独授权。2022 年 12 月发布的《互联网信息服务深度合成管理规定》（“《管理规定》”）、2023 年 4 月发布的《生成式人工智能服务管理办法（征求意见稿）》[16]（“《征求意见稿》”）等部门规章及规范性文件，对侵犯他人肖像权的深度伪造内容提出了不同程度的规制要求。

譬如，《管理规定》定义的“深度合成技术”包括人脸生成、人脸替换、人物属性编辑、人脸操控、姿态操控等生成或者编辑图像、视频内容中生物特征的技术[17]。《管理规定》对深度伪造提出较有针对性的规制。首先，因 deepfake 涉及人脸等生物识别信息的显著编辑功能，《管理规定》第 14 条第 2 款规定，deepfake 服务提供者和技术支持者应当提示服务使用者依法告知人脸被编辑的个人（即肖像权人），并取得其单独同意[18]。此外，为避免 deepfake 可能导致公众混淆或者误认，《管理规定》第 17 条第 1 款第 3 项规定，服务提供者应当在生成或者编辑的信息内容的合理位置、区域进行显著标识，向公众提示深度合成情况[19]。

另，《征求意见稿》中的部分规定针对深度伪造也进一步加以了规制。譬如第 4 条及第 18 条明确规定，人工智能生成内容应当尊重他人合法利益，不得损害肖像权、名誉权和个人隐私[20]。第 13 条规定，服务提供者应当建立用户投诉接收处理机制；在发现知悉生成的内容侵害他人肖像权时，应当采取措施，停止生成[21]。

---

[15] Hamilton v. Speight, 2021 U.S. LEXIS 3166 (U.S., June 21, 2021).
[16] 2023 年 7 月，国家网信办联合国家发展改革委、教育部、科技部、工业和信息化部、公安部、广电总局正式公布《生成式人工智能服务管理暂行办法》，并自 2023 年 8 月 15 日起施行。
[17] 《互联网信息服务深度合成管理规定》第 23 条。
[18] 《互联网信息服务深度合成管理规定》第 14 条第 2 款。
[19] 《互联网信息服务深度合成管理规定》第 17 条第 1 款第 3 项。
[20] 《生成式人工智能服务管理办法（征求意见稿）》第 4 条、第 18 条。
[21] 《生成式人工智能服务管理办法（征求意见稿）》第 13 条。

（二）涉及深度伪造的中国案例

1. 古风汉服网红魏某诉“AI 换脸”App 系列案件 [22]

2022 年 8 月，成都铁路运输第一法院审理了一批使用“AI 换脸”App 程序侵害他人肖像权的一审案件。原告魏某是一名古风汉服网红，在短视频平台上发布古风视频。2022 年 7 月，魏某分别起诉四家运营 AI 换脸软件的公司，认为被告在原告未授权的情况下，在其运营的 App 中上传包含原告肖像的视频作品，生成 AI 换脸视频，侵犯了其肖像权。

一审法院认为：（1）肖像权属于自然人的人格权，未经本人同意，不得擅自使用他人肖像；（2）被告经营的软件以“换脸”为主要功能和卖点，利用基于他人形象拍摄制作的素材引流，诱导用户点击下载、付费。被告的行为具有营利目的，不属于合理使用范畴；（3）涉案软件通过 AI 换脸技术替换了原视频中的自然人形象，供其用户使用，属于《民法典》第 1019 条规定的“以利用信息技术伪造的方式”侵害他人肖像权的行为，侵犯肖像权人的人格尊严。综上，成都铁路运输第一法院在四起案件中作出判决，支持原告主张（因被告已删除涉案视频），判令被告侵权成立。

2. 中国其他类似案例

2022 年 12 月，杭州互联网法院也审理了一起类似的“AI 换脸”App 程序利用深度伪造技术侵害他人肖像权的案件 [23]。

该案中，杭州互联网法院分析了深度伪造视频中肖像的可识别性，认定 AI 合成的视频与特定自然人之间能够建立对应联系。法院认为，原告楼某某为古风汉服模特，经常发布古风汉服照片和视频，对案涉模板视频及替换合成后视频中所载对应形象的人物肖像均享有肖像权。对于换脸后的视频，楼某某仅留存身体形象，但对比原视频素材，普通人仍能通过未被修改的相应场景和细节识别出身体形象对应主体为楼某某。因此，法院认为被告利用深度合成技术将其他用户提供的人脸替换至涉案视频中，并生成形象逼真的伪造肖像视频的行为侵害了原告楼某某的肖像权。最终，杭州互联网法院判决 App 开发者赔礼道歉并赔偿损失人民币 5000 元，双方当事人均服判息诉，判决已生效。

上述案例表明，国内各地法院对 deepfake 的侵权认定及判赔标准大致统一。开发运营者利用 deepfake 技术侵害他人肖像权，在删除侵权视频后，需承担赔礼道歉、赔偿损失的责任。赔偿金额根据涉案视频肖像的商业价值、涉案侵权视频数量、被告的过错程度等因素综合酌定。

---

[22] 成都铁路运输第一法院（2022）川 7101 民初 5502 号、5472 号、6350 号、6349 号民事判决书。

[23] 《杭州互联网法院判决一公司使用 APP 提供“换脸”服务侵害他人肖像权》，《人民法院报》2022 年 12 月 15 日，第 3 版，https://www.chinacourt.org/article/detail/2022/12/id/7065781.shtml。

结语

当生成式 AI 的发展使得深度伪造模糊了真与假的界限之时，如何保护个人形象权给法律带来了挑战。深度伪造技术的商业应用必须获得肖像权人的明确同意，并显著标识。技术应当有边界，平台应当及时删除并停止生成违反法律法规的深度伪造内容。

为促进人工智能技术的合法和有益使用，深度伪造技术的负面影响应当受到法律的规制，以更好地应对 AIGC 时代的挑战。随着 deepfake 的潜在问题及法律风险日益受到重视，未来，法律规制必将愈加明确与严格，我们将对此保持关注。

感谢律师王默，实习生陈玮聪对本文作出的贡献。

论图片生成式 AIGC 平台在侵权纠纷中的角色与责任边界

瞿淼 钱琪欣 焦晨恩

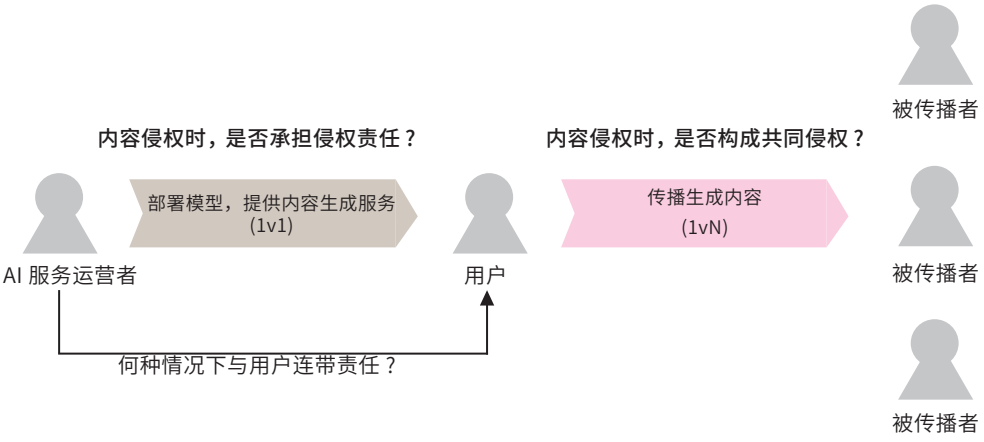
作为人工智能技术的前沿领域，生成式人工智能（“AIGC”）目前已经成为引发广泛热议的科技话题。AIGC 技术的发展速度令人惊叹，其应用场景也愈发丰富多元。在高效促进经济社会发展的同时，AIGC 技术的发展亦将不断产生如何用现有制度调整新型网络服务的问题。

本文以图片生成式 AIGC 平台为例，从平台责任的视角，根据服务中不同阶段下平台的法律角色（ICP 还是 ISP）及侵权责任边界进行分析，并据此为 AIGC 平台的产品设计及运营提供风险防范建议。

图片生成式 AIGC 服务过程分为内容生成和内容传播两个阶段：

1. 阶段 1—生成交付阶段（1 v 1）：AIGC 平台在服务器端生成内容（“AIGC”）并通过网络交付给用户；

2. 阶段 2—AIGC 使用阶段（1 v N）：用户在接收 AIGC 后进行使用。



如图所示，两阶段的参与主体、行为和影响范围均有所不同，因此不同阶段下 AI 服务运营者的角色与责任边界需要逐一分析。

一、阶段 1（内容生成与交付）下 AIGC 平台的责任边界

在该阶段，用户向 AI 工具发出指令，AI 工具根据用户指令生成内容并发送至用户端展示或供用户下载，其运营者在网络侵权案件中的角色偏向于内容服务提供者（ICP），而非技术服务提供者（ISP）。

（一）AIGC 平台可能被界定为 ICP

ISP 的界定和避风港原则在我国法律框架中最早适用于处理著作权领域的信息网络传播权侵权纠纷之中，随后被侵权责任法以及民法典吸收，被广泛适用于各种类型的网络侵权纠纷中。到目前为止，《信息网络传播权保护条例》是最为清晰地列举了典型 ISP 类型的法规。此外，虽然《民法典》并没有给出 ISP 的明确定义，但最高院在《民法典理解与适用》系列关于第 1194 条的释义中指出：第 1194 条的“网络服务提供者”应包括网络技术服务提供者和网络内容服务提供者：网络技术服务提供者不直接向用户提供信息，其只是提供通道或平台，本身并不对传输或存储的信息进行主动编辑、组织或者修改，全部内容都是由网络用户提供；网络内容服务提供者自身直接向网络用户提供内容或产品服务，其提供的内容和产品是该网络服务提供者自己主动编辑、组织、修改或提供的。

AIGC 工具生成图片过程中，用户发出指令后，AIGC 系由服务提供者控制的服务器生成并根据服务提供者预设的条件决定是否、以及何时将生成物交付给用户。即 AI 工具参与了 AIGC 的编辑和修改，因此我们更倾向于 AIGC 平台可能被认定为内容服务提供者而不当然受到避风港原则的保护。

- **AI 工具对生成内容不可控不影响提供者被界定为 ICP**

在区分网络服务提供者类型时，应着重考察网络服务提供者对于内容的影响力和决定权。虽然现有的 AIGC 原理决定了生成内容具有天然的不可控性，但 AIGC 平台实际上仍然可以通过选择合适的训练素材、增加过滤机制等手段影响和改变向用户交付的生成结果。

因此，我们认为，虽然 AI 工具的生成内容具有难以消除的不可控性，但并不能因此否认 AI 工具对生成物的影响和贡献，更不代表服务提供者没有能力或权利改变和决定向用户交付的内容。因此，AI 工具（即便是对内容控制力最弱的 AIGC 工具）很可能仍被界定在 ICP 的范畴内。

- **AI 能力的来源不影响提供者被界定为 ICP**

目前大量 AIGC 服务系基于第三方来源的模型提供。相比于自研模型提供服务的运营者，以第三方模型提供服务的运营者更加难以从技术上控制 AIGC 的内容。

对此，我们认为，正如产品生产者也需要为其选择、集成的第三方零件承担产品质量责任，在 AIGC 服务场景下，网络服务提供者对于如何选择、使用第三方能力具有自主决定和控制能力，也应当承担因其选择、管理 AI 能力不当而引发的法律责任。因此，AIGC 平台的 AI 能力来源并不影响其对外承担侵权、违法责任时的角色定位。并且，如果 AIGC 服务系运营者基于开源大模型进行二次训练形成的 AIGC 的技术能力，其作为训练者和提供者更应对其提供的工具承担相应责任。

（二）生成、交付 AIGC 行为的侵权定性思路

根据目前国内法律法规对 ICP 的内容审查和监管要求，在明确 AIGC 平台属性和角色更偏向于 ICP 的前提下，AIGC 平台应当对其提供内容的真实性和合法性承担责任。我们以网络侵权的领域内最为常见的两种诉由，网络著作权侵权和网络肖像权侵权为例来分析生成、交付 AIGC 这一服务过程的行为性质。

- **网络著作权侵权：**将与他人作品实质性相似的生成物生成并交付给用户可能侵犯他人的复制权、改编权或者信息网络传播权等著作权。在该情况下，由于著作权侵权遵循无过错原则，侵权定性难以避免。但是，根据《著作权法》第 59 条规定的原理和精神，AIGC 平台在证明自身不具有过错或过错程度较低的情况下，仍有机会免除或降低赔偿责任。
- **网络肖像权侵权：**肖像权侵权的认定应遵循过错责任原则，但实践中未经授权的使用行为往往即视为存在过错。因此，AIGC 平台对内容生成后交付前是否采取过滤和拦截措施影响某个特定侵权行为是否存在，或者是否采取了防范措施降低侵权的可能性，将主要影响最终侵权责任的比例、大小，而非侵权定性。

当然，在特定侵权争议中，AIGC 服务的提供者虽然与生成物的产出密切关联，但由于其交付给用户的动作往往不是公开可识别和固定的，且该阶段 AIGC 的传播范围主要限制在用户本人，因此在 AIGC 平台不能知晓或预见用户后续传播或使用范围，甚至并不知悉后续传播或使用方式的情况下，该等纠纷实际发生并导致 AIGC 平台承担赔偿责任的可能性仍然较低。

二、阶段 2（用户使用、传播 AIGC）下 AIGC 平台的责任边界

在该阶段，用户既可能将 AIGC 直接发布于 AIGC 平台内进行传播，亦可能将其发布于第三方平台或线下场景。在这两种不同传播场景下，AIGC 平台的角色定位和责任边界有所不同。

（一）用户将 AIGC 存储、发布于 AIGC 平台进行传播

如果根据产品的商业模式设计，用户随后将涉嫌侵权的 AIGC 直接存储或发布于 AIGC 平台，后续通过技术手段，供其他用户获取。在此种情况下，AIGC 平台便兼具有内容生成和内容传播属性，故 AIGC 平台仍有较大可能被认定为属于内容服务提供者。

进一步地，用户和 AIGC 平台对于被控侵权元素的贡献程度（尽管实践中难以界定）更多影响的是两者内部如何分摊侵权责任，而不影响对外共同承担侵权责任。当然，在此过程中仍有较多因素可能影响 AIGC 平台行为定性，例如 AIGC 平台对于用户后续利用方式的了解程度、平台与用户的交互过程、生成工具与发布平台的关联等。

（二）用户在平台之外使用或传播 AIGC

在该情形下，用户自行决定如何使用、传播 AIGC。该等行为引起的侵权纠纷已不符合通常所说的网络用户利用网络服务提供者的服务实施侵权行为的行为范式，故在此场景下 AIGC 平台不会作为网络服务提供者承担责任。但这并不意味着 AIGC 平台在此场景下绝对免责。在权利人知晓侵权内容系来自 AIGC 平台的情况下，权利人完全有可能主张 AIGC 平台与用户构成共同侵权，并要求 AIGC 平台与用户承担连带赔偿责任。

就该等主张能否成立，可能会有三种立场：

1. **AIGC 平台承担连带赔偿责任：**尽管不知晓行为细节，但 AIGC 平台仍可在一定程度上预见用户后续的使用、传播行为的发生及相应的侵权可能性。因而双方对于被控侵权的使用行为已存在共谋，构成共同侵权，且应就全部损失承担连带赔偿责任。
2. **AIGC 平台不构成共同侵权而无需承担连带责任：**AIGC 平台不干预、不决定用户如何使用 AIGC，亦不清楚用户的后续行为方式、数量和范围等，故双方主观上对侵权行为未达成合意，客观上亦未共同实施侵权行为，因此不构成共同侵权。
3. **AIGC 平台应根据过错程度按比例承担连带赔偿：**综合考虑 AIGC 平台在侵权内容的贡献程度、是否采取必要的防范或过滤措施、是否提示用户、对于用户行为的知晓程度等因素，判断 AIGC 平台是否存在过错而构成帮助侵权，进而根据过错程度按比例承担连带赔偿责任。

从根本上来看，造成三种不同结论的关键原因在于三种立场采用了不同的共同侵权认定标准，即以何种标准认定双方是否存在共同侵权的意思联络。但目前司法实践中并没有形成非常统一的意见，均是根据具体案件情况做出个案判断，故而在 AIGC 场景下亦难以

得到统一的结论。

但从结论对行业的影响来看，我们认为，第一种立场会将 AIGC 平台长期置于与无法预测的用户共同侵权并连带赔偿的不确定性下，且该种风险难以在技术上消除和控制，由此可能会对 AIGC 产业造成负面影响；第二种立场则同样过于绝对，可能会导致 AIGC 平台怠于采取预防措施，造成低成本生产的侵权内容泛滥。因此，针对 AIGC 这样的新型网络服务，我们更倾向于较为折衷的第三种立场，根据过错程度确定 AIGC 平台的责任边界，以引导和鼓励 AIGC 平台积极采取预防措施减少侵权内容的出现，更好地实现各方利益平衡。

三、AIGC 平台可采取的风险控制措施

虽然整体上有可能被认定为 ICP 而需要与用户承担帮助或者共同侵权的风险，但考虑到目前 AIGC 平台的属性尚无明确定论，且有效的事前预防措施和对于侵权内容采取相应措施仍然有可能在诉讼中降低赔偿金额。因此，在产品设计过程中，AIGC 平台仍然可以考虑在用户协议中进行提示和告知，明确服务内容、平台的权限等因素，并加强对用户的风险提示。

同时，考虑到生成式 AI 的技术特性，较难从模型本身出发直接消除内容风险。AIGC 平台如果希望进一步控制风险，也可从加强合法数据训练、限制用户输入侧元素或者限制内容输出侧等角度加以控制。

简析 AI 绘画著作权风险及规制建议

刘迎 吴涵 何光远 张浣然

引言

2024 年 2 月 16 日，OpenAI 公司发布视频生成模型 Sora，极大拓展了人工智能在视频生成领域的能力。用户仅需输入简单的文字描述，Sora 即可快速生成时长约一分钟的视频，Sora 的横空出世再次震撼了整个世界，助推生成式人工智能（Artificial Intelligence Generated Content，以下简称“AIGC”）成为社交圈近期广泛讨论的热点话题。

AI 绘画是 AIGC 技术的一项分支应用，具有丰富的商业应用场景和广阔的市场发展空间，是各大 AI 科技公司均在布局的新赛道。目前国外较为流行的 AI 绘画工具主要是 Stable Diffusion 和 Midjourney，国内同样有文心一格、360 智绘、通义万相等类似功能的 AI 绘画工具。根据国泰君安研报的预测数据，未来 5 年，AI 绘画在图像内容生成领域的渗透率将达到 10%-30%，市场规模或将超过 600 亿元[1]。

AIGC 技术的发展速度令人惊叹，但新技术的颠覆性变革也对现有法律制度体系造成冲击。伴随着 AI 绘画市场的迅速扩张，相应的法律问题也层出叠见，比如在著作权法领域，AI 绘画模型未经授权抓取版权图片训练 AI 模型的行为引起了艺术家和图片公司的警惕；在人格权领域，AI 绘画工具强大的“一键换脸”功能被广泛使用并传播，可能涉嫌侵犯人格权的问题。有鉴于此，本文简要介绍 AI 绘画的生成及运行机制，重点分析 AI 绘画可能存在的著作权风险及特点，并据此提供风险规制建议。

一、AI 绘画的生成及运行机制

就 AIGC 而言，《生成式人工智能服务管理暂行办法》（“《AIGC 暂行办法》”）第二条将其定义为“利用生成式人工智能技术向公众提供生成文本、图片、音频、视频等内容的服务。”AI 绘画，首先是 AI 从海量图文对应的数据中学习到了“语言描述”与“艺术画面”的关联。在此基础上，当用户输入一段语言描述，希望创作一幅新的画作时，AI 将调动以上学习到的知识和能力开始创作，经过数百轮不断修正画作，每一轮都会仔

[1] 陈义：《AI 绘画成新晋“流量密码” 产业发展机遇与风险并行》，载《通信信息报》2022 年 12 月 14 日，第 006 版。

细检查草稿与语言描述的一致性，以期让作品与输入的语言描述具有准确的关联。在这个修正的过程中，整体的构图不断明晰，最终形成在审美上与人类经验与知识高度一致的成品[2]。

各类 AI 绘画工具均需借助 AI 算法模型实现，而 AI 算法模型则需要对海量图片进行深度学习，AI 算法模型各有不同且都有优缺点，从微观角度解释 AI 算法模型会稍显晦涩难懂，但是从宏观层面而言，AI 绘画工具的生成及运行过程主要分为图片输入、深度学习和内容输出三个阶段，以下将分阶段介绍 AI 绘画的生成及运行机制。

（一）图片输入

图片输入阶段的主要任务是导入图片并建立模型，AI 算法模型开发者会通过爬虫等工具在互联网上搜集抓取数亿级别的海量图片，并对拟输入的图片进行筛选、分类、标记从而建立深度学习需要的图片数据库，输入图片的数量和质量最终也将影响生成图片的质量。AI 算法模型可分为算法和参数两个部分，AI 算法模型开发者基于不同需要对算法和参数进行调整，在图像识别领域一般使用卷积神经网络或者卷积神经网络图像分类算法模型[3]。AI 算法模型开发者在输入阶段将海量图片作为训练样本“投喂”给 AI 算法模型学习，并通过调整算法和参数进行迭代更新。

（二）深度学习

在深度学习阶段，AI 算法模型创作的机器学习原理类似于人脑的思考过程。AI 算法模型对训练样本图片进行分析处理，总结这些图片的像素值规律，再依据这些规律生成新的图片，AI 算法模型开发者会对生成图进行校验并作出反馈，AI 算法模型则会依据开发者的反馈修正其所总结的规律、再依据这些规律进行新一轮的图片生成，如此循环往复直至 AI 生成的图片满足了开发者设置的标准，此时可认定 AI 算法模型掌握了这一生成规律[4]。

（三）内容输出

在内容输出阶段，AI 绘画使用者会先将个人对图片的需求以提示词（prompt）或参考图的形式输入 AI 绘画工具，同时设置采样迭代步数、提示词相关性等参数，AI 算法模型基于使用者输入的文字搜寻对应的图片，再利用其在深度学习阶段掌握的规律对这些图片进行分析处理、最终生成数张符合要求的图片，AI 绘画使用者可以在其中挑选中意的图片[5]。

---

[2] 侯伟：《AI 作画，版权归谁？》，载微信公众号“中国知识产权报”，2022 年 10 月 25 日。
[3] 彭飞荣：《论算法创作中涉数据的著作权侵权风险及其化解》，载《法律适用》，2023 年第 4 期。
[4] 张赟洁：《AI 绘画生成物的著作权争议研究（一）：AI 绘画的运行机制与争议梳理》，载微信公众号“水木网络法学”，2023 年 12 月 21 日。
[5] 同上注。

二、AI 绘画创作的著作权风险分析

（一）输入阶段

在输入阶段，AI 绘画算法模型训练时需要爬取海量图片“投喂”给算法模型，这些图片难以避免会包括他人享有著作权的作品，甚至是大家耳熟能详的知名卡通形象作品（比如奥特曼、米老鼠、蜘蛛侠等）。根据《著作权法》第三条规定，美术作品、摄影作品及设计图等图形作品均为著作权法的保护对象。结合《著作权法》第十条规定，上述作品的著作权人依法享有发表、复制、改编、信息网络传播等著作权。AI 算法模型输入图片时，需要对图片信息及特征进行数字化处理，从而转化为 AI 算法模型可以理解的标准格式，因此在数据输入阶段无法避免对原图片信息和特征的复制，存在构成侵犯原图片著作权的风险。

根据《AIGC 暂行办法》第七条规定，AIGC 服务提供者应当依法开展预训练、优化训练等训练数据处理活动，应当使用具有合法来源的数据和基础模型，涉及知识产权的，不得侵害他人依法享有的知识产权。因此，AI 算法模型开发者在训练模型阶段需要遵守知识产权合规性要求，尊重他人享有著作权的作品。

关于 AI 算法模型在输入阶段使用他人享有著作权的作品是否构成侵权，国内司法实践暂未出现相关判例，但国内学者对此持有不同观点，比如王利明教授认为“如果生成式人工智能使用相关数据训练模型的时候未经著作权人授权，有可能会侵害著作权[6]”，也有部分学者认为“著作权人是否能够对人工智能学习和训练行为主张权利，在各个国家依然是有争议的问题，这个问题有待法律明确，目前不宜轻易得出该行为是否构成侵权应立即予以禁止的结论。虽然目前我国强调加强著作权的保护，但不能因此排除对出于人工智能学习、训练目的的数据挖掘行为适用合理使用条款的可能性[7]。”

这在其他司法辖区也是如此，例如近一年来美国存在多起相关诉讼，原告包括美国作家协会、Getty Images 和 New York Times 等，被告则为 Open AI、Stability AI 等 AI 算法模型开发者和运营者。但在这一系列案件中，对于合理使用抗辩的主张不尽相同，法院的认定标准是否可能趋同也有待观察。其中，在美国加州联邦法院 Stability AI 集体诉讼[8]一案中，原告 Anderson 指控 Stability 公司直接侵犯著作权的主要依据是：购买数亿受版权保护的图片作品副本并将其用于训练 Stable Diffusion 模型。Anderson 从输出端（output）入手进行举证，在专门查询图片作品是否被 AI 公司扫描的第三方网站 https://haveibeentrained.com 上输入自己的名字进行查询，发现自己的一些（而非所

---

[6] 王利明：《生成式人工智能侵权的法律应对》，载《中国应用法学》，2023 年第 5 期。
[7] 习睿：《又有 AI 产品被质疑侵权，AI 绘画与隐私保护如何平衡？》，载微信公众号“Tech 星球”，2023 年 8 月 18 日。
[8] Andersen v. Stability AI Ltd., 2023 U.S. Dist.

有）登记作品已被用作 AI 训练图像，该证据得到采信。法官最终认定 Stability 公司获取图片副本以训练 Stable Diffusion 模型的行为构成直接侵犯著作权。

（二）学习阶段

在学习阶段，AI 算法模型需要将图片语言通过机器翻译转化为机器语言，如此方可将海量图片数据输入预先设定的算法模型进行训练。该阶段学习的图片以输入阶段的图片数据为基础，如果输入阶段的图片存在著作权侵权风险，则学习阶段当然也构成侵权。但输入阶段不存在侵权的情况下，学习阶段对图片的处理是否同样会引起著作权风险？

根据学习阶段的 AI 算法模型原理，输入的图片在 AI 学习过程中会被无数次的复制、模拟、再复制，在此过程中图片数据只是一种暂存状态。在 Cartoon Network LP,LLLP v. CSC Holdings, Inc. 案中，美国联邦第二巡回上诉法院认为，尽管被告服务器缓存中确实呈现出原告的作品，但是持续时间非常短（该案中为 1.2 秒），因此不符合美国《版权法》中对侵犯复制权的规定[9]。临时复制本身不符合著作权法意义上复制的特性，将临时复制纳入复制权范畴，将不合理地扩大著作权人的权利，不利于他人对信息的获取。因此，著作权法并不承认临时复制是著作权法上的复制行为。换言之，在 AI 学习过程中，临时复制并不构成对复制权的侵犯。但非临时复制符合侵害复制权的构成要件，构成侵权。

（三）输出阶段

在输出阶段，AI 绘画工具根据使用者的指令输出成果，此类 AI 绘画是否属于著作权法意义上的“作品”，目前学术界对于该问题的争议较大。有学者认为 AI 绘画生成物具有独创性，是著作权法意义上的“作品”[10]。也有学者认为现阶段的人工智能属于弱人工智能，机器学习等均是以一定算法或模型为基础，由此产生的生成物不是著作权法上的作品[11]。近日北京互联网法院 AI 绘画著作权侵权纠纷作出一审判决，该案成为我国 AI 生成图片著作权侵权第一案。关于是否构成“作品”的问题，北京互联网法院在（2023）京 0491 民初 11279 号案件中重点论述了 AI 绘画图片属于“智力成果”以及具有“独创性”两个要件，并据此认为 AI 绘画图片构成作品，应当受到著作权法保护。

如果认可 AI 绘画生成物是作品，那么该作品的权利归属需要进一步明确，以便确定侵权时的责任主体。学者关于 AI 绘画生成物的权利归属如何认定主要存在以下几种不同意见：认为是归 AI 算法模型开发者或 AI 绘画运营者，或是归 AI 绘画使用者，或者归 AI 本身。著作权制度的立法初衷之一是保护自然人的利益，通过法定方式赋予作者一定时间内享有专有权，填补作者的付出，以达到促进文化发展与传播的目的。AI 算法模型或计

---

[9] 何隽：《大数据知识产权保护与立法：挑战与应对》，载《中国发明与专利》，2018 年第 3 期。
[10] 吴汉东：《人工智能时代的制度安排与法律规制》，载《法律科学（西北政法大学学报）》，2017 年第 5 期。
[11] 王迁：《论人工智能生成的内容在著作权法中的定性》，载《法律科学（西北政法大学学报）》，2017 年第 5 期。

算机仅仅是人类的附属物，不具有独立的法律人格。如果认为 AI 绘画生成物是著作权法意义上的作品，其权利归属应当是自然人，既是当下社会环境的诉求，也是自然秩序的要求[12]。北京互联网法院在（2023）京 0491 民初 11279 号案件中同样认为人工智能模型设计者仅是创作工具的生产者，AI 绘画使用者是直接根据需要对涉案人工智能模型进行相关设置，并最终选定涉案图片的人，涉案图片是基于 AI 绘画使用者的智力投入直接产生，且体现出了 AI 绘画使用者的个性化表达，故 AI 绘画使用者享有涉案图片的著作权。

如果 AI 绘画生成物与输入阶段用以训练的图片作品构成实质性相似，则该等 AI 绘画生成物可能构成对原作品的著作权侵害。输出内容存在侵权时 AI 算法模型开发者或 AI 绘画运营者是否应当承担责任。目前中国司法实践已有相关案例出现，广州互联网法院在（2024）粤 0192 民初 113 号案件中认为 AI 绘画运营者生成包含奥特曼版权元素的内容侵犯了奥特曼原作的复制权和改编权，并判令案涉 AI 绘画运营者承担停止侵权及赔偿损失等民事责任。在美国加州联邦法院 Stability AI 集体诉讼[13] 一案中，美国法院以原告 Anderson 未能证明 AI 绘画生成物与原告 Anderson 作品相似的仿冒内容这一事实驳回了原告 Anderson 的该项诉讼请求。

三、AI 绘画的著作权侵权特点

AI 绘画的生成数量和创作速度均远远超过艺术家的平均水平，因此 AI 绘画涉嫌侵权的作品数量庞大。与传统的著作权侵权行为相比，AI 绘画的著作权侵权存在以下特点：

（一）侵权行为认定困难

对于著作权侵权行为的认定，司法实践中一般采用“接触 + 实质性相似”原则，该原则具体指只有证明涉嫌侵权作品与受著作权保护的作品构成实质相似，同时作品权利人又有证据表明被告在此前具备了接触原作品的机会或者已实际接触了原作品，才能判定为著作权侵权。但是 AI 绘画对“接触 + 实质性相似”原则提出了如下挑战：

对“接触”要件而言，AI 绘画使用者只是输入了文字指令，并未接触在先受著作权保护的作品。事实上“接触”在先作品的是 AI 绘画算法模型，这就涉及对“接触”要件的理解和判断，AI 绘画算法模型对作品的接触可否视为 AI 绘画使用者对绘画的接触。

对“实质性相似”要件而言，AI 绘画生成物对每部作品色彩、线条等要素的使用量较低，呈现零散和碎片化的特点。AI 绘画生成物可能同时侵权多个著作权人的多部作品，但侵权内容对每部作品而言所占的数量和比例均较低，这就导致“实质性相似”的判断标准难以明确。对原作品著作权人维权而言，AI 算法模型训练时抓取的图片数量巨大，提取的

[12] 刘友华、魏远山：《机器学习的著作权侵权问题及其解决》，载《华东政法大学学报》，2019 年第 2 期。
[13] 同注释 8。

图片元素维度多种多样，如何对成千上万张 AI 绘画生成物与自身作品的相似度进行认定，对于著作权人显然有较大难度。

（二）侵权的责任主体认定及责任划分困难

AI 算法模型本身不是法律主体，也无法承担相应法律责任。AI 算法模型开发者、AI 绘画运营者及 AI 绘画使用者，参与 AI 绘画的开发与使用。如上文所述，AI 绘画生成物的产出通常要经过内容输入阶段、 机器学习阶段和内容输出阶段 , 三个阶段都面临着一定的侵犯著作权的风险，可能涉及到 AI 算法模型开发者对画作数据抓取程序及软件自我学习的设计，也可能涉及到 AI 绘画运营方对画作数据库的管理，也包含 AI 绘画用户自身思想与情绪的表达。原画著作权人维权时应该向哪些主体主张权利以及上述主体的责任如何划分，同样值得讨论。

对于 AI 算法模型开发者而言，其无法控制 AI 绘画使用者的使用目的，是用于个人学习欣赏还是其他商业目的，这完全取决于 AI 绘画使用者。在无法证明 AI 算法模型开发者存在明显过错的情况下，如果要求其承担侵权责任，则会极大伤害开发者的研发热情，不利于整个 AI 绘画行业的未来发展。

对于 AI 绘画运营者而言，虽然广州互联网法院在（2024）粤 0192 民初 113 号案件中已判令 AI 绘画运营者承担侵权责任，但同时广州互联网法院在判决书中也强调了“不宜过度加重 AIGC 服务提供者的义务”，呼吁建立一个平衡、包容、兼容创新与保护的中国式人工智能治理体系。

对于 AI 绘画使用者而言，如果认为其享有 AI 绘画生成物的著作权，则基于权利义务相一致的原则，其有可能需要对该等作品的侵权问题承担责任。但如果 AI 算法模型在训练阶段就存在较大侵权隐患，使用者只是将 AI 绘画作为一种创作工具时，由使用者承担侵权责任也有失公允。

因此，一旦发生 AI 绘画生成物侵权事件，该由哪个主体承担侵权责任以及如何划分责任比例及大小将会成为一个新的难题，后续有待司法实践的案例进一步明确上述问题。

四、AI 绘画的著作权风险规制建议

基于以上对 AI 绘画著作权风险与侵权特点的分析，为有效防范 AI 绘画创作在各个阶段的著作权风险，我们提供以下规制建议，供 AI 绘画行业的相关参与主体及监管部门参考：

（一）在内容输入阶段，面对 AI 绘画在前期的数据收集阶段可能存在的著作权侵权

风险，AI 算法模型开发者可以在现有法律制度下寻求并采用合法获取数据的方法，尽可能避免因未经著作权人许可使用其权利作品而引起侵权的情形。但是，面对大量的图片数据，AI 算法模型开发者很难识别哪些图片属于作品，作品的真实权利人以及权利人是否有权授权该图片等，后续需要整个行业共同解决如何创建更为便捷的授权方式。尽管理论上可以通过著作权集体管理组织获取授权，但训练所需图片超乎寻常的量级及其对应授权规模，将为开发者带来难以负荷的授权成本。此外，在未建立延伸性集体管理制度的情况下，同样会面临多数作品未纳入集体管理的授权障碍。

就这一实践困境而言，国家有关部门也可以考虑在现有制度上进行创新，为 AI 算法模型训练学习阶段的作品使用设置合法性依据。目前业内讨论较多的主要有两种思路：一是在著作权法中增设 AI 算法模型训练学习的合理使用例外或数据挖掘例外。二是设置法定许可制度来优化 AI 绘画中图片数据的使用授权机制，即 AI 算法模型开发者可以不事先获得作品权利人的许可直接使用作品，仅需向权利人支付合理报酬。

此外，参照近日全国网络安全标准化技术委员会正式发布的《生成式人工智能服务安全基本要求》，我们理解 AI 算法模型开发者在输入阶段可能有必要遵循进一步拓宽的防止侵犯著作权的注意义务，包括但不限于：1）设置语料及生成内容的知识产权负责人，并建立知识产权管理策略；2）建立对语料中的主要知识产权侵权风险进行识别的机制等。

（二）对于 AI 绘画运营者而言，参考广州互联网法院在（2024）粤 0192 民初 113 号案件中的观点，其作为 AIGC 服务提供者，在内容输出时应积极履行合理注意义务，主要包括：1）应按照《AIGC 暂行办法》第十五条规定，建立投诉举报机制，使得权利人可以通过投诉举报机制来保护其著作权。2）应按照《AIGC 暂行办法》第四条规定，通过服务协议等方式提醒 / 教育用户尊重他人知识产权，并在充分考量自身技术特性与权益诉求的情况下，参考目前通行的行业实践，向用户告知生成内容的知识产权相关风险，并就知识产权侵权风险的责任与义务进行约定。

（三）对于 AI 绘画使用者而言，为减少侵权发生的几率，使用者向 AI 绘画输出指令时应尽量回避与知名艺术家或知名作品相关的提示词。如果使用者指令 AI 绘画模仿知名画家的风格或者相关作品时，应特别重视 AI 绘画生成物与该画家作品是否构成“实质性相似”的比对，以降低侵权风险。

AIGC：合规引领探索之路

张逸瑞 冯宝宝 朱佳蔚 张津豪

引言

从 GPT 3.5 的问世、GPT4.0 的革新到 Open AI 最近推出的 Sora 多模态 AI 生成软件，生成式人工智能（AIGC）在全球范围内的奇点时刻似乎愈来愈近。在中国，AIGC 的应用也已经深入到金融、医疗、交通等多个关键行业。然而，随着 AIGC 技术的广泛应用，其法律和伦理问题也日益凸显。为了规范这一新兴领域，我国出台了若干相关法律和规定，其中的针对性法规包括《互联网信息服务算法推荐管理规定》[1]（“《算法推荐规定》”）、《互联网信息服务深度合成管理规定》[2]和《生成式人工智能服务管理暂行办法》[3]（“《AIGC 暂行办法》”）。这些规定不仅为 AIGC 的应用提供了法律框架，也为行业的合规操作提出了具体要求。

对于生成式人工智能平台（“AIGC 平台”）而言，合规运营不仅保护平台自身的权益，也有助于维护消费者的权益，同时促进技术的健康发展。本文将从资质合规、内部合规管理体系及制度、互联网应用关键条款完善、外部商业合作等方面初步绘制 AIGC 平台合规管理的全景图像，为 AIGC 服务提供者（指《AIGC 暂行办法》中规定的，利用生成式人工智能技术提供生成式人工智能服务的组织、个人，“AIGC 服务提供者”）在探索快速发展的 AIGC 领域提供合规操作指南。

一、资质合规

AIGC 服务提供者作为互联网信息服务提供者，应当根据《互联网信息服务管理办法》[4]和《中华人民共和国电信条例》[5]，申请办理 B25 类信息服务业务的增值电信业务经营许可证（“ICP 证”）；在一些服务领域或业务场景内，AIGC 平台还需取得专门的许可证。同时，如 AIGC 服务提供者提供的服务具有舆论属性或者社会动员能力，AIGC 服务提供者在向公众提供服务前，应当进行安全评估，并按照《算法推荐管理规定》履行

[1] 2021.12.31 发布，2022.03.01 实施。
[2] 2022.11.25 发布，2023.01.10 实施。
[3] 2023.07.10 发布，2023.08.15 实施。
[4] 2011.01.08 发布，2011.01.08 实施。
[5] 2016.02.06 发布，2016.02.06 实施。

算法备案手续。具体而言：

（一）互联网信息服务业务经营许可证 / 备案（ICP 证 / 备案）

根据《互联网信息服务管理办法》，互联网信息服务可分为经营性和非经营性两类。经营性互联网信息服务，是指通过互联网向上网用户有偿提供信息或者网页制作等服务活动。非经营性互联网信息服务是指通过互联网向上网用户无偿提供具有公开性、共享性信息的服务活动。国家对经营性互联网信息服务实行许可制度；对非经营性互联网信息服务实行备案制度。未取得许可或者未履行备案手续的，不得从事互联网信息服务。因此，针对需提供经营性互联网信息服务的 AIGC 平台，需取得 B25 类增值电信业务经营许可证（即 ICP 证）；若为提供无偿服务的 AIGC 平台，需申请 ICP 备案。

特别需要注意的是，根据《工业和信息化部关于开展移动互联网应用程序备案工作的通知》[6]，APP 主办者提供互联网信息服务的，应当向省级通信管理局履行备案手续，未履行备案手续的，不得从事 APP 互联网信息服务。因此，对于需要提供 App、小程序等移动互联网信息服务的 AIGC 平台，应当注意及时向省级通信管理局履行备案手续。

（二）在线数据处理与交易处理业务许可证（EDI 证）

根据《中华人民共和国电信条例》，经营增值电信业务需要根据业务覆盖范围是否跨省、自治区、直辖市分别向国务院信息产业主管部门或省、自治区、直辖市电信管理机构申请办理《跨地区增值电信业务经营许可证》或《增值电信业务经营许可证》。在线数据处理与交易处理为增值电信业务的一种，从事该业务需要取得在线数据处理与交易处理业务证（即 EDI 证）。因此，对于需要从事数据处理与交易处理的 AIGC 平台，需取得 EDI 许可证。

（三）公安联网备案

根据《计算机信息网络国际联网安全保护管理办法》[7] 等相关规定，互联单位、接入单位和使用计算机信息网络国际联网的法人或其他组织，都应当自网络正式联通之日起 30 日内，向所在地的省、自治区、直辖市人民政府公安机关指定的受理机关办理备案手续，实践中各网站通常登录“全国公安机关互联网站安全管理服务平台”提交公安联网备案申请。对于 AIGC 平台而言，通常需要通过互联网向用户提供信息服务，因此应当及时进行备案。

（四）算法备案

根据《算法推荐管理规定》《深度合成管理规定》《AIGC 暂行办法》规定，具有舆

[6] 2023.07.21 发布，2023.07.21 实施。
[7] 2011.01.08 发布，2011.01.08 实施。

论属性或者社会动员能力的 AIGC 服务提供者必须就 AIGC 所使用的算法进行算法备案。目前我国法律法规和相关规定中仅明确了“具有舆论属性或社会动员能力的互联网信息服务”（即开办论坛、博客、微博客、聊天室、通讯群组、公众账号、短视频、网络直播、信息分享、小程序等信息服务或者附设相应功能以及开办提供公众舆论表达渠道或者具有发动社会公众从事特定活动能力的其他互联网信息服务），而对于何为具有舆论属性或社会动员能力的算法推荐服务、深度合成服务、生成式人工智能服务则暂时并未给出进一步定义。实务中，对于何为“具有舆论属性或社会动员能力”的判断相对较为宽泛，几乎涵盖了所有具备信息共享功能的服务。因此，除少部分功能简单，并且完全不具备任何信息互通以及用户交流能力的平台以外，其余大部分具备一定信息传递以及用户聚合能力的 AIGC 平台均需由 AIGC 服务提供者进行算法备案。

在算法备案所需提交材料中，尤以《算法安全自评估报告》最为复杂。该报告要求 AIGC 服务提供者提供包括算法风险研判、算法风险防控情况的描述，并要求 AIGC 服务提供者完成包括风险防范机制、用户权益保护体系、内容生态治理体系、模型安全保障机制、数据安全防护体系在内的风险管理制度建设。

（五）安全评估

目前我国多部法律法规和相关规定中均对“具有舆论属性或社会动员能力的互联网信息服务”提出了安全评估的要求。如上文所述，AIGC 平台服务很有可能涉及具有舆论属性或社会动员能力的互联网信息服务，即需要按照《具有舆论属性或社会动员能力的互联网信息服务安全评估规定》通过全国互联网安全管理服务平台完成安全评估。按照《AIGC 暂行办法》等法律法规和相关规定，对于 AIGC 平台服务还需进行新技术新应用安全评估（“双新评估”）。2024 年 3 月 1 日，全国网络安全标准化技术委员会发布了《生成式人工智能服务安全基本要求 (TC260-003)》[8]，作为全国网络安全标准化技术委员会发布的技术文件，该要求为《AIGC 暂行办法》中关于 AIGC 平台安全性的评估提供了细化标准参考以及技术支撑，为 AIGC 服务提供者在现行人工智能治理框架下对自身的服务安全进行评估提供了指引。

（六）《计算机软件著作权登记证书》《App 电子版权证书》或《软件著作权证证书》

根据安卓、苹果等应用市场的 APP 上架规则[9]，服务提供者通常被要求提交《计算机软件著作权登记证书》《App 电子版权证书》《软件著作权认证证书》之一作为著作权归属证明供运营商审核。其中，《计算机软件著作权登记证书》由中国版权保护中心登记

[8] 2023.10.11 发布。

[9] 苹果 APP 上架规则：https://developer.apple.com/cn/app-store/review/guidelines/；华为 APP 上架规则：https://developer.huawei.com/consumer/cn/doc/app/50104。

颁发、《App 电子版权证书》由易版权平台认证颁发、《软件著作权认证证书》由中国版权协会认证颁发。因此，如需在 App 应用上架，建议及时向相应的机构或平台申请著作权归属登记认证。

（七）其他业务所需资质

通常 AIGC 平台业务场景较为广泛，很有可能涉及多个行业的监管，从而需要获得特定行业的相关证照才能够合法运营，例如，在涉及图文、视听节目的情形下，往往还涉及《网络文化经营许可证》《网络出版服务许可证》《信息网络传播视听节目许可证》等行业监管角度的证照。届时应当根据 AIGC 平台具体的业务模式综合判断其是否需办理相关资质证照。

二、内部合规管理体系及制度

（一）科技伦理审查制度

2019 年 7 月，中央全面深化改革委员会第 9 次会议审议通过《国家科技伦理委员会组建方案》。2019 年 10 月，中共中央办公厅、国务院办公厅印发通知，成立国家科技伦理委员会，并先后成立了人工智能、生命科学、医学三个分委员会。2023 年 12 月 1 日，科学技术部、教育部、工业和信息化部等多部门联合发布的《科技伦理审查办法（试行）》[10]（“《科技伦理审查办法》”）正式实施，该办法明确从事生命科学、医学、人工智能等科技活动的单位，研究内容涉及科技伦理敏感领域的，应设立科技伦理（审查）委员会。即，AIGC 服务提供者涉及开展《科技伦理审查办法》适用范围内的科技活动，需自行设立科技伦理（审查）委员会或委托其他单位的科技伦理（审查）委员会对于所涉科技活动进行科技伦理审查 [11]。

（二）AI 内容安全基础审查制度

2024 年 3 月 1 日，全国网络安全标准化技术委员会发布《生成式人工智能服务安全基本要求》，为面向境内公众提供生成式人工智能服务的提供者提高服务安全水平、提供者自行或委托第三方开展安全评估或相关主管部门评判生成式人工智能服务的安全水平提供参考。该文件对于语料安全、模型安全、安全措施、安全评估等提出统一要求。就 AIGC 的整体内容安全而言，AIGC 服务提供者应当履行下述基础义务：

1. AIGC 平台的语料安全合规义务

（1）语料来源安全审查：对语料来源进行管理，在面向特定语料来源进行采集前，应对该来源语料进行安全评估，语料内容中含违法不良信息超过 5% 的，不应采集该来源

[10] 2023.09.07 发布，2023.12.01 实施。
[11] 详见本书《人工智能（AI）：科技伦理治理走起》一文，本文不多做赘述。

语料；面向特定语料来源进行采集后，应对所采集的该来源语料进行核验，含违法不良信息情况超过 5% 的，不应使用该来源语料进行训练；

（2）**语料内容安全审查：**对训练语料内容应进行过滤，采取关键词、分类模型、人工抽检等方式，充分过滤全部语料中违法不良信息；对训练前语料的知识产权侵权情况应进行识别，建立知识产权管理策略，设置语料以及生成内容的知识产权负责人；对包含个人信息的语料，应在充分满足合法使用该个人信息的条件后进行使用；

（3）**语料标注安全审查：**对标注人员，应自行进行考核、职能划分，以及预留充足、合理的标注时间；对标注规则，应至少包括标注方法、质量指标等内容，并对功能性标注和安全性标注分别制定标注规则；为确保标注内容准确性，应对功能性标注和安全性标注设立不同审核通过标准。

2. 针对 AIGC 平台的模型安全要求包括：

（1）**模型备案要求：**AIGC 服务提供者如使用第三方基础模型进行研发，则应当确保该等基础模型已经经过或至少正在申请主管部门备案[12]，而不建议使用未经主管部门备案的基础模型；

（2）**模型生成内容安全要求：**在 AI 的训练过程中，AIGC 服务提供者应将生成内容安全性作为评价生成结果优劣的主要考虑指标之一；每次对话中，应对使用者输入信息进行安全性检测，引导模型生成积极正向内容。

3. 针对 AIGC 平台的安全措施要求包括：

模型适用安全措施要求：AIGC 服务提供者根据服务领域，应充分论证应用必要性、适用性以及安全性；根据服务场合，应具备与风险程度以及场景相适应的保护措施；根据服务人群，应采取不同技术或管理措施。

（三）数据安全制度

结合《AIGC 暂行办法》《中华人民共和国网络安全法》《中华人民共和国数据安全法》《中华人民共和国个人信息保护法》等法律法规，我们梳理了数据输入、模型开发、部署使用以及运营监控四个周期内 AIGC 服务提供者所需履行的数据安全合规义务，AIGC 服务提供者应当依据其提供的服务内容及类型制定有效的数据安全制度。具体而言，

1. 数据输入周期。在该周期内，AIGC 服务提供者将完成训练数据的采集和预处理（包括数据标注、数据清洗等），并最终形成完整的预训练语料库与数据集。该周期内，AIGC 服务提供者所需要遵守的合规义务包括：

[12] 《互联网信息服务算法推荐管理规定》第二十五条规定，国家和省、自治区、直辖市网信部门收到备案人提交的备案材料后，材料齐全的，应当在三十个工作日内予以备案，发放备案编号并进行公示；材料不齐全的，不予备案，并应当在三十个工作日内通知备案人并说明理由。

（1）数据采集合规义务，即使用具有合法来源的数据和基础模型，至少需要涵盖知识产权、个人信息保护、反不正当竞争法等相关的方面。

通常而言，数据来源的类型主要包括自行采集型、数据交易型和开放数据爬取型，其中，自行采集是指通过 APP、传感器等方式直接采集数据，数据交易是指通过合法的交易方式从数据提供方处获取相关数据，开放数据爬取则是指通过数据爬虫等方式从第三方获取开放的数据。

根据数据来源的类型不同，确保数据来源合法性的常规操作亦有所不同。在自行采集型与数据交易型中，AIGC 服务提供者并非数据的原始采集人，无法直接了解该等数据的原始采集过程，因此其保证数据来源合法性的重点在于（i）确保取得相关数据权利主体的授权，建立数据许可规范；（ii）在有关数据供应协议或软件许可协议中要求供应商 / 许可方对其提供的数据 / 模型不侵犯第三方权利做出陈述与保证 （具体详见下文第四部分：外部商业合作）；以及（iii）在对外宣传或用户协议中注明其所使用的数据或基础模型的来源方，并进行免责声明。而当 AIGC 服务提供者以开放数据爬取的方式收集数据时，则应当重点关注数据爬虫行为本身是否满足上述合法性要求，例如不得违反 Robots 协议和网站公开的声明 / 协议等文件、不得干扰网站的正常运行等。

总体而言，建议 AIGC 服务提供者针对数据采集确立一套合规流程及风险评估标准，以确保数据收集的合法合规性。

（2）数据处理合规义务，即依法进行数据清洗、数据标注等数据处理活动，主要内容包括：

（a）涉及个人信息的，应当取得个人同意或者符合法律、行政法规规定的其他情形；

（b）采取有效措施提高训练数据质量，增强训练数据的真实性、准确性、客观性、多样性，包括但不限于制定并执行数据筛选规范、数据清洗规范等数据质量优化工具；针对数据标注，制定可操作的标注规则，包括但不限于前文所述语料标注规则。

2. 模型开发周期。该周期为 AI 模型的核心周期。在该周期内，AIGC 服务提供者将完成 AI 模型的训练，基于 AI 模型拟完成的任务对 AI 模型进行微调，确保模型能够灵活应对不同任务需求，提高模型的泛化性与迁移性。该阶段 AIGC 服务提供者将对 AI 模型

进行一系列技术验证测试及安全测试，并最终形成可部署 AI 模型。该周期内，AIGC 服务提供者所需要遵守的合规义务至少包括：

（1）算法安全及伦理合规义务。包括但不限于：

（a）建立算法机制机理审核制度，定期审核、评估、验证算法机制机理、模型、数据和应用结果，禁止研发并及时制止诱导用户沉迷、过度消费等违反法律法规或者违背伦理道德的算法模型；

（b）建立算法反歧视制度，在算法设计、训练数据选择、模型生成和优化、提供服务等过程中，防止因为 AI 模型的算法偏见而导致产生民族、信仰、国别、地域、性别、年龄、职业、健康等算法歧视现象；

（c）建立算法安全制度，包括（i）算法鲁棒性监测制度，通过红队测试，定期模拟对抗样本攻击、数据污染等，监测、评估、验证并优化算法鲁棒性；（ii）算法透明度监测制度；（iii）算法信息过滤监测制度，在程序、算法方面设置过滤、发现或抵制机制；（iv）模型可控性监测制度；（v）算法安全事件应急处理相关规则等。

（2）网络安全义务。包括但不限于：

（a）按照网络安全等级保护制度，开展网络安全认证、检测、风险评估、风险管理等活动，制定网络安全事件应急预案，及时处置系统漏洞、计算机病毒、网络攻击、网络侵入等安全风险；

（b）开展数据安全教育培训，采取相应的技术措施和其他必要措施，保障数据安全；

（c）明确数据安全负责人和管理机构，落实数据安全保护责任。

3. 部署使用周期。在该周期内，AIGC 服务提供者将完成 AI 模型及下游应用系统的定型、系统部署与运行测试，跑通在线服务，并最终形成以 AI 模型为基础，面向最终用户端提供完整操作体验的可商用化 AI 系统。该周期内，AIGC 服务提供者所需要遵守的合规义务主要包括：

（1）建立 AIGC 公平竞争机制，杜绝利用算法共谋方式形成垄断、排除市场竞争，遵循反垄断、反不正当竞争相关法律规定；

（2）建立 AIGC 透明度支持团队，在主管部门依对 AIGC 的服务开展监督检查时予以配合，并按照主管部门的要求对训练数据来源、规模、类型、标注规则、算法

机制机理等予以说明，并提供必要的技术、数据等支持和协助。

4. 运营监控周期。该周期是AIGC服务提供者将成熟的AI系统进行商业化落地的周期。在该周期内，AI模型将正式投入市场使用。AIGC服务提供者须实时监控用户及市场反馈，及时处理各类应急情况并与有关政府部门合作。在发生危害网络安全的事件时，AIGC服务提供者须立即启动应急预案，采取相应的补救措施，并按照规定向有关主管部门报告。同时，应当对数据处理活动定期开展风险评估，并向有关主管部门报送风险评估报告，以维护AI系统的稳定性。

（四）内容生态治理制度

根据《网络信息内容生态治理规定》[13]《算法推荐管理规定》《深度合成管理规定》等规定的要求，建议AIGC平台应当以下列合规要点为抓手，建立并完善平台内容生态治理制度体系：

1. 制定网络信息内容生态治理细则，健全用户注册、账号管理、信息发布审核、跟帖评论审核、版面页面生态管理、实时巡查、应急处置等制度，同时设立网络信息内容生态治理负责人，配备与业务范围和服务规模相适应的专业人员；

2. 建立内容知识产权侵权内容处置流程，并在平台规则中设置“通知 - 删除”相关条款，及时屏蔽、下架相关侵权内容链接，并对侵权用户采取封禁等措施；

3. 如采用个性化算法推荐技术推送信息的，应设置符合规定的推荐模型，建立健全人工干预和用户自主选择机制；

4. 编制网络信息内容生态治理工作年度报告，年度报告应当包括网络信息内容生态治理工作情况、网络信息内容生态治理负责人履职情况、社会评价情况等内容；

5. 加强深度合成内容管理，采取技术或者人工方式对深度合成服务使用者的输入数据和合成结果进行审核；

6. 建立辟谣机制，投诉、举报机制以及舆情监控机制，并配备相应的外部操作入口以及内部标准处理流程，及时受理并处理用户的投诉、举报以及所发现的平台相关舆情，保存有关记录，并向有关主管部门报告，对相关使用者依法依约采取警示、限制功能、暂停服务、关闭账号等处置措施；

7. 建立健全用于识别违法和不良信息的特征库，完善入库标准、规则和程序，记录并留存相关网络日志；

8. 提供模拟自然人进行文本的生成或者编辑服务、语音生成或者显著改变个人身份

[13] 2019.12.15 发布，2020.03.01 实施。

特征的编辑服务、人物图像、视频生成或者显著改变个人身份特征的编辑服务等具有生成或者显著改变信息内容功能的服务时，可能导致公众混淆或者误认的，应当在生成或者编辑的信息内容的合理位置进行显著标识，向公众提示深度合成情况。

（五）用户权益保护制度

根据《互联网用户账号信息管理规定》[14]《个人信息保护法》《网络信息内容生态治理规定》《算法推荐管理规定》等法律法规的要求，为确保 AIGC 平台的用户权益，AIGC 服务提供者应当采取的合规管控措施至少包括：

1. 建立用户实名认证制度；

2. 建立用户个人信息保护制度；

3. 明确用户权利义务；

4. 保证用户知情权；

5. 保证用户选择权；

6. 明确用户投诉处理路径；

7. 指导用户认识 AIGC 技术，并防范未成年沉迷及过度依赖。

三、互联网应用关键条款完善

由于 AIGC 在内容生成物属性方面的不确定性以及对于数据安全的潜在影响力，建议 AIGC 平台在用户协议条款中加入了不同于传统互联网应用的安排，以便最大程度防范潜在的版权纠纷以及数据风险。

1. 输入内容的权利安排。在 AIGC 服务中，用户输入的内容通常包含一定的知识产权。根据《中华人民共和国著作权法》[15] 及其实施细则，由用户创造的、用户输入内容的原始权利应当归用户所有，但同时用户亦需要为其所输入的内容承担相应的责任，例如保证不侵犯第三方权利，保证不得含有违反法律法规或公序良俗的信息。

2. 输入内容的使用限制。由于用户输入内容的权利属于用户，而 AIGC 服务提供者往往需要使用输入内容进行模型优化，因此，AIGC 服务提供者需要就原始输入内容的后续使用获得用户的明确授权。AIGC 服务提供者可以在用户协议中明确规定，服务提供者使用用户输入内容的范围和目的。

3. 生成内容的权利安排。生成内容的权利安排相较于用户输入内容则更为复杂，目前，行业内采取的权利安排主要分为用户独占，或用户所有 + 平台无偿使用两种类型。我

[14] 2022.06.27 发布，2022.08.01 实施。
[15] 2020.11.11 发布，2021.06.01 实施。

们理解，赋予用户对生成内容的权利有助于 AIGC 行业吸引更多用户，同时允许 AIGC 服务提供者在一定范围内使用这些内容，例如用于展示、宣传或模型优化，则可以同时满足 AIGC 平台的模型自我学习与进化的需求。

4. 生成内容的使用限制。AIGC 服务提供者作为生成内容的主要责任人之一，有责任确保其服务不被用于非法目的，因此需要在用户协议以及平台准则中限制生成内容的使用，以避免用户将生成内容用于从事违法活动或侵犯他人权利的行为之中。

四、外部商业合作

除了上文所提及的资质证照、内部管理体系及制度以及互联网应用所必需的外显文件外，AIGC 平台在对外合作过程中还会与诸多合作方签署不同的商业合作协议。该等商业合作协议根据 AIGC 平台自身的性质与合作事项的不同包括诸多类型。AIGC 平台需要根据不同的协议类型以及自身的商业安排与合作方就各自的权利义务达成一致。以下，我们主要就相关协议中 AIGC 平台需要关注的重点内容进行简要提示。

（一）与技术支持方签署的 AI 模型开发 / 许可协议

倘若 AIGC 平台作为平台运营方，在与技术支持方签署的 AI 模型开发或许可协议中，AI 模型本身的知识产权权属问题通常是双方的关注要点之一。AI 模型作为计算机软件，其核心技术成果包括代码及经训练后获得的参数。其中，AI 模型代码可以作为计算机软件作品获得版权保护，AI 模型代码对应的算法在符合商业秘密构成要件的情况下也可以作为商业秘密获得保护，因此，与传统软件开发 / 许可协议相同，协议双方通常会针对合作过程中所产生的上述相关权属安排进行提前约定。

而 AI 模型参数是用于定义模型的可调整变量，主要基于大量输入训练数据集、调整模型的原始参数值从而输出最佳结果。因此，AI 模型参数的取得与训练数据的数量和质量紧密相关，这也导致协议双方在约定 AI 模型参数的知识产权权属时可能会产生一定的争议，我们建议在合作前期即对相关权属进行提前约定。

（二）与数据提供方签署的数据交易 / 数据训练协议

通常情况下，技术支持方在开发 AI 模型的过程中将会使用到大量的训练数据，该等训练数据可能由 AIGC 平台自行采集、开放数据爬取或通过数据交易的方式从其他数据提供方处取得。对于通过数据交易方式取得的数据，一方面，AIGC 平台应当确保数据提供方提供的数据具有合法来源，在数据交易协议中要求数据提供方对相关数据的收集和对外提供等处理行为进行陈述保证，例如遵守相关法律法规、不侵犯第三方的合法权益等，并

约定相应的法律责任。另一方面，考虑到高质量的训练数据对模型的优化至关重要，因此 AIGC 平台还应当在数据交易协议中明确数据的采购要求及验收标准。

此外，数据的使用方式以及相关权益的归属也是协议双方应当关注的要点之一。就数据的使用方式而言，技术支持方通常需要对数据提供方提供的数据进行提取、清洗、筛选等衍生处理以用于模型训练或其他商业用途，为了避免后续产生纠纷，双方应当在协议中就如何使用数据进行明确约定。就数据相关权益的归属而言，数据虽然在权利属性方面尚存争议，但从实践来看，双方仍可以就数据相关权益的归属、后续利用范围和限制等予以明确约定。特别地，对于经过 AIGC 平台衍生处理后的数据形成的相关权益，AIGC 平台亦可以在协议中予以主张。

结语

在探索 AIGC 这片广袤而充满潜力的新海洋时，合规便是企业航船的罗盘。合规经营不仅是企业遵循法律的基本要求，更是其长远发展和市场竞争力的关键所在。

其实，合规不仅是对现有规定的被动遵循，它更应是一种主动的战略布局，需要企业不断更新对法律环境的认知，主动预测和应对法规的变化，构建与时俱进的合规机制，制定前瞻性的合规计划，从而为自身赢得战略优势。在 AIGC 领域，合规更要求研究机构、上下游企业以及法律合规团队等参与主体不断地更新知识体系、积极参与行业对话，方能行稳致远。

感谢实习生缪逸泓、张文溢、何一辰、徐若宸对本文作出的贡献。

“侵权包赔”：AIGC 知识产权赔偿条款之对比分析

宋海燕 张永洁

一、背景引入

在生成式人工智能（Generative AI）广泛应用的过程中，人工智能生成内容（AI Generated Content，以下简称“AIGC”）所引发的侵权风险引起了各方重视。全球知名 AIGC 公司如 OpenAI[1]、微软[2]、Anthropic[3] 等因侵权纠纷频遭被诉。虽然目前大多数被诉主体都是 AIGC 服务提供者，但 AIGC 服务的使用者（即用户）也同样面临侵权风险。据人工智能内容治理公司 Acrolinx 于 2023 年 8 月对 86 家财富 500 强企业的调查结果，约 30% 的企业将“知识产权侵权风险”视为应用 AIGC 技术时最为关注的问题[4]。为消除市场忧虑，以微软为代表的多家 AI 公司推出了知识产权赔偿条款（Indemnity clause）。本文将围绕微软、OpenAI、谷歌、Amazon、Adobe、Shutterstock、Canva 等公司推出的 AIGC 知识产权赔偿条款进行简单介绍，以期对其异同点进行梳理对比。

二、各大 AIGC 公司的知识产权赔偿条款之对比分析

本文所介绍的“知识产权赔偿条款”（以下简称“赔偿条款”），是指 AI 公司承诺其用户：若用户因使用其提供的 AIGC 产品或服务而面临第三方侵权索赔，AI 公司同意承担相应的赔偿责任。2023 年 9 月 7 日，微软发布了 Copilot 版权承诺书（Copilot Copyright Commitment），表示若商业用户（Commercial customer）因使用 Copilot 或其生成输出而遭第三方基于版权的侵权诉讼，微软将同意为该用户辩护并支付因案件而产生的赔偿金——前提是该用户已使用产品中内置的防护装置（Guardrails）和内容过滤器（Content filters）并遵守了其他条款[5]。此外，Shutterstock[6]、Getty Images[7]、

---

[1] 2023 年 9 月 19 日，美国作家协会以及包括《权力的游戏》原著作者乔治·R·R·马丁（George R.R. Martin）在内的 17 位美国著名作家提起集体诉讼，指控 OpenAI 未经允许使用原告作品训练 ChatGPT 模型。参见 Authors Guild et al. v. Open AI Inc. et al. Case No. 1:23-cv-08292.

[2] 2023 年 10 月 17 日，微软、Facebook 母公司 Meta 等企业遭到了美国阿肯色州前州长 Mike Huckabee 和其他几位作家的起诉，被指控未经授权使用原告的书籍训练其 AI 模型。参见 Huckabee v. Meta Platforms, Inc. Case No. 1:23-cv-09152.

[3] 2023 年 10 月 18 日，环球音乐集团等音乐出版商起诉 Anthropic，称其 AI 模型 Claude 未经允许使用受版权保护的歌词作为训练数据，并向用户提供并传播了与受版权保护的歌词相同或相似的生成内容，构成版权侵权。参见 Concord Music Group, Inc. v. Anthropic PBC. Case No. 3:23-cv-01092.

[4] 参见 Charlotte Baxter-Read. “Enterprise Insights into Generative AI Adoption” Acrolinx, 8 Aug.2023,www.acrolinx.com/blog/enterprise-insights-into-generative-ai-adoption/. Last visited on December 27, 2023.

[5] “Microsoft announces new Copilot Copyright Commitment for customers.” Microsoft, 7 Sep. 2023, blogs.microsoft.com/on-the-issues/2023/09/07/copilot-copyright-commitment-ai-legal-concerns. Last visited on December 27, 2023.

[6] “Introducing Indemnification for AI-Generated Images: An Industry First.” Shutterstock, 29 Aug.2023,www.shutterstock.com/blog/ai-generated-images-indemnification. Last visited on December 27, 2023.

[7] “Getty Images Launches Commercially Safe Generative AI Offering.” Getty Images, 25 Sep.2023,newsroom.gettyimages.com/en/getty-images/getty-images-launches-commercially-safe-generative-ai-offering. Last visited on December 27, 2023.

谷歌[8]、OpenAI[9]、Anthropic[10]等公司也陆续向用户作出了类似承诺，表明将为用户承担因使用 AIGC 服务而产生的知识产权侵权赔偿责任。

虽然各大公司的赔偿条款在具体形式上存在差异，但大多涵盖了适用主体、索赔类型、保护额度、具体程序以及限制条件这五个关键要素。本文将对此进行逐一梳理。

（一）适用主体

目前，大多数 AI 公司推出的赔偿条款仅适用于付费用户，就具体条款而言可分为以下情形：

（1）只有额外付费购买才可获得赔偿权利。譬如，Adobe Firefly 在官网介绍中表明赔偿条款适用于企业用户，同时声明企业用户需要额外付费购买，才能获得相应的 IP 赔偿（IP indemnification）权利[11]。

（2）明确规定适用主体是付费用户。例如，OpenAI 在服务条款中规定其赔偿条款适用于付费用户，包括 API 用户和 ChatGPT 企业用户[12]。而 ChatGPT 数亿的免费用户，则不会受到赔偿条款的保护。

（3）明确规定适用的产品范围是付费产品。例如，微软声明其赔偿条款适用于 Copilot 商业服务和 Bing Chat Enterprise 的付费版本（Paid versions of Microsoft commercial Copilot services and Bing Chat Enterprise）。这表明免费版 Bing 的用户不在微软赔偿条款的保护范围之内[13]。另，Google 也指出其赔偿条款适用的产品范围为 Google Workspace 中的 Duet AI 以及 Google Cloud 等一系列付费产品，不包括 Google 对标 ChatGPT 而推出的免费 AI 聊天机器人——Google Bard[14]。

虽然各 AIGC 公司关于适用主体的表述存在差异，但保护付费用户是目前 AIGC 赔偿条款的共性趋势，且多数公司特别强调对企业用户的保障。究其原因，我们认为可能存在以下几点考量：首先，付费用户是 AIGC 公司重要的收入来源，AIGC 公司有动力为其提供额外的赔偿保障，以维持用户忠诚度，并进一步扩大用户市场。其次，若企业用户使用 AIGC 进行大规模商业活动，将会面临更高的侵权风险。因此，提供赔偿条款有助于缓解企业用户在使用 AIGC 产品或服务时所面临的潜在法律风险，确保其商业活动的顺利进行。

---

[8] “Shared fate: Protecting customers with generative AI indemnification.” Google, 13 Oct. 2023,cloud.google.com/blog/products/ai-machine-learning/protecting-customers-with-generative-ai-indemnification. Last visited on December 27, 2023.

[9] “Business terms.” OpenAI, 14 Nov. 2023,openai.com/policies/business-terms. Last visited on December 27, 2023.

[10] “Expanded legal protections and improvements to our API.” Anthropic, 19 Dec. 2023, www.anthropic.com/index/expanded-legal-protections-api-improvements. Last visited on December 27, 2023.

[11] “Adobe Firefly for enterprise.” Adobe, www.adobe.com/sensei/generative-ai/firefly/enterprise.html. Last visited on December 27, 2023.

[12] “Service terms.” OpenAI, 6 Nov. 2023, openai.com/policies/service-terms. Last visited on December 27, 2023.

[13] “Microsoft announces new Copilot Copyright Commitment for customers.” Microsoft, 7 Sep. 2023, blogs.microsoft.com/on-the-issues/2023/09/07/copilot-copyright-commitment-ai-legal-concerns. Last visited on December 27, 2023.

[14] “Shared fate: Protecting customers with generative AI indemnification.” Google, 13 Oct. 2023, cloud.google.com/blog/products/ai-machine-learning/protecting-customers-with-generative-ai-indemnification. Last visited on December 27, 2023.

（二）索赔类型

在 AIGC 的数据训练和应用中，存在两个关键阶段可能引发侵权：首先，在数据训练（Training data）阶段，AIGC 公司若未经授权收集、使用他人受知识产权保护的数据和内容用于训练 AI，则有可能构成著作权侵权。其次，在生成输出（Generated output）阶段，若 AI 生成输出包含受知识产权保护的作品或与他人作品构成实质性相似，也可能构成侵权。此外，在深度合成技术的应用场景中，AI 生成输出还可能侵犯肖像权、名誉权、隐私权与个人信息等权益。

目前多数 AIGC 公司的赔偿条款主要针对 AI 生成输出的侵权索赔。这些赔偿条款涵盖的索赔类型不尽相同，可分为以下三种模式：

（1）只涵盖版权（Copyright）。如微软 Copilot 赔偿条款明确仅适用于生成输出的版权侵权，不包括商标侵权等 [15]。

（2）涵盖更广泛的知识产权（Intellectual Property Rights），如版权、专利、商业秘密。值得注意的是，商标大多不在知识产权赔偿条款的涵盖范围之列。例如，Canva[16] 赔偿条款明确适用于版权、专利或商业秘密的侵权索赔。Google、[17]OpenAI[18] 以及 Amazon[19] 在服务条款中约定：其赔偿义务仅适用于知识产权索赔。但是，这三家公司还约定，如果 AIGC 用户在商业活动中使用 AIGC 产品或服务涉及侵犯他人商标或相关权利，将不会得到赔偿，这实际上限制了 AIGC 用户对商标索赔享有的赔偿权利。

（3）涵盖知识产权、个人形象权（Right of Publicity）、隐私权（Right of Privacy）等人格权。某些 AIGC 公司在提供知识产权侵权赔偿的同时，进一步将其涵盖范围延伸到其他合法权益。如 Adobe Firefly 赔偿条款不仅适用于 AIGC 侵犯任何第三方的专利、版权、商标，还包括针对个人形象权及隐私权的索赔 [20]。

从上述政策中可见，基于版权的豁免是 AIGC 公司提供的知识产权赔偿条款中最常见的类型。各 AIGC 公司在提供知识产权赔偿条款上的差异主要源于其业务定位和产品特性的不同。以生成图片的 AIGC 产品为例，这类产品通常依赖大量图像数据进行训练，其输出更易涉及人物肖像和其他敏感个人信息，进而触及个人形象权和隐私权等法律问题。因此，Adobe、Getty Images 和 Shutterstock 此类产生 AIGC 图像或视频内容的公司，更

---

[15] “Microsoft announces new Copilot Copyright Commitment for customers.” Microsoft, 7 Sep. 2023, blogs.microsoft.com/on-the-issues/2023/09/07/copilot-copyright-commitment-ai-legal-concerns. Last visited on December 27, 2023.

[16] “Canva for Teams Subscription Service Agreement.” Canva, 26 Jul. 2022, www.canva.com/policies/enterprise-ssa. Last visited on December 27, 2023.

[17] “Google Workspace Service Specific Terms.” Google, 20 Nov. 2023, workspace.google.com/intl/en/terms/service-terms. Last visited on December 27, 2023.

[18] “Service terms.” OpenAI, 6 Nov. 2023, openai.com/policies/service-terms. Last visited on December 27, 2023.

[19] “AWS Service Terms” Amazon, 15 Dec. 2023, aws.amazon.com/cn/service-terms/?nc1=h_ls. Last visited on December 27, 2023.

[20] “Firefly Legal FAQs-Enterprise Customers.” Adobe, 13 Sep. 2023, www.adobe.com/content/dam/cc/us/en/products/sensei/sensei-genai/firefly-enterprise/Firefly-Legal-FAQs-Enterprise-Customers-2023-09-13.pdf. Last visited on December 27, 2023.

倾向于将赔偿条款扩展到知识产权之外的人格权。

（三）保护额度

多家 AIGC 公司在推出赔偿条款时，声明会为用户提供全额赔偿。如 OpenAI 表示：若第三方向其 AIGC 用户提出索赔涉及知识产权侵权，OpenAI 将会为用户支付由有管辖权的法院最终判定的任何损害赔偿金额以及支付给第三方的和解金额[21]。Getty Images 同样表示，在满足条件的情况下，其 AIGC 用户有可能会得到无上限的赔偿（Uncapped indemnification）[22]。

（四）具体程序

用户若想获得赔偿，必须遵循各公司赔偿条款所规定的具体程序。以 Amazon 的 AWS 服务条款为例，仅当用户完成以下既定程序时，Amazon 才承担赔偿责任：第一，立即向 AWS 发出索赔书面通知；第二，允许 AWS 控制索赔的辩护；第三，在必要的范围内保留并提供足够的记录，以评估用户能够获得赔偿的资格；第四，在索赔的辩护及和解中与 AWS 合理合作（费用由 AWS 承担）[23]。

其他公司的具体实施程序与之相似，通常包括以下步骤：

（1）用户需及时通知，并积极配合。在面临第三方的侵权索赔时，用户有责任及时通知 AIGC 服务提供者并配合提供相关信息。例如，Shutterstock 要求用户应在知道或应当知道被诉之后的 5 个工作日之内，以书面方式通知 Shutterstock，这是获得赔偿的必要前提[24]。用户的及时通知和配合有助于 AIGC 公司掌握案件具体情况，制定有力的诉讼策略。这是整个赔偿流程的关键起点。

（2）AIGC 公司介入并控制诉讼流程。如 OpenAI 在其商业条款中明确规定了提供赔偿一方应当“对诉讼程序拥有完全控制权”[25]。这意味着将由 AIGC 公司负责选择律师团队、制定诉讼策略，并做出是否接受和解的决策。

AIGC 公司大多具备应对知识产权侵权诉讼的丰富经验，由公司控制诉讼流程，既可以减轻用户的诉讼成本，也可以确保案件得到高效和专业的处理，从而最大限度地保护用户和 AIGC 公司自身的利益。

（五）限制条件

AIGC 公司通常会在赔偿条款中设置若干限制性条件，旨在确保用户负责任地使用其

---

[21] “Business terms.” OpenAI, 14 Nov. 2023, openai.com/policies/business-terms. Last visited on December 27, 2023.
[22] “Getty Images Launches Commercially Safe Generative AI Offering.” Getty Images, 25 Sep. 2023, newsroom.gettyimages.com/en/getty-images/getty-images-launches-commercially-safe-generative-ai-offering. Last visited on December 27, 2023.
[23] “AWS Service Terms” Amazon, 15 Dec. 2023, aws.amazon.com/cn/service-terms/?nc1=h_ls. Last visited on December 27, 2023.
[24] “Shutterstock License Agreement.” Shutterstock, 17 May 2022, www.shutterstock.com/license. Last visited on December 27, 2023.
[25] “Business terms.” OpenAI, 14 Nov. 2023, openai.com/policies/business-terms. Last visited on December 27, 2023.

AIGC 产品或服务，降低滥用风险。若 AIGC 用户违反了限制条件，将不能享有赔偿权利。AIGC 赔偿条款的“限制或例外条件”大致可归纳为以下五种：

（1）用户故意或放任侵权。如果用户利用 AIGC 故意侵权，例如用户知道使用特定 AI 生成内容可能导致侵权，或用户输入的材料本身侵权，或用户在收到侵权指控后继续使用这些内容，则此类用户将不能适用赔偿条款。

（2）用户对 AIGC 的生成内容进行修改。根据赔偿条款声明或用户协议，赔偿条款的适用前提是 AIGC 用户须按原样（As is）使用 AI 输出内容，不得对其进行修改或编辑[26]。

（3）用户绕开 AIGC 产品内置的安全过滤器。AIGC 产品通常内置有各种安全保护和内容过滤机制，以防止 AIGC 的生成侵权。若用户故意绕开这些机制，表明他们有意增加侵权风险，因此不符合赔偿条款的适用条件。如 Google 表示用户必须使用 Google 提供的引用来源（Source citations）、过滤器（Filters）等工具，才能适用赔偿条款[27]。

（4）用户侵犯了他人商标权。OpenAI 等公司专门规定，如果 AIGC 用户在商业活动中使用 AIGC 产品或服务时侵犯了他人的商标权，则不能适用赔偿条款。

（5）用户违反用户协议的规定或者违约使用 AIGC 产品。具体的违规行为包括 AIGC 用户未经合法授权使用 AIGC 产品，或使用 AIGC 产品从事非法活动等。这种“兜底式”的规定既赋予 AIGC 公司灵活调整空间，以防止 AIGC 产品或服务被滥用，也鼓励用户按照约定负责任地使用产品或服务，以减少侵权风险。

结语

AI 广泛应用提升了生产力的同时，也引发了法律风险。知名 AIGC 企业纷纷推出了知识产权赔偿条款，这反映出 AIGC 行业对法律风险和用户权益的高度重视。目前推出赔偿条款的大多是资金雄厚的跨国科技公司，他们凭借充足的财务预算和丰富的知识产权侵权应对经验，敢于向用户作出“侵权包赔”的承诺。

尽管赔偿条款等机制为 AIGC 服务使用者提供了一定程度的保护，但并不能涵盖所有潜在的法律风险。AIGC 的知识产权归属以及侵权责任界定等问题依然悬而未决。这些法律问题亟需各国法律法规和司法判例的持续更新来加以应对。

感谢律师王默、赵怡冰，实习生陈颖思对本文作出的贡献。

[26] “Business terms.” OpenAI, 14 Nov. 2023, openai.com/policies/business-terms. Last visited on December 27, 2023.
[27] “Google Workspace Service Specific Terms.” Google, 20 Nov. 2023, workspace.google.com/intl/en/terms/service-terms. Last visited on December 27, 2023.

Sora 或者 ChatGPT：AI 生成的内容究竟归谁

张逸瑞 钱琪欣 冯宝宝 蔡文苑 朱佳蔚

引言

随着 Sora 的诞生，人们已经越来越意识到人工智能生成内容（Artificial Intelligence Generated Content，“AIGC”）可以处理更加复杂的数据类型和任务，其生成内容的效率和逼真度也将着实可期。与此同时，AIGC 的法律定性、权益分配、责任承担等问题成为司法界、实务界及学界讨论的热点。本书《ChatGPT 出品：谁是作者？》《论图片生成式 AIGC 平台在侵权纠纷中的角色与责任边界》等文中也对相关问题进行了解读和评述。

在中国，全国首例“AI 文生图”著作权侵权案（“S 案”）一审判决生效，引发了新一轮关于 AIGC 的可版权性以及权利归属问题的热烈讨论。在该案中，法院赋予利用人工智能生成的图片著作权法的保护，并肯定了 AI 使用者“创作者”身份，对 AIGC 引发的诸多著作权难题进行了探索和尝试，也为后续类似案件的处理提供了参考和借鉴。

本文将结合 S 案判决以及国内外相关实践，对 AIGC 的可版权性以及权利归属问题作进一步探讨和分析。

一、AIGC 的可版权性问题

（一）我国相关规定及司法实践

1. 相关法律规定

根据《中华人民共和国著作权法（2020 修订）》[1]（“《著作权法》”）第 3 条及《中华人民共和国著作权法实施条例》[2]（“《著作权法实施条例》”）第 2 条规定，受《著作权法》保护的“作品”，是指文学、艺术和科学领域内具有独创性并能以一定形式表现的智力成果。从上述规定可见，AIGC 想要构成受《著作权法》保护的“作品”，需要满

[1] 2020 年 11 月 11 日发布，2021 年 6 月 1 日生效。
[2] 2013 年 1 月 30 日发布，2013 年 3 月 1 日生效。

足“属于文学、艺术和科学领域”、“能以一定形式表达”、“具备独创性”，以及“属于人类智力成果”四个构成要件（“作品四要件”）。

就 AIGC 是否“属于文学、艺术和科学领域”而言，由于 AIGC 的表达形式主要为文字、图片、视频、音频等，因此往往能够被认为具备“属于文学、艺术和科学领域”这一构成要件。

就 AIGC 是否“能以一定形式表达”而言，主要的判断标准包括：（1）相关内容是否能够被人类所感知，并以一定形式被复制；以及（2）相关内容是否与“思想”存在区别，构成了具象化的“表达”。此处“思想”与“表达”的区别主要在于创作时能够供作者进行选择的表述范围之大小，即如果针对某种概念只有唯一一种或有限的表述形式，则这些表述应被视为“思想”，而非“表达”，不能够受到著作权的保护[3]。由于 AIGC 往往以数据的形式进行传输和复制，并且能够生成和人类创作的普通作品外观无异的内容，因此往往能够被认为具备“能以一定形式表达”这一构成要件。

基于此，是否符合“独创性”和“智力成果”要件就成了 AIGC 能否构成“作品”的关键。我国理论和实务界普遍认为，基于鼓励创作的目的，《著作权法》应当只保护人类的智力成果，“作品”的独创性需体现人的智力选择与判断。由于在部分 AIGC 产品中，人类仅需要输入简单的提示词（prompts），AI 就能生成富有细节的内容（如下图，用户仅输入“日出”的提示词，即可得到复杂的体现日出的图片），人类对最终生成内容的贡献似乎微乎其微，这也是将 AIGC 纳入《著作权法》保护面临的最大的挑战。

[3] 雷献和、赵琪与张晓燕的其他著作权权属侵权纠纷案，最高人民法院（2013）民申字第 1049 号民事裁定书，最高人民法院（指导性案例 81 号）。

2. S 案中法院对 AIGC 法律属性的认定

（1）案情简介

原告李某（“原告”）使用某款知名 AIGC 软件（“S 软件”）通过输入提示词的方式生成了一张图片（“涉案图片”）并在社交平台上共享。被告刘某（“被告”）未经许可，于其注册使用的百家号账号上发布了一篇文章并于其中使用了涉案图片。原告起诉称，被告未获得原告许可，且截去了原告在小红书平台的署名水印，使得相关用户误认为被告为该作品的作者，侵犯了原告对涉案图片享有的著作权专有权利（包括署名权及信息网络传播权）。

（2）法院对 AIGC 是否构成作品的分析

由于该案中原告主张享有著作权的图片系由 AI 生成，相关图片是否受《著作权法》保护就成了案件审理的基础和争议焦点。法院按照上述《著作权法》的规定，对涉案图片是否符合作品四要件逐一进行了详细审查和分析，在认定涉案图片的外观与通常的照片、绘画无异，符合“属于文学、艺术和科学领域”、“能以一定形式表达”两个要件的基础上，进一步对涉案图片是否符合“独创性”和“智力成果”要件作出了有意义的探索，在一定程度上回应了社会各界的关切。

（a）关于 AI 生成图片是否属于智力成果

法院首先对 S 软件的原理进行了考察，认为“S 软件可以根据文本指令，利用文本中包含的语义信息与图片中包含的像素之间的对应关系，生成与文本信息匹配的图片。该图片不是通过搜索引擎调用已有的现成图片，也不是将软件设计者预设的各种要素进行排列组合”“该模型的作用或者功能类似于人类通过学习、积累具备了一些能力和技能，它可以根据人类输入的文字描述生成相应图片，代替人类画出线条、涂上颜色，将人类的创意、构思进行有形呈现”。

在此基础之上，法院对原告使用 S 软件生成涉案图片的过程进行了分析，指出原告为了实现其希望获得的图片，先输入了关于图片艺术类型、主体、环境、人物呈现方式等的提示词并设置了相关参数，根据初步生成的图片，又增加了提示词、调整了参数，最终选择了一幅自己满意的图片。在此过程中“从原告构思涉案图片起，到最终选定涉案图片止，这整个过程来看，原告进行了一定的智力投入，比如设计人物的呈现方式、选择提示词、安排提示词的顺序、设置相关的参数、选定哪个图片符合预期等等”，最终认定涉案图片体现了原告的智力投入，具备“智力成果”要件。

（b）关于 AI 生成图片是否具备独创性

法院强调，“利用人工智能生成图片，是否体现作者的个性化表达，需要个案判断，不能一概而论”“人工智能生成图片，只要能体现出人的独创性智力投入，就应当被认定为作品，受到著作权法保护”。在该案中，“原告对于人物及其呈现方式等画面元素通过提示词进行了设计，对于画面布局构图等通过参数进行了设置，体现了原告的选择和安排。另一方面，原告通过输入提示词、设置相关参数，获得了第一张图片后，其继续增加提示词、修改参数，不断调整修正，最终获得了涉案图片，这一调整修正过程亦体现了原告的审美选择和个性判断。”因此，涉案图片能够体现本案原告的主观选择与个性化表达，满足“独创性”要求。

（3）既往案例的内在逻辑：包含人类独创性贡献的 AIGC 可以受《著作权法》保护

事实上，在 S 案之前，我国已经有两例因机器生成物著作权问题引发的案例。虽然这两个案件的判决结果不同，但法院的论述却体现出内在的一致性，即：由自然人创作是获得著作权法保护的必要条件，机器生成物只有在具有人类独创性贡献的情况下才可能成为著作权法意义上的作品 [4]。

- F 案 [5]：在该案中，法院明确表示自然人创作完成仍应是《著作权法》上作品的必要条件。涉案分析报告系威科先行库利用用户输入的关键词与算法、规则和模板结合自动生成，虽然在其生成过程中有两个环节（软件开发环节和软件使用环节）有自然人参与，但最终生成的结果并没有传递自然人（软件开发者和软件使用者）的思想、感情的独创性表达。由于涉案分析报告不是自然人创作的，因而不构成“作品”，无法获得《著作权法》的保护。
- D 案 6：法院审查后认为，在原告使用 D 软件生成涉案文章的过程中，文章框架模板的选择和语料的选定等均由原告主创团队相关人员选择与安排，该等选择与安排符合《著作权法》关于创作的要求，从而认定涉案文章属于《著作权法》所保护的作品。

S 案实际上继承了前述案件的内在逻辑和价值取向，在对 AIGC 是否构成作品的判断上，仍然强调个案中人类对生成结果的贡献和投入。值得关注的是，相比于否定 AI 生成内容作品属性的 F 案，S 案所涉及的 AI 产品更加智能，能够根据用户输入的简单提示词生成富有细节的图片，理论上使用该产品生成的内容获得作品保护的难度更高。为何在 S

---

[4] 具体案例分析详见《ChatGPT 出品：谁是作者？》，https://mp.weixin.qq.com/s/vHNlN3S3WnLy-UpAahyV4w?version=4.1.20.6024&platform=win，本文仅为对法院核心观点的概括。

[5] 参见北京知识产权法院（2019）京 73 民终 2030 号民事判决书。

[6] 参见深圳市南山区人民法院（2019）粤 0305 民初 14010 号民事判决书。

案中法院作出完全相反的认定呢？从判决书来看，一个关键原因在于两案中用户创作过程存在较大差异：在 F 案中，原告在制作涉案分析报告的过程中仅仅只输入了简单的搜索关键词；而在 S 案中，原告首先输入了涉及图片的艺术类型、主体、环境等多方面要素的提示词，并在获得第一张图片后，继续增加提示词、修改参数，不断调整修正，最终获得涉案图片。由此，用户输入的提示词越全面、越具体，被认定为对最终生成结果具有独创性贡献的可能性越高。而对生成内容的反复修正调整也将为用户的贡献提供有力佐证。

虽然 S 案判决已经生效，但作为全球首例明确认定 AI 生成图片能够受到版权保护的案件，该案的判决仍然存在一定探讨空间。根据《著作权法实施条例》第 3 条规定，创作是指直接产生文学、艺术和科学作品的智力活动。为他人创作进行组织工作，提供咨询意见、物质条件，或者进行其他辅助工作，均不视为创作。即现行关于著作权的法律规定强调“直接产生”。我们理解，“直接产生”的含义是考察人类决定表达内容的自由意志与作品之间联系的紧密程度。只有当人类的智力投入与最终的表达之间具备的因果关系强到可以认定最终表达直接由人类决定时，该等表达才能够被认定为满足“智力成果”的要求。而 S 案中原告所输入的指令、提示词、后续修正等，只是对构成作品的表达产生了间接影响，似乎不能达到直接决定表达本身的紧密程度。作品的表达仍然是由执行指令的主体（即 AI 本身）自行选择判断并以个性化的方式实现的。在 S 案中，对于原告在本案中的智力投入与最终表达之间如何构建直接因果关系并未进行详细的论证。

（二）其他国家相关规定及实践

虽然关于 AIGC 的立法和实践不同国家和地区存在一定差异，但总体来看，除了英国等少数国家和地区通过单独立法，明确为计算机生成作品（从定义来看也可涵盖 AIGC 产品）提供著作权保护外，包括欧洲大陆、美国、日本等在内的国家或地区都或多或少强调或考量了人类的贡献这一因素。

1. 相关立法规定

国际组织或国家	文件名称	相关内容
国际保护知识产权协会（AIPPI）	《人工智能生成作品版权问题》	只有在作品的创作过程中存在人为干预贡献并且满足其他保护条件的情况下，所有生成的作品才有资格受到版权保护。没有人为干预的情况下，所有生成的作品不应受到版权保护[7]。

[7] Copyright in artificially generated works, 2019 AIPPI World Congress, September 18, 2019, https://aippi.soutron.net/Portal/Default/en-GB/RecordView/Index/35.

国际组织或国家	文件名称	相关内容
美国	《美国版权局实践纲要》	• 第 306 条 美国版权局将登记一部作为作者的原创作品，前提是该作品是由人创作的[8]。 • 第 313.2 条 要获得“作者身份”作品的资格，作品必须由人类创造。 版权局拒绝注册由机器或纯粹的机械过程产生的作品，该过程在没有人类作者的任何创造性输入或干预的情况下随机或自动运行[9]。
	《版权登记指南：包含人工智能生成材料的作品》	版权保护的是人类创意的产物。AI 生成内容是否能够取得版权注册取决于创作的具体情况，**特别是 AI 工具如何操作，以及它如何被用来创建最终的作品**： • 如果一部作品的传统创作要素是由机器制作的，那么这部作品就缺乏人类的创作，版权局将不予注册。例如当技术工具仅根据人类的提示（Prompt）产生复杂作品，则该作品的创作要素就是由机器制作的，而非人类。 • 如果人类艺术家以足够有创意的方式选择或安排 AI 生成的材料，或者艺术家修改了 AI 生成的材料，该等修改能达到版权标准，则版权局**可以保护这些作品中由人类所创作完成的部分**。 • 人类仍然可以使用技术工具进行创作，比如使用 Photoshop 来编辑图片，也仍然构成作者。关键在于人是否**对最后的作品进行了创作的控制、是否完成了创作的要素。关键的问题在于“‘作品’是否基本上是人类原创，计算机（或者其他设备）仅仅为一种辅助工具，或者作品中的传统创作元素（文学，艺术或音乐的表达，或选择、安排等元素）是否实际上不是由人类而是机器所构思、执行。”**[10]
欧盟	《关于发展人工智能技术的知识产权的报告》	人工智能生成的内容必须受知识产权法律框架的保护，以鼓励对这种形式的创造进行投资，提高公民、企业的法律确定性，而且因为它们目前是人工智能技术的主要使用者之一。**人工代理和机器人自主创作的作品可能不符合版权保护的条件，以遵守与自然人相关的原创性原则，而且“智力创作”的概念涉及作者的人格**[11]。

[8] Compendium of the U.S. Copyright Office Practices § 306, US Copyright Office （3d ed. 2021）.
[9] Compendium of the U.S. Copyright Office Practices § 313. 2, US Copyright Office （3d ed. 2021）.
[10] Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence, US Copyright Office, Mar 16, 2023, https://www.federalregister.gov/documents/2023/03/16/2023-05321/copyright-registration-guidance-works-containing-material-generated-by-artificial-intelligence.
[11] Intellectual property rights for the development of artificial intelligence technologies, European Parliament, https://www.europarl.europa.eu/doceo/document/TA-9-2020-0277_EN.html.

国际组织或国家	文件名称	相关内容
日本	《著作权审议会第 9 小委员会（计算机创作物关联）报告书》	• 关于计算机创作物的著作权财产 （1）根据《版权法》，“作品”被定义为“属于文学、科学、艺术或音乐范围内的思想或情感的创造性表达”。 这一定义基于《版权法》的立法目的，假定它是一个人思想和情感的创造性表达，属于一个综合性知识和文化概念的范围。 （2）人们传统上使用工具来创作作品（书写工具、打字机、文字处理机等），而**计算机创作的作品也让人们能够创造性地表达自己的思想感情。作为作品实现目的的“工具”，那么其版权性就会得到肯定**。 （3）为了使某人被认定为使用计算机系统作为工具创作了受版权保护的作品， 首先，必须有利用计算机系统表达想法和感受的创造性意图。但这种创作意图通常可以从使用计算机系统的实际行为中推断出来，并不需要事先对具体结果的形式有明确的意图，现阶段只要有一个意图就足够了。意图“使用计算机创建具有某种表达方式的产品，可以将其视为个人个性的表达。” 其次，在创作过程中，该人必须做出足以被视为创造性贡献的行为，以获得具体的结果。虽然需要根据具体情况来判断什么样的行为才算足以认定为创造性贡献，但判断的主要标准是创作类型、行为主体、因此，我们将在“二、计算机创作作品特有的版权问题”中对此进行进一步讨论。 此外，最终产品需要具有足够的外观，以便能够客观地评价其作为思想和情感的创造性表达，即使是不使用计算机的正常创作，这一点也不会改变。 （4）一般而言，创意作品要被认定为受版权保护的作品，必须满足（3）中的所有要求，但不使用计算机的普通创作当就事物而言，一个人的创作意图和行为通常被认为是理所当然的，而在实践中，版权的归属往往只能通过评估最终的产品来确定。对于计算机创作的作品，考虑到计算机系统干预创作过程的性质，有必要审查人的创作意图以及是否存在创作行为 [12]。
英国	1988 年《版权、外观设计和专利法》	• 第 9 条 如果在没有人类作者的情况下由计算机生成的设计，则**计算机生成作品的著作权归“创作作品所需安排（the arrangements necessary）的人”所有** [13]。

[12] 著作権審議会第 9 小委员会（コンピュータ创作物関係）報告書，https://www.cric.or.jp/db/report/h5_11_2/h5_11_2_main.html#0.
[13] Copyright, Designs and Patents Act 1988, https://www.legislation.gov.uk/ukpga/1988/48/contents.

2. 相关案例

在 AI 技术更为成熟和发达的美国，已经出现数例因 AIGC 能否受版权保护引发的案件。在这些案件中，美国司法 / 执法机构均采取了较为保守的态度，认为只有在人类对 AI 最终生成的结果具有控制力、可预见性的情况下，其才能成为受版权法保护的作品。

例如，在 THALER v. PERLMUTTER et al，Civil Action No. 22-1564（BAH）一案中，原告泰勒创作了一个名为“Creativity Machine”的 AI 系统，泰勒声称该系统自行生成了一件虚拟艺术作品，标题为“通往天堂的捷径”，因此其试图以该 AI 系统为作者，向美国版权局申请作品登记。

泰勒在向美国版权局申请登记过程中，将系统“Creativity Machine”列为作者，理由是“该作品由机器上运行的算法自主创建。”而原告作为该系统的所有者，机器生成的作品应被视作他雇佣机器完成的雇佣作品，故原告泰勒应获得作品的著作权。美国著作权局于 2019 年 8 月驳回了该著作权申请，拒绝为其注册，其主要理由是“著作权法仅适用于人类创作的作品。”遭遇拒绝后，泰勒于 2020 年、2022 年多次要求重新考虑他的申请，但美国著作权局均以类似理由拒绝。最终，原告希望通过起诉的方式完成著作权的注册，但哥伦比亚特区地区法院驳回了原告的动议。

法院认为，虽然著作权与时俱进，但人类的创造力是可著作权性的核心条件。该案中的原告泰勒在申请著作权时，并未提及自己独创性的智力劳动，故其作品难以通过著作权登记审核。虽然原告在庭审中试图改变表述方式，如该作品的产生是“由他提供指令，且 AI 完全由他本人控制与运作。”但法院认为，这些陈述直接与原告申请著作权时的行政记录相矛盾，故不予采信。

再如，在美国版权局审查委员会复议维持拒绝登记人工智能生成内容《SURYAST》图像为作品的决定（SR # 1-11016599571; Correspondence ID: 1-5PR2XKJ）中，Ankit Sahni 利用人工智能软件 RAGHAV，以其拍摄的日落照片为底稿，参照梵高的《星月夜》风格，并结合输入的风格强度变量提示生成了二维图像“SURYAST”。Ankit Sahni 于 2021 年 12 月就“SURYAST”提交版权登记申请，并将其与 RAGHAV 共同列为“SURYAST”的作者。其中，根据 Ankit Sahni 的描述，RAGHAV 因委托作品获得作者身份。2022 年 6 月，美国版权局首次作出决定拒绝 Ankit Sahni 的申请。Ankit Sahni 则于 2022 年 9 月及 2023 年 7 月两度申请复议。2023 年 12 月，美国版权局复议后维持了拒绝登记“SURYAST”图像为作品的决定，主要理由在于：《版权登记指南：包含人工智能生成材料的作品》中规定，只有当作品包含足够的人类创作因素时，该作品才能够

受到版权保护；如果作品的所有“传统作者元素”（文学、艺术或音乐领域的表达、选择或编排等）都是由人工智能生成的，则其缺乏人类作者身份。本案中，Ankit Sahni 仅向 RAGAV 提供了三个输入：一个基本图像、一个风格图像和一个决定风格转移量的可变值，并由 RAGHAV 而确定如何根据风格传递值对基础图像和风格图像进行插值，生成新的二维图形。图像中的具体元素（日落、云和建筑等元素的出现与否及具体的位置、颜色）的去向都不是由 Ankit Sahni 控制，而是由 RAGHAV 生成，因此 Sahni 先生对 RAGAV 创作作品的创作控制不足，无法进行注册。

（三）共识和挑战

从国内外的实践可见，虽然对于 AIGC 能否获得版权保护众说纷纭，但在现阶段，一个相对被广泛采用的分析思路是，以人类对最终生成结果的控制力和贡献作为 AIGC 能否受版权保护的判断标准。这一思路似乎为解决 AIGC 的可版权性问题提供了一条可行的路径，但如何界定 AI 生成结果中是否存在人类的控制力和贡献，以及如何划定控制力与贡献的界限，在个案适用过程中仍然存在一定困难。

相较于 Word、PPT、Photoshop、Premiere 等传统创作辅助软件，AI 能力的介入使得 AIGC 场景下人类的贡献与最终生成结果之间的关系变得疏离和抽象。AIGC 产品所固有的生成结果随机性、跨模态等特征，决定了人类无法再像使用传统软件那样掌控 AI 生成内容的每一处细节。如果对人类的参与程度采取相对宽松的标准，认为人的智力投入并不需要达到完全决定最终表达的程度，而只需要通过对提示词、参数等的选择安排对表达产生一定程度的影响，大概率会推导出 AIGC 构成作品的结论；反之，如果认为人类的智力投入需要完全决定最终表达，两者之间需要具有一一对应的关系，则更可能会推导出 AIGC 不构成作品的结论。这种判断尺度的差异，也是导致当前对 AIGC 法律属性产生争议的重要原因。

二、AIGC 归属于 AI 技术开发者还是 AI 服务使用者？

根据《著作权法》第十一条的规定，著作权属于作者，也就是创作作品、为作品贡献创造性智力劳动的人（包括自然人和组织创作的法人）。在 AIGC 生成的过程中，有自然人介入和参与的部分大体集中在产品开发和使用两个环节。在 AIGC 构成作品，即 AIGC 中存在人类独创性贡献的情况下，理论上，AIGC 的权益归属可能存在以下几种情况：

1. AIGC 的内容来源于开发者的预设。在此情况下，AI 最终生成的内容均由开发者预先设定，使用者只能按既定的规则和流程使用 AI 产品，不能随意增添开发者没有设定的素材。由于使用这类 AI 产品生成的内容都属于开发者预设的范围，体现的是开发者对素

材的取舍、安排，因此，AIGC 的著作权被认定为归属于开发者的可能性较大。

2. AIGC 的内容来源于使用者的输入。在此情况下，AI 产品的主要功能是辅助用户进行创作，用户可以按照自己的想法控制生成内容的具体表达。与之相应，AIGC 的著作权被认定为归属于使用者的可能性较大。

3. AIGC 的内容来源于开发者、使用者双方。这类 AI 产品生成内容的过程与两人合作创作的过程类似，可能会与使用者存在多轮交互，根据使用者的需求分别提供开发者预设的不同素材，共同形成新的内容。此时，AIGC 作为开发者和使用者“合作完成”的作品，其著作权被认定为由开发者、使用者共同享有的可能性较大。

但从当前主流 AIGC 产品的形态来看，在 AI 技术研发者、AI 使用者两者中，AI 技术研发者往往更关注构建并优化算法本身，而缺乏对于 AI 在使用时输出的海量具体生成内容在创造性方面的思考和要求，尽管在内容生成的过程中，AI 技术研发者会考虑通过算法设计规避包括政治宗教、伦理道德、公共秩序等方面的话题讨论，亦可能使用人工识别数据标签训练额外的违规内容识别模型，将其内置于 AIGC 的算法中，以此剔除 AIGC 中违法违规、违反伦理道德和公序良俗的内容从而影响最终的内容，但该等行为更多是为了遵守普遍的法律法规与公序良俗的要求，很难解释为 AI 技术研发者对于 AIGC 的独创性智力活动，因此，AI 技术研发者成为作者的可能性相对较小[14]，而在前述 AIGC 侵权案等国内司法判例中，法院亦认定各案中的涉案内容是基于 AI 使用者的智力投入直接产生，且体现出了 AI 使用者的个性化表达，故 AI 使用者是涉案内容的作者，享有涉案内容的著作权。

尽管如此，在当事人之间有约定的情况下，应尊重当事人之间的意思表示，优先适用当事人之间的约定。在当前 AIGC 相关法律制度和规则还不甚明朗的情况下，通过用户协议等方式对AIGC的权益归属进行约定有助于AIGC的后续利用和争议解决（在前述S案中，法院在判断涉案图片的权利归属时亦对产品协议的相关约定进行了考虑），也是实践中较为常见的做法。

从当前国内的实践来看，各平台的用户协议等规定中都会对基于 AI 使用者输入内容所自动生成的输出结果的知识产权归属进行约定。其中，约定知识产权归属平台所有的情况较为少见，更为常见的是约定 AIGC 相关的知识产权归属 AI 使用者所有，但 AI 使用者需要许可平台在全球范围内享有免费的、永久的、可分许可的使用权利，包括用于对 AIGC 进行复制、修改、进一步利用等。

[14] 详见《ChatGPT 出品：谁是作者？》，https://mp.weixin.qq.com/s/vHNlN3S3WnLy-UpAahyV4w?version=4.1.20.6024&platform=win。

除此之外，部分平台选择将 CC0 1.0 协议[15]纳入平台协议体系中，用以处理 AIGC 的知识产权归属问题。CC0 1.0 协议是适用于全球范围的一套版权共享协议，该协议规定，在法律允许的范围内，将 CC0 1.0 协议适用于其作品的主体即放弃了其根据版权法在世界范围内对作品享有的所有权利，包括所有相关权和邻接权，从而将作品放入公有领域。任意主体均可复制、修改、传播和表演该作品，甚至用于商业目的，而无需取得权利人的任何许可。值得注意的是，CC0 1.0 协议并不影响任何人的专利权或商标权，亦不影响权利人对于该作品或如何使用该作品所拥有的权利，如公开权或隐私权。除非另有明示声明，否则在适用法律允许的最大范围内，该作品的权利人不对作品作任何保证，亦不对作品的所有使用负任何责任。

结语

整体而言，随着 AIGC 技术的发展和应用，相关的法律和政策也在不断更新和探索，以保持与时代发展的同步。除了在《著作权法》框架下的分析外，有一个更为原点的问题同样值得我们思考：我们究竟是否需要对 AIGC 赋予法律保护，为什么要对 AIGC 赋予法律保护？可能当我们对这个问题得出相对确定的结论之后，在对 AIGC 法律属性的分析上，我们会有一个更加清晰的认识和逻辑起点。

感谢实习生缪逸泓、张文溢、何一辰对本文作出的贡献。

[15] 参见链接：https://creativecommons.org/publicdomain/zero/1.0/。

ChatGPT 许可应用，知识产权和数据怎么看？

张逸瑞 吴之洲 张津豪

随着以 ChatGPT 为杰出代表的人工智能（Artificial Intelligence，“AI”）软件一次次“火爆出圈”，针对 AI 软件功能、价值、意义等多领域的讨论也从原先的仅限于技术圈内部扩散至社会全域。2023 年 2 月 27 日，中共中央、国务院印发《数字中国建设整体布局规划》（“《规划》”），《规划》指出，要全面赋能经济社会发展，推动数字技术和实体经济深度融合，在农业、工业、金融、教育、医疗、交通、能源等重点领域，加快数字技术创新应用。AI 作为支撑数字经济发展的重要基础设施，正在与各行业典型应用场景相融合，将为我国数字经济发展提供核心驱动力。

商业实践中，AI 软件发挥作用的方式通常体现为 AI 企业将其研发的 AI 软件许可给使用者，以收取许可费的形式盈利。如何合理安排 AI 软件许可协议中双方的权利义务，特别是知识产权和数据相关条款如何设计，在该等业务模式下至关重要。以下，我们将基于 AI 软件与传统软件的区别，对 AI 软件许可协议中知识产权和数据条款设计所应当包含的要素进行探讨。

一、AI 软件与传统软件的区别

（一）软件开发方式

对于传统软件而言，软件开发者更关注的是软件的功能需求，即软件必须实现的功能。因此，软件开发者需要通过使用各种模型对相关功能需求进行描述，数据处理等规则往往已经被事先设计确定。而对于 AI 软件而言，功能需求相对并不那么重要，模型训练则十分关键，模型开发者通过使用大量的数据对待训练模型进行持续训练，使之归纳出处理新数据的规则。待训练模型通过学习知识成为具有推理和决策能力的训练后模型，从而实现智能化。因此，相比于传统软件，AI 软件开发者更关注的是模型、训练模型的数据以及支撑模型训练的算力。

（二）数据使用方式

在传统软件开发过程中，由于没有模型训练的环节，软件开发者一般不需要收集并使用大量的数据。而在 AI 软件的开发过程中，软件开发者则必须借助大量且高质量的数据对模型进行训练，并在训练过程中不断优化参数以提高运行效率和准确性。训练数据通常根据具体的应用场景进行确定。以计算机视觉应用场景为例，尽管利用一些现有的开源数据也可以对模型进行训练，但是这些数据通常不能很好地满足特定的视觉应用场景需求，解决上述问题的关键在于如何采集足够多的来自于实际应用场景的真实图像或视频数据，并对这些数据进行一定的处理，例如数据清洗、数据标注等。

（三）软件部署方式

从软件使用者角度出发，AI 软件的安装部署方式与传统软件可能并无明显差异，但是从运营方式和商业模式来看，二者还是存在一定区别。对于传统软件而言，其对算力的要求相对较低，因此通常是由企业购买后安装在其自有服务器上，相关数据也通常存储在本地计算机或服务器中。而对于 AI 软件而言，新兴应用场景产生的海量数据对 AI 算力的需求持续加大，例如云游戏、自动驾驶等对数据传输的速度和量级都提出了更高的要求，而通过云计算和云部署的方式便可以在很大程度上解决上述问题。在该等情形下，相关数据则被传输并存储在云端。

二、AI 软件许可协议知识产权关注要点

鉴于上述提到的区别，相比于传统软件许可协议，AI 软件许可协议在知识产权条款的设计方面也存在特殊的安排，尤其是在许可标的、知识产权权属、侵权风险以及责任承担方面。

（一）许可标的及其知识产权权属

为了明确软件许可协议中不同知识产权的权属安排，我们有必要先对软件许可中常见的许可标的进行梳理。

1. 软件许可标的

在传统软件许可协议中，关于许可标的的安排一般会区分源代码和目标代码。源代码是由程序员用人类可读的语言编写的用于执行某些任务的代码，然后将文件保存为规定的格式，但该等代码未经编译无法被机器直接执行；而目标代码则是通过编译器将源代码转换而成的机器可直接执行的代码。由于目标代码通常难以被人类所理解，因此倘若需要对软件进行修改，例如增加定制化的功能模块，则往往需要对源代码进行修改。实践中，如

果被许可方对软件的需求仅涉及运行和使用，一般不涉及源代码的交付；但是如果被许可方对软件的维护、调整、改进和升级有特定需求，许可方通常还需要向被许可方交付软件的源代码，并授予其源代码层面的许可。

如上文所述，在处理传统软件相关许可标的时，一种常见的思维模式是“程序员编程→源代码→编译→目标代码→机器执行”；而在面对 AI 软件时，上述思维模式可能需要予以进一步调整，这是因为还需要考虑到 AI 模型在整个软件开发过程中的作用。不同于传统软件通常直接由程序员编写源代码赋予功能，AI 软件通常由算法工程师编写的训练程序训练而来，训练程序通过执行一定的算法，从训练数据中归纳出某些“推理规则”，这些“推理规则”代码化后便构成了训练后的 AI 模型。从上述意义上说，模型是程序产生的程序。

基于上述比较，回到 AI 软件许可协议许可标的的层面，应当专门对 AI 模型予以特别约定——如果被许可方仅需利用许可方已有的训练后模型，则被许可方根据许可协议取得训练后模型一定的使用权即可；但在很多场景下，被许可方需要的并非已有的训练后模型，而是定制化的训练后模型，对于该等定制化的训练后模型的权利归属、使用条款，双方有必要在许可协议中予以进一步约定。

2. 知识产权权属

在传统软件许可协议中，无论许可标的是目标代码还是源代码，双方均应当对相关知识产权的权属安排进行提前约定，以免后续产生纠纷。一般而言，软件许可协议的知识产权归属安排会根据时间顺序采用“三段式”的叙述逻辑，即背景知识产权、前景知识产权和改进知识产权。其中，背景知识产权是指协议一方在履行协议前拥有或取得的技术成果及相关知识产权，前景知识产权是指在双方合作期间产生的知识产权，而改进知识产权则是指对前景知识产权进行的修改、改编或提升，包括但不限于对前景知识产权相关的功能、性能、部件或模块的变更等。

在传统软件许可协议的谈判过程中，以前景知识产权为例，若许可方向被许可方提供目标代码或源代码层面的许可，相关前景知识产权的安排一般需要考虑双方的谈判地位。强势的一方通常会要求前景知识产权全部归其所有，在某些情形下可以考虑后续免费或附条件地许可另一方使用。倘若双方之间的谈判地位相当，则一般会约定由做出实质性贡献的一方享有相关前景知识产权。

而在 AI 软件许可协议中，由于许可标的涉及 AI 模型，相关前景知识产权在形成与权属约定方面则与传统软件许可协议存在诸多差异。如上文所述，模型是由训练程序从训练

数据中归纳出的某种“推理规则”，在此过程中，训练数据的质量和标注精度对模型的准确性起到至关重要的作用，换言之，训练程序输入不同的训练数据后所输出的模型也不尽相同。一般而言，模型的训练分为静态训练（static training）和动态训练（dynamic training）两种，因此，模型也分为静态模型与动态模型。对于静态模型，模型训练好则长期投入使用，而对于动态模型而言，随着新数据的不断输入，通过对这些数据的整合，模型也将不断进行更新迭代。

因此，在 AI 软件许可中，若许可方许可的仅是静态模型，则被许可方在具体的应用场景下使用该等模型，模型不会在被使用时同步自我演化或改进，被许可方只能通过许可协议要求许可方向其定期提供更新后的模型。但是，若被许可方获得的是动态模型的许可，由于被许可方持续不断地向模型输入实际应用场景的数据，模型也将被不断训练进而形成新的版本。在该等情形下，由于模型在使用被许可方所提供的数据过程中实现了自我改进，被许可方本身便可以对该等改进所形成的前景知识产权主张相应的权利。即使在许可方较为强势进而主张相关前景知识产权为自己单独所有的情况下，被许可方也可以考虑要求许可方就最新版本的模型向自己提供一项免费的许可，对此，双方还应当在许可协议中进一步明确许可费、更新维护等相关事项。

3. AIGC 的保护

在传统软件许可中，许可方基于目标代码进行研发或创作的成果一般归属于被许可方，例如被许可方利用 Word 软件编写的文档在构成作品的前提下受到著作权法的保护。但是，在 AI 软件许可中，则面临关于人工智能生成内容（Artificial Intelligence Generated Content，“AIGC”）可版权性的讨论，对该问题的具体分析可以参见本书中《ChatGPT 出品：谁是作者？》一文。整体而言，在现行法律体系下，AIGC 很可能难以通过著作权进行保护，以合适的方式向 AI 使用者明确告知其享有的相关权益至关重要，例如在 AIGC 构成作品情况下的著作权归属、通过 AIGC 进行二次创作情况下的相关权益分配等。

（二）知识产权侵权风险

当前，对 AI 知识产权相关问题的讨论更多围绕在 AIGC“是否构成作品”以及“权利归属”等问题上，然而事实上，模型训练中可能产生的潜在知识产权侵权风险同样不能忽视。2023 年 2 月 15 日，《华尔街日报》记者弗朗西斯科·马可尼发布推文称，ChatGPT 模型的训练未经授权使用了大量主流媒体的新闻数据，包括路透社、纽约时报、卫报、BBC 等，但从未支付任何费用[1]。

[1] 《陷入侵权风波！ OpenAI 遭媒体指责：白用我们的文章训练 ChatGPT！》，财联社，https://m.cls.cn/detail/1270005，最后访问日期：2024 年 3 月 19 日。

仅从我国著作权法相关法律法规（“著作权法”）来看，倘若 ChatGPT 模型对训练数据的使用行为无法满足作品“合理使用”构成要件，在未获得相关著作权人许可的情况下，可能构成著作权侵权。

1.“合理使用”的适用困境

根据一般的著作权法理论，“合理使用”是指在特定情况下使用作品，可以不经著作权人许可，不向其支付报酬，但应当指明作者姓名或者名称、作品名称，并且不得影响该作品的正常使用，也不得不合理地损害著作权人的合法权益。这是因为著作权法的立法目的在于通过授予著作权人垄断权利来鼓励文学、艺术和科学领域的创作和传播，但有一些事项在立法者眼中具有更高的价值位阶，著作权人的垄断权利需要让位于这些事项（例如社会运行过程中对于知识和信息的最基本需求）。以我国著作权法为例，“合理使用”的事由包括但不限于“个人使用”“适当引用”“在时事新闻报道中使用”“在课堂教学和科学研究中使用”等。

尽管相关立法已经创设了多种可以适用“合理使用”的场景，但当我们将 AI 软件与“合理使用”的相关标准进行比对时，可能依然很难找到可以完全适用的条款。现行立法中与 AI 软件情形较为接近的合理使用情形主要包括“个人使用”“适当引用”和“科学研究”三类，但在适用时均存在一定的困难：首先，AI 软件大多数是面向不特定主体提供服务，难以符合“个人使用”的适用条件；其次，“适当引用”的前提是“为介绍、评论说明某一作品”或“说明某一问题”，ChatGPT 等 AI 软件对作品的商业化使用行为也不符合上述目的；最后，“科学研究”对作品的使用必须是为了“学校课堂教学或者科学研究”，以及“供教学或者科研人员使用”，此外对使用的作品仅能“少量复制”，而 AI 模型训练由于需要使用大量的数据，对相关作品的复制并非“少量”，因此也难以满足上述要求。综上，仅从我国著作权法来看，利用已有作品进行 AI 模型训练的行为似乎很难构成“合理使用”。

2.“许可使用”的现实障碍

若 AI 软件对作品的使用不构成“合理使用”，则必须取得相关作品著作权人的许可。但是，对于 AI 模型的训练数据而言，确保训练数据中包含的作品全部获得作品著作权人的许可在现实中并非易事。一方面，AI 软件开发者需要花费大量的时间和成本将可能受保护的作品从训练数据中识别出来；另一方面，针对识别出来的受保护的作品，AI 软件开发者还需逐一地与作品的著作权人进行协商取得其许可，并支付许可费用。考虑到不同作品许可谈判的难度以及 AI 软件开发的时效性，在实践中逐一取得相关作品著作权人许

可的可行性可能并不高。

由此可以看出，AI 软件开发过程中模型训练的特殊性不可避免地导致其可能存在侵犯第三方知识产权的风险。因此，在许可标的涉及 AI 模型的软件许可协议中，双方有必要对该等潜在的知识产权侵权风险以及双方的责任分配作出明确约定。知识产权不侵权保证条款是软件许可协议中的常见条款，一般而言，被许可方应当要求许可方就使用许可软件行为不侵犯第三方知识产权作出陈述与保证，并约定在侵犯第三方知识产权引发赔偿的情况下许可方所应承担的责任。

三、AI 软件许可协议数据关注要点

AI 软件许可标的的不同使得 AI 软件许可协议知识产权条款的设计应当有特殊的考量，而 AI 模型本身对数据天然的依赖性则要求协议双方在协商谈判时还应当特别关注数据的使用和权属、数据安全合规等在内的相关问题。

（一）数据使用与权属

1. 数据使用

为了不断提升模型的性能，ChatGPT 等 AI 模型一般还需要使用用户提供的数据作为模型训练的新数据来源。ChatGPT 的使用条款规定：我们可能会使用您通过 ChatGPT 输入的内容以及 ChatGPT 基于您输入的内容输出的内容来提供、维护、开发和改进我们的服务[2]；而 OpenAI API 的使用条款规定：我们不会将您通过 OpenAI API 输入的内容以及 OpenAI API 基于您输入的内容输出的内容用于开发或改进我们服务[3]。在《如何使用您的数据来提高模型性能》文档中，模型开发者进一步明确了使用相关数据的目的：AI 模型最有用和最有前途的特性之一是它们可以随着时间的推移而改进。我们通过科学和工程突破以及接触现实世界的问题和数据不断改进我们的模型[4]。

许可方可以使用被许可方提供的数据对模型进行实时的训练以提高模型的准确性，但是在实践中，并非所有的被许可方均希望将自己收集的数据作为训练数据提供给许可方。与传统软件许可相比，被许可方的数据更容易被许可方当作训练数据用于其他模型开发，特别是，倘若许可方将利用被许可方数据开发的模型提供给被许可方的竞争对手，那么将会对被许可方的市场竞争产生巨大影响。因此，部分被许可方会在许可协议中明确约定许可方不能使用相关数据进行模型训练。即使被许可方获得的是动态模型的许可，也通常会对许可方使用相关数据的目的和范围进行限制，例如，若许可方不允许被许可方访问其全

---

[2] https://openai.com/terms/，最后访问日期：2024 年 3 月 19 日。
[3] https://openai.com/policies/business-terms，最后访问日期：2024 年 3 月 19 日。
[4] https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance，最后访问日期：2024 年 3 月 19 日。

部客户群的聚合数据，则被许可方同样可以要求许可方就被许可方的数据对其他客户施加相同的访问限制。

2. 数据权属

在 AIGC 的可版权性成为人们的讨论焦点之余，AI 相关数据的权属安排也是 AI 软件许可协议中双方绕不开的话题之一。总体而言，数据可以分为原始数据和衍生数据。原始数据是数据采集时提供的、反映客观事物属性的记录，是不经过任何加工、创作或提取、编辑的数据。衍生数据是指基于特定的商业目的、通过运用一系列技术手段对数据进行筛选、分析、处理从而形成的数据。

AI 软件的使用过程中可能涉及的数据主要有三类，包括模型训练阶段使用的训练数据以及模型使用阶段的输入数据和输出数据。其中，训练数据又包括原始训练数据和训练数据集。原始训练数据是指模型开发者直接收集的数据，理论上来说，AI 模型接受的训练数据越多，其自我进化也会更快，但是这种情况必须建立在训练数据没有任何错误的基础上。因此，模型开发者往往会在原始训练数据的基础上进行一定的处理，例如数据清洗、数据标注、数据分组等，从而形成高质量和高精准的训练数据集用于模型训练。模型使用阶段的输入数据一般是具体应用场景下的原始数据，例如使用者的个人信息、受著作权保护的作品等；输出数据即为上文提及的 AIGC。在动态模型训练中，模型使用阶段的输入数据和输出数据也有可能成为新的训练数据以进一步改进模型。模型开发者为了避免在 AI 模型实际应用过程中发生训练阶段无法预期的事件，可能会要求将使用阶段采集的数据作为训练数据来生成新的精度更高的模型。

目前为止，针对数据权属的问题虽尚未形成清晰的解决方案，但是一般认为对于原始数据权利的确认并不代表否认原始数据主体的权利。因此，就训练阶段使用的原始训练数据而言，相关数据主体应当对其享有相关权益。而对于经过模型开发者处理形成训练数据集而言，在模型开发者经过充分授权对原始训练数据进行了收集、清洗、标注等衍生开发后，模型开发者对经过自己合法数据活动形成的数据集合原则上应当享有占有、使用、收益和处分的权利。因此，在对衍生数据进行界定的前提下，各方可以基于自身的谈判地位以及各自的商业需求对衍生数据的权属进行安排。与原始训练数据类似，对于模型使用阶段的输入数据，相关权益也应当归属于输入数据主体。但是对于输出数据，在其法律属性界定尚存在争议的情况下，建议双方在协议中对相关数据的权益归属、使用方式等进行明确约定。

（二）数据安全合规

如上文所述，AI 模型训练、应用中涉及大量的数据，从行业维度来看，这些数据可

以分为金融数据、交通数据、自然资源数据、卫生健康数据、科技数据等；从数据载体维度来看，这些数据可以分为音频数据、视频数据、图像数据、文字数据等；而从数据主体维度来分，上述数据又可以分为个人数据、企业数据和公共数据等。在 AI 模型训练和后续的许可中，无论是许可方还是被许可方，均应当特别注意数据的来源合规问题。此外，在确保数据来源合规的前提下，双方还应当就如何使用相关数据，使用相关数据所应当采取的安全保护措施等进行明确约定。

1. 数据来源合规

考虑到在 AI 软件许可协议中，数据的使用场景主要包括模型训练阶段对训练数据的使用以及模型使用阶段对输入数据的使用，且模型使用阶段收集的数据后续也可能成为新的训练数据，因此，无论是对于许可方还是被许可方，均应当确保自身使用的数据具有合法来源。一般而言，对于 AI 模型而言，获取数据的方式主要包括数据交易、自行采集和开放数据爬取。数据交易是指通过合法的交易方式从数据提供方处获取相关数据，自行采集是指通过 APP、传感器、相机等方式直接采集数据，开放数据爬取则是指通过数据爬虫等方式获取开放的数据。对于数据交易和自行采集两种获取方式而言，最重要的是要确保如何取得相关数据权利主体的授权。而对于开放数据爬取而言，则更应当关注数据爬虫行为本身是否合法，例如爬虫所采取的技术手段是否突破数据访问控制、数据爬虫的使用目的是否正当等。

对于许可方而言，例如，在收集和使用个人数据进行模型训练时，可能存在的风险包括但不限于侵犯人格权和个人信息权。《中华人民共和国民法典》第一百一十条规定：“自然人享有生命权、身体权、健康权、姓名权、肖像权、名誉权、荣誉权、隐私权、婚姻自主权等权利。”第一百一十一条规定：“自然人的个人信息受法律保护。任何组织或者个人需要获取他人个人信息的，应当依法取得并确保信息安全，不得非法收集、使用、加工、传输他人个人信息，不得非法买卖、提供或者公开他人个人信息。”《中华人民共和国个人信息保护法》第二条规定：“自然人的个人信息受法律保护，任何组织、个人不得侵害自然人的个人信息权益。”以个人信息为例，除法律另有规定，许可方只有在取得个人信息主体同意的前提下才能处理相关个人信息。在该等个人信息来源于其他第三方的情况下，许可方至少还应当要求相关个人信息的提供方保证其提供的个人信息获得了个人信息主体的同意。

对于被许可方而言，一方面，其作为 AI 模型的使用方，可以在协议中要求许可方对其提供的模型不侵犯第三方权利作出陈述与保证，常见的陈述与保证条款，例如：“模型

的开发系根据适用法律法规的要求进行，模型的许可不会侵犯任何第三方的合法权益”。但是考虑到在不同的场景下双方谈判地位可能存在的差距，许可方同样也可以对己方的某些义务进行免除，由许可方作出的、典型的该等陈述与保证条款，例如：“模型按‘现状’和‘可获得’方式予以授权，不附带任何种类的明示或默示保证，许可方对模型的使用不承担任何责任”。另一方面，在被许可方将模型使用阶段获取的数据提供给许可方以对模型进一步训练改进的情形下，被许可方同样需要履行相关合规审查义务，包括其向许可方提供数据的行为是否已获得了数据主体的充分授权，是否违反其应当履行的保密义务等。

2. 数据安全保护

由于 AI 软件的云计算和云部署等特点，在 AI 软件许可协议中，许可方的数据安全保护能力往往是被许可方关注的重点。如前文所述，在 AI 模型的使用阶段，其会采集各行业领域的不同类型的数据，这些数据中可能包括敏感个人信息，国家重要数据等对安全保护有特殊要求的数据。

以自动驾驶为例，智能驾驶汽车上集成的摄像头、激光雷达、导航仪等各类传感器，每时每刻都在收集车主本人、乘车人、驾驶人等的个人信息、车辆的环境信息以及车辆行驶信息等。根据《汽车数据安全管理若干规定（试行）》，车辆行踪轨迹、音频、视频、图像和生物识别特征等信息属于敏感个人信息，而军事管理区、国防科工单位以及县级以上党政机关等重要敏感区域的地理信息、人员流量、车辆流量等数据、汽车充电网的运行数据等则属于重要数据[5]。若汽车数据处理者对收集的上述数据进行不当使用，将可能导致个人信息主体的人身、财产安全以及国家安全受到损害。对此，法律法规规定汽车数据处理者在处理敏感个人信息时，应当符合特定要求，例如应具有直接服务于个人的目的，包括增强行车安全、智能驾驶、导航等；在处理重要数据时，应当按照规定开展风险评估并形成风险评估报告、报送汽车数据的安全防护和管理措施，包括保存地点、期限等[6]。

因此，在 AI 软件许可协议中，被许可方应当要求许可方对数据的采集、存储、使用、传输等各方面均采取充分的数据安全保护措施，防止数据被窃取、滥用、篡改或毁损，并对可能因数据安全问题导致的责任承担进行明确约定。此外，在 AI 软件许可领域，由于许可方很有可能是境外主体，在该等情形下，数据出境可能引发的数据安全相关问题应当引起被许可方的特别关注。倘若在使用 AI 软件过程中确实涉及数据出境，被许可方应当在协议中明确要求许可方遵守数据出境的合规要求和履行数据出境申报义务。例如，被许可方可以在协议中要求许可方承诺其对相关数据的使用应当遵守中国关于数据出境的相关法律法规。

---

[5] 《汽车数据安全管理若干规定（试行）》第三条。
[6] 《汽车数据安全管理若干规定（试行）》第九条、第十条、第十三条。

结语

2022 年 7 月 29 日，科技部等六部门印发的《关于加快场景创新 以人工智能高水平应用促进经济高质量发展的指导意见》提出，要着力打造人工智能重大场景、提升人工智能场景创新能力、加快推动人工智能场景开放以及加强人工智能场景创新要素供给。应用场景需求是技术进步的重要推动力，而如何合理安排“开发者”与“使用者”双方的权利义务则是人工智能应用场景落地的重要保障和关键一步。本文重点对 AI 软件许可协议中的知识产权和数据条款如何设计进行了探讨，在此基础上，交易双方可以结合具体的交易场景和交易类型进行量身打造，从而最大程度维护自身利益。

浅析 ChatGPT 训练数据之合理使用

宋海燕 陈佩龄

引言

ChatGPT，一款由美国科技公司 OpenAI 于 2022 年 11 月 30 日发布的 AI 聊天机器人，一经面世便引发全球热议。随着其热度不断升高，与之相关的诸多版权争议受到广泛关注，训练数据侵权问题便是其中之一。

作为语言生成式模型，ChatGPT 训练数据由大量文本数据组成。目前各国对生成式 AI 训练数据的使用仍未单独制定成文法规定，但域外对文本与数据挖掘（Text Data Mining，后称“TDM”）技术的法律规制却具有重要借鉴意义。TDM 指的是利用自动分析技术分析文本与数据的模式、趋势以及其他有价值的信息，是以计算机为基础的，从文本或数据导出或组织信息的过程 [1]。从技术原理来看，ChatGPT 训练数据库的建构与 TDM 均以文本和数据输入为基础，二者在著作权法上具有相似意义。而在法律层面上，基于制度衔接与法律秩序稳定性的考量，针对使用主体、使用目的、使用方式、限制条件等问题，二者的法律适用应当存在一定程度上的延续与联系。因此，本文将围绕 ChatGPT 训练数据之合理使用展开分析，从比较法视野分析英国、欧盟、美国及中国对 TDM 所制定的合理使用制度，继而分析现行法律框架下 ChatGPT 所实施的数据挖掘行为是否具有合法性依据。

一、ChatGPT 数据挖掘原理与侵权风险

ChatGPT 是一种基于自然语言处理（NLP）的 AI 系统，使用了深度神经网络和自然语言处理技术来生成文本，其工作原理可分为三个阶段：数据输入——机器学习——结果输出。自然语言处理 AI 的训练数据通常由大量文本数据组成，当中包含了语言的各种形式和用法。

ChatGPT 训练数据的使用流程 [2]：

---

[1] "IFLA Statement on Text and Data Mining (2013)." IFLA, www.ifla.org/publications/ifla-statement-on-text-and-data-mining-2013/. Accessed 22 Apr. 2023.

[2] "ChatGPT and Data Annotation." 23 Feb. 2023, labelyourdata.com/articles/data-annotation-for-training-chatgpt. Accessed 22 Apr. 2023.

1. 数据收集：从各种来源收集原始数据。

2. 数据预处理：将原始数据进行清洗和标准化，以便后续处理和分析。

3. 数据标注：将数据进行标注，为机器学习提供训练数据。

4. 特征提取：从标注好的数据中提取特征。

5. 模型训练：对训练数据进行分析和学习。

6. 结果生成：输出生成物。

ChatGPT 的训练过程中涉及到大量文本数据的使用。尽管 OpenAI 并未公开当前版本 ChatGPT 所使用的 GPT-3.5 语言模型数据量，但从公开数据来看，GPT-3 语言模型由 1750 亿参数训练而成，由此迭代而来的 GPT-3.5 语言模型显然需要更庞大的数据量作为支撑 [3]。

ChatGPT 主要依赖于两种文本数据源，即用户输入内容和训练数据库。关于用户输入内容，根据《使用条款》规定，用户输入的内容将作为 ChatGPT 学习的文本数据之一。如果用户不同意此使用方式，可以通过邮件等方式拒绝授权而不会影响其正常使用 [4]。关于 ChatGPT 的训练数据库，其数据来源可大致分为三种：第一种，来源于公有领域的内容。公有领域内容指的是不属于私人所有，任何人可以不受限制地使用和加工的数据，包括本身便不受法律保护的内容及已过著作权保护期间进入公有领域的内容；第二种，通过签订合同获得合法授权的内容，即通过与权利人签订合同从而获得有效授权，合法使用相关数据及内容；第三种，未经授权的信息及内容。该来源指的是数据及内容本身为受著作权保护的客体，而 ChatGPT 在未经授权的情况下对相关内容进行挖掘使用，其获取渠道通常为利用“爬虫”技术获取网络数据及内容、非法获取数据库内容以及未经许可数字化非电子数据内容等方式。通过上述方式所构建的训练数据库，由于涉及未经授权使用受著作权保护的数据及内容，因此天然具有著作权侵权风险。

在我国现行《著作权法》框架下，ChatGPT 训练数据使用过程的不同行为均可能存在著作权侵权风险。首先，在数据内容收集阶段，训练数据的使用或构成复制权侵权。数据收集的方式有两种形式，分别是将非数字格式的原内容转化为计算机可读的数据格式，即“原件扫描”，或是对他人已有数据进行访问和获取文本内容。训练数据的输入过程必然伴随着相应的复制行为。目前学界认为，ChatGPT 数据挖掘过程中的复制行为不属于因数字环境传输中“暂时的”、“在技术过程中必然发生的”，且“不具有独立经济价值”

[3] "GPT-4 Is Coming – What We Know So Far." Forbes, Bernard Marr, 24 Feb. 2023, www.forbes.com/sites/bernardmarr/2023/02/24/gpt-4-is-coming--what-we-know-so-far/?sh=11045dd86c2d. Accessed 7 Apr. 2023.

[4] "Introducing ChatGPT." OpenAI, openai.com/blog/chatgpt. Accessed 22 Apr. 2023.

的“临时复制”情形，因此除非存在法定豁免情形，否则 ChatGPT 数据内容挖掘行为可能构成复制权侵权。实际上，ChatGPT 在数据挖掘过程中，往往并非只将数据短暂复制于系统中，而是需要将作品数据长时间停留，继而便可能涉及到对作品的永久性复制。尽管当前各国对“临时复制”的法律性质存在争议，但对永久性复制应当归入复制权规制范围却存在共识。

其次，在数据预处理阶段，训练数据的使用或构成演绎权侵权。演绎权指的是在原作品创作思想表达的基础上演绎创作新作品的权利。我国《著作权法》并未采纳“演绎权”这一术语，而是在演绎权的概念上进一步分解为翻译权、改编权、汇编权等权利。但归根结底，演绎权及其分解而来的权利核心在于原作品的主要思想表达并未因创作语言、题材、种类或形式的变化而改变。而 ChatGPT 在数据预处理阶段，涉及对所收集数据进行清洗、标准化、标注与特征提取等步骤，存在侵犯改编权等权利的风险。

最后，在结果生成阶段，训练数据的使用也可能构成与传播相关的权利侵权。因 ChatGPT 会依据训练数据自动化生成结果，并以可视化方式表现，过程中需要将数据或文本通过互联网进行传输，从著作权法角度显然也会涉及侵权。

ChatGPT 数据挖掘行为本身具有高度复杂性，在所涉著作权内容不同的情形下，其侵犯的权利也会有所不同，上述仅对可能涉及的侵权风险作非穷尽式列举。关于 ChatGPT 在实际应用场景下的侵权形态与风险，应作个案具体分析。

二、域外视角下的 TDM 合理使用制度——英国、欧盟、美国与中国之比较

著作权合理使用制度，即在符合特定条件情形下，法律允许他人可自由使用受著作权保护的客体而无须经由著作权人同意，抑或是向著作权人支付相应报酬，为著作权限制与例外的核心制度之一。当 ChatGPT 训练数据未经授权使用受著作权保护内容，并且涉及到作者的独创性表达及造成公众传播效果时，便可能构成著作权侵权。此时需要进一步判断其是否构成合理使用。合理使用的制度目的在于平衡著作权人对作品享有的专有权利与公众获取作品的需求，促进创新和文化多样性发展，保障公众基本利益。就生成式 AI 训练数据的合理使用规则而言，大多数国家尚未单独制定成文法规定。但各国针对 TDM 所制定的合理使用规则，对当下 ChatGPT 数据挖掘行为的合理使用制度适用具有重要指引作用[5]。

（一）英国

作为最早制定《版权法》的国家之一，英国是合理使用制度的开创国，也是最早对

[5] 除本文所列举的英国、欧盟与美国外，日本与新加坡同样对 TDM 制定了版权例外规则。日本 TDM 版权例外规则详见《著作權法》：https://www.cric.or.jp/english/clj/cl2.html；新加坡 TDM 版权例外制度详见《COPYRIGHT ACT 2021》：https://sso.agc.gov.sg/Acts-Supp/22-2021/Published/。

TDM 通过立法方式确定其合法性的欧洲国家。2014 年修订的《版权法》新增了第 29A 条 TDM 版权例外规则条款，当中允许为了非商业性研究的文本和数据挖掘目的，利用计算机分析技术对已经合法获得访问的任何版权材料进行复制[6]。可见英国通过立法形式明确为 TDM 应用赋予了合法性，以防止版权成为阻碍相关技术创新发展的阻力。不过值得注意的是，英国同样为相关立法设置了许多限制条件。英国虽未对行为主体设定限制，却在客体上将 TDM 合理使用范围限定为“合法获得访问的版权材料”，即行为人本身应当具备合法访问相关版权材料的资格。英国也对“使用目的”作出了限制，规定只有基于“计算机分析”和“非商业性使用”目的的 TDM 属合理使用范围，即排除了不以计算机处理、分析数据为目的的行为及具有盈利性质的商业性使用。具体来看，ChatGPT 的技术厂商 OpenAI 最初的定位虽为非营利组织，但其自 2019 年起便开始转型为营利性组织，ChatGPT 的数据挖掘、使用行为难以被定性为“非商业性使用”。在使用行为上，ChatGPT 训练数据的使用过程可能涉及版权意义上的多个行为，包括复制、改编与传播。而英国《版权法》第 29A 条只针对 TDM 的复制行为提供了合法性支持，对其他行为则未设置侵权豁免，因此相关行为仍存在侵权风险。除上述条件外，英国也规定了权利限制的例外情形，指出未经版权人授权将 TDM 过程中产生的复制件进行交易，包括出售、出租、许可等行为仍会构成侵权。

整体而言，英国《版权法》给予 TDM 一定的实施空间，同时也兼顾了版权人的利益。但由于 ChatGPT 等生成式 AI 在数据挖掘与使用行为上所具有的复杂性，其在英国《版权法》框架下未必能够适用合理使用规则。

（二）欧盟

欧盟委员会在 2016 年 9 月公布了《欧盟数字化单一市场指令》提案，随后于 2019 年 3 月通过了《数字化单一市场版权指令》（“《版权指令》”），对 TDM 的使用制定了版权例外规则。

---

[6] Copyright, Designs and Patents Act 1988, 29A: Copies for text and data analysis for non-commercial research (1)This section applies where— (a)a person has lawful access to a copy of a copyright work for the purposes of research to which this section applies, and (b)the copy is retained by the person on a secure electronic network for the purposes of carrying out text and data analysis for those research purposes. (2)The making of a copy of the work by the person who has lawful access to the work under subsection (1) does not infringe copyright in the work provided that— (a)the copy is made by an automated process, (b)the copy is used only for the purposes mentioned in subsection (1)(b), and (c)the person satisfies the other conditions in this section. (3)The other conditions are that— (a)the person making the copy has reasonable grounds for believing that doing so is necessary for the purposes of the research mentioned in subsection (1)(b), (b)the person does not use the copy to compete with the owner of the copyright in the work, (c)the person does not supply the copy to any other person except for the purposes mentioned in subsection (1)(b), and (d) the use of the copy is accompanied by a sufficient acknowledgement. (4)This section applies to research for any purpose, except for commercial purposes.

欧盟《版权指令》第 3 条 7	第 3 条以科学研究为目的的文本和数据挖掘： 1. 成员国应当规定，科研机构和文化遗产机构为科学研究目的进行文本和数据挖掘，对其合法获取的作品或其他内容进行复制与提取的行为，属于 96/9/EC 指令第 5 条 (a) 项与第 7 条第 1 款，2001/29/EC 指令第 2 条以及本指令第 15 条第 1 款所规定的权利的例外。 2. 第 1 款所规定的作品或其他内容的副本应以适当的安全等级储存，可保留作科学研究之用，包括为验证研究结果之用。
欧盟《版权指令》第 4 条 8	第 4 条文本和数据挖掘的例外或限制： 1. 成员国应规定，以文本和数据挖掘为目的，对合法获取的作品或其他内容进行复制与提取的行为，属于 96/9/EC 指令第 5 条 (a) 项与第 7 条第 1 款，2001/29/EC 指令第 2 条，2009/24/EC 指令第 4 条第 1 款 (a) 和 (b) 项，以及本指令第 15 条第 1 款所规定的权利的例外 2. 以进行文本和数据挖掘为目的，根据第 1 款复制和提取的作品或其他内容可保留到必要时为止。

《版权指令》第 3 条、第 4 条规定，基于“科学研究”与“数据分析”两种目的，并且作品为合法获取的情形下 TDM 具有正当性。从法律条文来看，欧盟同样通过封闭性规范的方式将 TDM 列入了合理使用范围，并且对 TDM 版权例外制度的适用设定了相应限制条件。针对以“科学研究”为目的的 TDM，欧盟将主体限制为科研及文化遗产机构。换而言之，由于 ChatGPT 的发行厂商 OpenAI 不符合相关主体要求，因此不能适用《版权指令》第 3 条所规定的版权例外规则。而针对以“数据分析”为目的的 TDM，《版权指令》第 4 条实际上赋予了 TDM 在数据处理阶段使用行为的合法性，该例外不存在主体限制或使用技术目的限制，即使是出于商业性使用目的也同样适用。在客体条件上，欧盟与英国都将其限定为以合法途径获取的作品。ChatGPT 在数据收集、预处理阶段可能涉及多种权利侵权，而该条中只为过程中的复制、提取行为规定了版权例外，但对其他类型行为则未明确说明。

整体而言，在欧盟《版权指令》合理使用制度下，行为人享有作品的阅读权即可享有作品的挖掘权，但需要为副本的保存提供安全措施。

（三）美国

不同于英国、欧盟以立法形式明确将 TDM 列入合理使用范围，美国采用了基于四要素标准与司法判例为指导的开放性立法。美国《版权法》第 107 条规定了合理使用制度，即以四要素作为判断标准，通过综合分析使用者的使用是否符合相关法定要素来判断该使

[7] DIRECTIVE (EU) 2019/790 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 17 April 2019，article 3.
[8] DIRECTIVE (EU) 2019/790 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 17 April 2019，article 4.

用是否为合理使用。具体而言，四要素标准指的是：（1）使用目的和性质。第一要素包括该使用是商业性使用或者是基于非营利性教育目的之使用。一般而言，若被告对版权作品的使用为商业性使用，则不属于合理使用；但若使用是基于非营利性教育目的，则更有可能被视为合理使用。但自 Campbell 案[9]后，美国法院认为在商业性使用能明显增进社会效益时，应当以“转换性使用”作为判断标准而忽略商业性使用目的。而“转换性使用”的内涵是对原创作品进行某种程度上的改编、转化或转换，以产生新的表达形式、意义或价值的行为。（2）版权作品的性质。法院在分析第二法定要素时，需要考虑的是究竟被使用的作品是描述事实的叙事作品或创作性很强的虚构作品。通常来说，作品的创作性越强，就越应受到法律保护。（3）被告的使用占版权作品的数量和质量。第三个法定要素要求对使用行为不仅要进行定量分析，还要进行定性分析。（4）被告的使用对版权作品市场的影响。第四个法定要素的重点在于，若被告的使用行为减少了版权人的收益，则被告的使用可能会被认定为不合理的使用[10]。

基于四要素标准，ChatGPT 对训练数据的使用有相当可能性被认定为转换性使用，继而受合理使用制度保护。事实上，从司法判例来看，美国法院整体也对 TDM 持相对开放的立场，当中最具代表性的案件便是美国“谷歌图书馆”案与“TVEyes”案。

1. 谷歌图书馆案（Authors Guild v. Google Inc）[11]

美国“谷歌图书馆”案是由谷歌图书馆计划引发的全球诉讼系列案件之一，谷歌公司在未经授权的情况下将图书数字化并收录到其搜索引擎中，这些书籍包括已出版的和未出版的作品。

2013 年，美国纽约地区法院对此案作出一审判决，认为谷歌扫描图书的行为构成合理使用，不构成版权侵权。2015 年 10 月，美国第二巡回法院确认了一审法院的判决，认为谷歌图书馆属于合理使用，不构成侵权。

在论及谷歌图书馆对原告作品的使用目的时，美国第二巡回法院认为谷歌未经授权将受版权保护的书籍进行电子扫描、设立搜索功能并将上述书籍的片段在网络上显示的行为属于非侵权式的合理使用。谷歌所采取的扫描行为是高度转换性的，其显示的文字数量是有限的，而向公众提供的只言片语也不会与原作品构成竞争或替代关系。因此，即便谷歌是一家追求利润的商业公司，这也不妨碍认定谷歌图书馆的行为属于合理使用[12]。

---

[9] Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569 (1994).
[10] 宋海燕：《娱乐法（第二版）》，商务印书馆 2018 年版，第 90-95 页。
[11] Authors Guild v. Google, Inc., 4F., 3d 202,209(2015).
[12] 宋海燕：《娱乐法（第二版）》，商务印书馆 2018 年版，第 100-101 页。

2. TVEyes 案（Fox News Network, LLC v. TVEyes, Inc）[13]

在“TVEyes”案中，TVEyes 公司提供的搜索引擎可以对主流电视节目进行实时监控和搜索，用户可以根据关键词检索快速定位到感兴趣的节目，并可查看不超过 10 分钟的节目剪辑内容，也可以对相关片段进行保存、下载。

福克斯新闻公司在 2014 年向法院提起诉讼，指控 TVEyes 公司侵犯其版权，并要求 TVEyes 停止提供相关服务。此后，其他电视网站也加入了诉讼行列。

最终，美国第二巡回法院认可 TVEyes 将大量电视节目片段复制归档，并向用户提供关键词搜索等服务的使用行为具有变革性，对原作品构成转换性使用。但却同时认为 TVEyes 允许用户对相关节目片段进行查看与下载并不合理，将可能对版权人的市场地位与许可收入造成实际损害。最终法院判决 TVEyes 对相关作品的传播构成版权侵权。

上述两个案件展现了美国司法裁判中对 TDM 合理使用的整体态度。在美国“谷歌图书馆”案中，美国法院认定谷歌公司基于向公众提供搜索和片段浏览服务目的而对原告作品进行全文复制的行为具有“目的转换性”，强调司法实践中应当对四要素标准作综合考量。“TVEyes”案则显示出在法院已将 TDM 前期阶段的使用行为认定为合理使用的情形下，倘若相关技术实施者未采取必要技术以降低对原作品权利人的替代性影响，仍可能存在侵权风险。

相较于美国“谷歌图书馆”案与“TVEyes”案中原告的使用行为，ChatGPT 经过对训练数据的学习而生成结果的使用行为显然更具有“转换性使用”意义。因此，在美国《版权法》合理使用制度框架下，ChatGPT 的数据挖掘行为有相当可能性得以构成合理使用。

（四）中国

区别于英国、欧盟与美国为 TDM 制定了合理使用规则，TDM 目前尚未被涵盖在我国《著作权法》第 24 条所列举的 12 种法定著作权例外情形中。换而言之，当前国内的著作权例外制度无法为 TDM 的实施提供合法性依据。

我国《著作权法》第 24 条规定了 12 种合理使用情形。然而，ChatGPT 对训练数据的使用难以被该 12 种法定情形所保护。ChatGPT 的数据挖掘行为并非为“个人学习”“教学或科研”“公共文化机构”所使用，且本质上属商业性使用，难以直接援引该条作为侵权抗辩。因此，值得进一步讨论的便是 ChatGPT 的数据挖掘行为能否落入《著作权法》第 24 条兜底条款的保护之中。从案例来看，兜底条款实际上同样难以为 ChatGPT 的数据挖掘行为提供法律依据，在部分案件中可见国内法院对 TDM 著作权侵权问题的整体态度。

---

[13] Fox News Network, LLC v. TVEyes, Inc., 883 F.3d 169, 179 (2d Cir. 2018).

1. A 作者诉 B 公司数字图书馆案 [14]

本案中，原告 A 作者是某书籍的作者及著作权所有人。被告 B 公司获得了涉案作品的纸件版本并将涉案书籍进行扫描。随后，B 公司将扫描的图书片段开放给旗下搜索引擎，从而使互联网用户从搜索结果中看到涉案作品的片段。

在讨论 B 公司数字图书馆的扫描书籍及通过搜索展示书籍片段的行为是否构成著作权的“合理使用”时，法院首先指出，B 公司的涉案复制行为并不属于《著作权法》（2010）第 22 条规定的合理使用行为，故应初步推定为构成侵权。但随后又提出，鉴于实际的司法实践中，法院已在部分案例中认定《著作权法》（2010）第 22 条规定之外的其他特殊情形也可构成合理使用，故如果 B 公司能够主张并证明其涉案复制行为属于合理使用的其他特殊情形，那么该行为也可被认定合理使用。

关于如何判断涉案的复制行为是否构成《著作权法》（2010）第 22 条规定之外的合理使用特殊情形时，法院提出应综合考虑以下相关因素，包括（1）使用作品的目的和性质；（2）受著作权保护作品的性质；（3）所使用部分的性质及其在整个作品中的比例；以及（4）被告的使用行为是否影响了原告作品的正常使用或不合理地损害著作权人的合法利益等。在综合考虑了上述因素之后，法院认为在本案中，B 公司未能针对上述因素提交相关事实证据，故驳回 B 公司关于合理使用的抗辩，认为其图书馆行为构成侵权 [15]。

从中美类案判决的对比来看，在面对相似案情与抗辩理由时，两国法院在判断相关行为是否构成合理使用时得出了相反结论。在中国“A 作者诉 B 公司数字图书馆”案中，法院认为在《著作权法》（2010）第 22 条规定的具体情形外认定合理使用，应当从严掌握认定标准，而被告应当对考量因素中的事实问题承担举证责任。但该案中被告 B 公司并未充分举证其涉案行为属合理使用，因此法院最终推定其使用行为构成侵权。由此可见，依据当前《著作权法》相关规定，ChatGPT 的数据挖掘行为被中国法院认定为合理使用具有难度，仍存在较高侵权风险。

上述观点也可以从 2023 年 4 月 11 日国家互联网信息办公室起草的《生成式人工智能服务管理办法（征求意见稿）》[16] 中得到印证。该征求意见稿回应了公众关注的生成式 AI 若干问题。其中第 7 条对生成式 AI 产品 / 服务的数据来源作出规制，明确规定提供者应当对生成式人工智能产品 / 服务的预训练数据、优化训练数据来源的合法性负责，不应

---

[14] （2013）高民终字第 1221 号。
[15] 宋海燕：《娱乐法（第二版）》，商务印书馆 2018 年版，第 101-102 页。
[16] 2023 年 7 月，国家网信办联合国家发展改革委、教育部、科技部、工业和信息化部、公安部、广电总局正式公布《生成式人工智能服务管理暂行办法》，并自 2023 年 8 月 15 日起施行。

含有侵犯知识产权的内容[17]。换而言之，该征求意见稿并未对 TDM 设置著作权例外，一定程度显示出我国立法动向仍对认可 TDM 合理使用持保留态度。

结语

信息获取与知识共享是数字经济的基础。随着人工智能领域高速发展，在可预见的将来仍会不断涌现涉及他人著作权作品的新型使用行为。当前部分国家已对数据挖掘、使用行为设定了著作权例外制度，尝试在科学技术的发展与著作权人的利益保障之间取得平衡。我国《著作权法》目前尚未对数据挖掘制定著作权例外规则，相关技术在中国的实施仍具有侵权风险。但数据挖掘作为人工智能时代的基础性技术，合理使用规则的缺失必然会限制信息自由流动与创新发展。为了促进科技领域发展，我国应当保障数据挖掘技术的流通与应用，平衡著作权人利益与公共利益的冲突，建构旨在驱动创新的合理使用制度。

感谢资深顾问王冬梅对本文作出的贡献。

[17] “国家互联网信息办公室关于《生成式人工智能服务管理办法（征求意见稿）》公开征求意见的通知 - 中共中央网络安全和信息化委员会办公室”中共中央网络安全和信息化委员会办公室 , https://www.cac.gov.cn/2023-04/11/c_1682854275475410.htm，最后浏览日期：2023 年 4 月 21 日。

谈 AIGC 的可版权性——美国、欧盟、英国与中国之比较

宋海燕 陈玮聪

前言

近期以来，AI 技术的高速“狂飙”，掀起了一波又一波 AIGC 相关应用的热潮。不论是 ChatGPT[1] 对待人类花样提问的灵活自如应答，还是 Midjourney[2] 根据人类的特定指令，生成的一幅幅令人惊叹不已的图像，均属于 AIGC 的范畴，对人类社会造成了变革性的冲击。

所谓 AIGC，即 AI Generated Content，是指利用人工智能技术来生成内容。AIGC 从最初面世以来，其是否属于版权法保护的客体，能否受到版权法的保护，这一问题便引起了社会各界的热烈讨论。

近期，美国艺术家克里斯蒂娜·卡什塔诺娃（Kristina Kashtanova）的漫画《Zarya of the Dawn》因涉及部分 AIGC 内容，在版权注册申请过程中面临了尴尬的遭遇。基于此，美国版权局于 2023 年 3 月 16 日发布了一份指南，强调必须由人类创作的作品才能获得版权保护。美国版权局对于 AIGC 版权问题的态度，让有关 AIGC 版权保护的探讨有了更进一步的发展。

本文以科幻漫画《Zarya of the Dawn》的注册历程为切入点，介绍了美国版权局对于人工智能生成物的认定态度，并通过域内外对比研究，介绍了欧盟、英国、中国关于人工智能生成物可版权性的最新司法实践和相关政策情况。

一、《Zarya of the Dawn》的尴尬遭遇及美国版权局的应对

美国艺术家克里斯蒂娜·卡什塔诺娃（Kristina Kashtanova，以下简称“卡什塔诺娃”）创作了一本名为《Zarya of the Dawn》的科幻漫画书，该漫画书的部分图像是由人工智能平台 Midjourney 根据卡什塔诺娃的指令创建。作者将该书提交给美国版权局申请注册登记，最初该书获得了版权登记。随后，美国版权局通过卡什塔诺娃的社交媒体发帖获悉，作者在写作该书时，使用人工智能平台 Midjourney 创作了该书的部分内容，美国版权局

---

[1] ChatGPT（全名：Chat Generative Pre-trained Transformer），美国 OpenAI 研发的聊天机器人程序，于 2022 年 11 月 30 日发布。作为人工智能技术驱动的自然语言处理工具，它能够通过理解和学习人类的语言来进行对话，还能根据聊天的上下文进行互动，真正像人类一样来聊天交流，甚至能完成撰写邮件、视频脚本、文案、翻译、代码，写论文等任务。百度百科：《ChatGPT》，载百度百科，https://baike.baidu.com/item/ChatGPT?fromModule=lemma_search-box，最后访问时间：2023.3.22。

[2] Midjourney 是一个由同名研究实验室开发的人工智能程式，可根据文本生成图像，于 2022 年 7 月 12 日进入公开测试阶段，使用者可透过 Discord 的机器人指令进行操作。维基百科：《Midjourney》，载维基百科：https://en.m.wikipedia.org/wiki/Midjourney，最后访问时间：2023.3.22。

随后以该作品缺少人类作者为由拒绝注册，原因是该局将版权法解释为仅仅保护“人类作者身份的作品”，如果该作品集并非由人类创作，那么该局将拒绝注册该作品[3]。

United States Copyright Office

Library of Congress · 101 Independence Avenue SE · Washington DC 20559-6000 · www.copyright.gov

February 21, 2023

Van Lindberg
Taylor English Duma LLP
21750 Hardy Oak Boulevard #102
San Antonio, TX 78258

Previous Correspondence ID: 1-5GB561K

Re: Zarya of the Dawn (Registration # VAu001480196)

Dear Mr. Lindberg:

The United States Copyright Office has reviewed your letter dated November 21, 2022, responding to our letter to your client, Kristina Kashtanova, seeking additional information concerning the authorship of her work titled *Zarya of the Dawn* (the “Work”). Ms. Kashtanova had previously applied for and obtained a copyright registration for the Work, Registration # VAu001480196. We appreciate the information provided in your letter, including your description of the operation of the Midjourney’s artificial intelligence (“AI”) technology and how it was used by your client to create the Work.

The Office has completed its review of the Work’s original registration application and deposit copy, as well as the relevant correspondence in the administrative record.[1] We conclude that Ms. Kashtanova is the author of the Work’s text as well as the selection, coordination, and arrangement of the Work’s written and visual elements. That authorship is protected by copyright. However, as discussed below, the images in the Work that were generated by the Midjourney technology are not the product of human authorship. Because the current registration for the Work does not disclaim its Midjourney-generated content, we intend to cancel the original certificate issued to Ms. Kashtanova and issue a new one covering only the expressive material that she created.

The Office’s reissuance of the registration certificate will not change its effective date—the new registration will have the same effective date as the original: September 15, 2022. The public record will be updated to cross-reference the cancellation and the new registration, and it will briefly explain that the cancelled registration was replaced with the new, more limited registration.

[1] The Office has only considered correspondence from Ms. Kashtanova and her counsel in its analysis. While the Office received unsolicited communications from third parties commenting on the Office’s decision, those communications were not considered in connection with this letter.

2023.2.21 美国版权局对于漫画《Zarya of the Dawn》的注册回应

然而，在结合作者的意见进一步考虑之后，美国版权局推翻了之前的决定，并于

[3] Zarya of the Dawn (VAu001480196) at 2 (Feb. 21, 2023), https://www.copyright.gov/docs/zarya-of-the-dawn.pdf. last visit：3/24/23.

2023 年 2 月 21 日发布了一项新的决定，准许漫画《Zarya of the Dawn》的整体注册，但缩小了注册范围，以明确排除通过人工智能技术生成的材料部分（即 Midjourney 根据卡什塔诺娃的提示指令自动生成的图像）。新的注册范围仅涵盖卡什塔诺娃在写作本书时所形成的“作者创作的文字和对人工智能生成的作品的选择、协调和安排”，而那些由 Midjourney 自动生成的图像则不予保护。

美国版权局做出上述决定的考量因素之一是人类参与创作过程的程度。Midjourney 自动生成的图像，卡什塔诺娃只是为此 AI 系统提供了目标生成图像的提示和参数，这种提示和参数并未证明卡什塔诺娃对 Midjourney 结果的输出有足够的控制权，无法使她有资格成为这些 AIGC 的作者（或合作作者），并认定这些图像不是具有创造性的、人工创作的艺术品，无法获得版权保护。

基于此，美国版权局驳回了使用人工创作的文本提示可以对生成的图像进行版权保护的观点，并指出这些人工智能系统的用户没有人为参与或控制图像的创建，因此，无法受到版权法的保护。

2023 年 3 月 16 日美国版权局发布了《版权登记指南：包含人工智能生成材料的作品》（“版权登记指南”），阐述了其在审查和注册包含人工智能技术生成材料的作品时所采取的做法 [4]。

其中一些要点包括：

- 版权只能保护人类创造力的产物——宪法和版权法中使用的“作者”一词不包括非人类。
- 科技工具可以是创作过程中的一部分，但作品表达的创造性必须是由人类控制的。如果只是 AI 技术根据人类的提示产生作品，则该作品缺乏人类作者身份，不受版权保护。如果人类艺术家以足够有创意的方式选择或安排 AI 生成的材料，以及艺术家修改 AI 生成的材料以符合版权保护标准，使得 AI 生成的作品包含足够的人类作者身份，则可以支持版权主张。
- 对于包含 AI 生成物的作品，美国版权局将考虑 AI 的贡献是“机械复制”的结果，还是包含作者“创造性的想法（智力活动），（由作者）赋予表现形式”的结果。答案将取决于具体情况，特别是 AI 工具如何运作以及作者如何使用 AI 工具创建最终作品。

这份版权登记指南也对申请者提出了版权注册的具体要求，部分内容如下：

---

[4] Copyright Registration Guidance for Works Containing AIi-Generated Material.https://www.govinfo.gov/content/pkg/FR-2023-03-16/pdf/2023-05321.pdf. last visit：3/24/23.

- 申请人有义务披露提交注册的作品中包含人工智能生成的内容，并简要说明人类作者对作品的贡献。例如，将 AI 生成的文本合并到更大的文本作品中的申请人应该声明文本作品中人工创作的部分。
- 如果已经提交申请的作品包含 AI 生成材料，那么申请人需要重新检查是否充分披露了这些材料，以便申请有效。如果未披露，那么申请者需要联系版权局进行补充注册。

美国版权局最后表示，其将持续监测涉及 AI 和版权的新事实和法律发展，并可能在未来发布与注册或该技术涉及的其他版权问题相关的其他指南。

这份版权登记指南，阐明了美国版权局对于 AIGC 的态度。当且仅当 AIGC 具备“作者的创造性想法（智力活动）、（由作者）赋予表现形式”时，才有可能获得版权法的保护。由此可见，美国版权局采取“独创性”为判断依据。

二、欧盟和英国

（一）欧盟

一般而言，根据欧盟版权法，要获得版权保护，必须满足两个条件：（1）创作必须是作品；（2）必须是上述作品的原作者或已通过转让获得版权。

2017 年，欧盟议会法律事务委员会（JURI）向 Commission on Civil Law Rules on Robotics 提出议案，该议案的提出背景是：机器人技术和人工智能技术已经成为本世纪最突出的技术趋势之一，它们的使用和开发的快速增加给我们的社会带来了新的困难和挑战，也引发了与所有这些领域相关的法律和伦理问题，需要欧盟层面的及时干预。其议案涵盖机器人和人工智能一般原则、民用机器人和人工智能发展总则、知识产权和数据流、标准化、安全和保障、自主运输工具、护理与医疗机器人等等多个方面。在该议案“EXPLANATORY STATEMENT”板块下，提及了关乎“AIGC 可版权性”的内容——要求为计算机或机器人制作的受版权保护的作品制定“自主智力创作（own intellectual creation）”的标准[5]。

近年来，面对人工智能空前发展的趋势，2020 年，欧盟委员会又发布了名为“Trends and Developments in Artificial Intelligence – Challenges to the Intellectual Property Rights Framework” 的报告[6]，该报告的整体结论认为人工智能当前的技术发展水平不需要欧洲的版权法和专利法立即发生实质性的变化，版权法和专利法的现有概念、规则足够抽象和灵活，可以应对人工智能当前的挑战，邻接权制度可能会扩展到各个领域

[5] Committee on Legal Affairs, REPORT with recommendations to the Commission on Civil Law Rules on Robotics 27.1.2017 - (2015/2103(INL))：https://www.europarl.europa.eu/doceo/document/A-8-2017-0005_EN.html，last visit: 2023.3.27.

[6] European Commission, Directorate-General for Communications Networks, Content and Technology, Hartmann, C., Allan, J., Hugenholtz, P., et al., Trends and developments in artificial intelligence : challenges to the intellectual property rights framework : final report, Publications Office, 2020, https://data.europa.eu/doi/10.2759/683128，last visit：3/24/23.

“未经授权”的人工智能生成物。

同时，该报告提出了“四步测试法”，即四个相互关联的标准，来判断 AIGC 是否符合“作品”资格：

Step 1 – 文学、艺术、科学领域；

Step 2 – 人类智力活动 ；

Step 3 – 独创性；

Step 4 – 表达。

根据欧盟发布的“四步测试法”，AIGC 能否符合“作品”资格，主要取决于是否满足第二步和第三步，即 AIGC 是否表达了人类的智力活动、是否具有独创性。

（二）英国

英国对于 AIGC 的立场是比较具有前瞻性、突破性的，同时，英国也是为数不多的几个在没有人类创作者的情况下保护由计算机生成作品的国家之一。其早在《1988 年版权、外观设计和专利法案》中便有针对计算机生成物的相关规定[7]：

- “computer-generated”（计算机生成物），是指在不存在任何人类作者的状况下，由计算机运作生成的作品。
- “对于计算机生成的文字、戏剧、音乐或艺术作品而言，作者应是对该作品的创作进行必要安排的人。”（“必要安排”的判断以“实质性贡献”为依据）

因此，根据英国法律，完全由 AI 生成的作品，可能获得版权。值得注意的是，在这种情况下，立法者将这种作品的保护期缩短为 50 年，而由人类作者创作的作品的保护期为 70 年。

2021 年 10 月 29 日起，伴随着人工智能技术的发展，英国知识产权局（UKIPO）公开就人工智能（AI）和知识产权互动以及人工智能对知识产权制度的影响征求意见，为期 10 周，截至 2022 年 1 月 7 日，所获得的信息将为其立法提供信息支持。本次社会咨询旨在就人工智能在专利和版权制度中应如何处理的一系列选择征求证据和意见。这一成果将支持人工智能战略，是英国政府在其最近推出的创新战略中的承诺之一，旨在帮助确保英国的知识产权环境继续领先世界。本次咨询主要聚焦以下三方面问题[8]：

- 没有人类作者的计算机生成作品的版权保护。这些目前在英国受到 50 年的保护。

---

[7] Copyright, Designs and Patents Act 1988.https://www.gov.uk/government/publications/copyright-acts-and-related-laws. last visit：3/24/23.

[8] https://www.gov.uk/government/consultations/artificial-intelligence-and-ip-copyright-and-patents/artificial-intelligence-and-intellectual-property-copyright-and-patents，last visit：3/24/23.

但是它们应该得到保护吗？如果应该，应该如何保护？

- 文本和数据挖掘的版权许可或例外，这在人工智能的使用和开发中通常很重要。
- 人工智能发明的专利保护。我们应该保护它们吗？

同时，该网站咨询页面显示，收到了对计算机生成作品保护的批评意见，这些批评意见认为[9]：

- 从法律的角度来看，计算机生成的作品必须是独创的才能受到保护。但独创性的法律概念是参考人类作者和人格、判断力和技能等特征来定义的。有人争辩说，法律不明确且自相矛盾。
- 从经济的角度来看，一些人认为对计算机生成的作品的版权保护是过度的。这是因为计算机不需要因产生新内容而获得奖励，但知识产权权利对第三方来说是有成本的。他们认为应该取消这种保护或将其限制在必要的最低限度。
- 从哲学的角度来看，一些人认为版权的根源在于人类的作者身份和创造性的努力，应该只适用于人类的创造。他们认为，保护计算机生成的作品可能会以牺牲人类创作为代价来促进这些作品，并贬低人类创造力。

英国知识产权局在 2022 年 6 月对上述人工智能版权和专利的咨询进行了回应[10]，与 AIGC 相关的部分回应如下：

- 对于计算机生成的作品，我们不打算修改法律。目前没有证据表明对计算机生成的作品的保护是有害的，AI 的使用仍处于早期阶段。因此，无法对任何一个（征求意见中所提供的）选项进行适当评估，任何更改都可能产生意想不到的后果。我们将不断审查法律，如果有证据支持，我们可能会在未来修改、替换或取消对计算机生成物的保护。
- 对于 AI 设计的发明，我们现在不打算修改英国专利法。大多数受访者认为人工智能还不够先进，无法在没有人类干预的情况下进行发明。但我们将继续审查这一领域的法律，以确保英国专利制度支持人工智能创新和人工智能在英国的使用。我们将寻求在国际上推进 AI 发明讨论，以支持英国的经济利益。

（三）中国

我国对于作品的相关法律规定主要体现在：

---

[9] https://www.gov.uk/government/consultations/artificial-intelligence-and-ip-copyright-and-patents/artificial-intelligence-and-intellectual-property-copyright-and-patents，last visit：3/24/23.

[10] https://www.gov.uk/government/consultations/artificial-intelligence-and-ip-copyright-and-patents/outcome/artificial-intelligence-and-intellectual-property-copyright-and-patents-government-response-to-consultation，last visit：3/24/23.

著作权法实施条例 (2013 修订)
第 2 条 著作权法所称作品，是指文学、艺术和科学领域内具有独创性并能以某种有形形式复制的**智力成果**。 第 3 条 著作权法所称创作，是指直接产生文学、艺术和科学作品的**智力活动**。为他人创作进行组织工作，提供咨询意见、物质条件，或者进行**其他辅助工作，均不视为创作**。
《北京高院侵害著作权案件审理指南》（2018）
2.1 条规定：审查原告主张著作权的客体是否构成作品，一般考虑如下因素：（1）是否属于在文学、艺术和科学范围内**自然人**的创作；（2）是否具有**独创性**；（3）是否具有一定的表现形式；（4）是否可复制。
《著作权法》（2020）
第 2 条第 1 款规定:“中国公民、法人或者非法人组织的作品，不论是否发表，依照本法享有著作权。”第 2 款至第 3 款规定了外国人、无国籍人的作品受《著作权法》保护的条件。 第 3 条规定：本法所称的作品，是指文学、艺术和科学领域内具有独创性并能以一定形式表现的智力成果。

目前，在我国司法实践过程中，与 AIGC 相关的案例仅有两例，分别是：

1. 中国人工智能生成内容著作权侵权第一案（2018-2019 年）：“北京 A 律师事务所诉 B 公司案”[11]。

一审中，原告 A 律师事务所认为，被告 B 公司未经许可使用法律统计数据分析软件（b 法律信息库）生成的分析报告侵害其著作权；被告 B 公司辩称涉案文章不构成作品，A 律师事务所不享有涉案文章的权利，B 公司亦未向公众提供涉案文章。

本案一审判决书中提及“根据现行法律规定，文字作品应由自然人创作完成。虽然随着科学技术的发展，计算机软件智能生成的此类“作品”在内容、形态，甚至表达方式上日趋接近自然人，但根据现实的科技及产业发展水平，若在现行法律的权利保护体系内可以对此类软件的智力、经济投入予以充分保护，则不宜对民法主体的基本规范予以突破。故本院认定，自然人创作完成仍应是著作权法上作品的必要条件。上述分析报告的生成过程有两个环节有自然人作为主体参与，一是软件开发环节，二是软件使用环节。软件开发者（所有者）没有根据其需求输入关键词进行检索，该分析报告并未传递软件研发者（所有者）的思想、感情的独创性表达，故不应认定该分析报告为软件研发者（所有者）创作完成。同理，软件用户仅提交了关键词进行搜索，应用“可视化”功能自动生成的分析报告亦非传递软件用户思想、感情的独创性表达，故该分析报告亦不宜认定为使用者创作完

[11] 参见：北京互联网法院 (2018) 京 0491 民初 239 号；北京知识产权法院 (2019) 京 73 民终 2030 号。

成。综上，软件研发者（所有者）和使用者均不应成为该分析报告的作者。……分析报告系 b 数据库利用输入的关键词与算法、规则和模板结合形成的，某种意义上讲可认定 b 数据库“创作”了该分析报告。由于分析报告不是自然人创作的，因此，即使 b 数据库“创作”的分析报告具有独创性，该分析报告仍不是著作权法意义上的作品，依然不能认定 b 数据库是作者并享有著作权法规定的相关权利……”

二审法院认为一审法院对上述部分的认定正确，予以确认。

2.“中国 AI 作品第一案”（2019 年）：C 公司诉 D 科技公司侵害著作权及不正当竞争纠纷案[12]。（全国首例认定人工智能生成的文章构成作品的生效案件）

某计算机软件系由原告 C 公司关联企业 E 科技（北京）有限公司自主开发的一套基于数据和算法的智能写作辅助系统。E 科技公司授权 C 公司在许可区域使用上述智能写作计算机软件，并约定运行使用授权软件所创作的作品的著作权归 C 公司所有。自 2015 年以来，C 公司主持创作人员使用该智能写作助手每年可以完成大约 30 万篇作品。2018 年 8 月 20 日，C 公司在某某网站上首次发表了标题为《午评：xxxxx》的财经报道文章（“涉案文章”），末尾注明“本文由某某智能写作助手自动撰写”。被告运营的网站于 2018 年 8 月 20 日发布了标题为《午评：xxxxx》的文章。经比对，该文章与 C 公司在本案中主张权利的涉案文章的标题和内容完全一致。

原告 C 公司认为涉案文章作品的著作权归 C 公司所有，被告的行为侵犯了 C 公司的信息网络传播权并构成不正当竞争。

法院认为：涉案文章是否构成文字作品的关键在于判断涉案文章是否具有独创性，而判断步骤应当分为两步：首先应当从是否独立创作及外在表现上是否与已有作品存在一定程度的差异，或具备最低程度的创造性进行分析判断；其次，应当从涉案文章的生成过程来分析是否体现了创作者的个性化选择、判断及技巧等因素。……从涉案文章的外在表现形式与生成过程来分析，该文章的特定表现形式及其源于创作者个性化的选择与安排，并由智能写作助手软件在技术上“生成”的创作过程均满足著作权法对文字作品的保护条件，法院认定涉案文章属于我国著作权法所保护的文字作品。

综合以上两份判决，A 律所诉 B 公司案中的大数据报告系 b 数据库利用输入的关键词与算法、规则和模板结合形成，而 C 公司诉 D 科技公司案中的网页文章蕴含了创作者团队个性化的选择与安排，满足了我国《著作权法》对于作品的相关规定要件。由此观之，我国司法实践中，对于 AIGC 的版权问题，依然延续“自然人、独创性”的思路，并将其作为判断作品资格的关键因素。

---

[12] 参见：广东省深圳市南山区人民法院（2019）粤 0305 民初 14010 号。

结语

以往全球范围内，与人工智能技术相关的司法实践中，人工智能生成物可版权性的案件相对较少，综合起来致使 AIGC 在立法与司法层面还有许多有待讨论的问题，比如可版权性的认定、人工智能主体资格的确定、人工智能生成物的保护与利益平衡问题等等。但随着近年来 ChatGPT、Midjourney 等人工智能迅速涌入，其生成物对于社会生活产生了翻天覆地的影响，“AIGC”侵权案件不断萌发，权利人版权意识愈发增强，AIGC 版权案件也势必会逐渐增多，理论界与实务界应对其予以密切关注与高度重视。

AIGC 能否成为版权保护的客体？对 AIGC 可版权性的认定一方面可以为与之相关的技术纠纷“定分止争”，另一方面，也会对文学、艺术、科学领域的版权发展起到指引作用。如果关于“可版权性”问题的答复是否定的，那么如何对 AIGC 提供保护，才能更好地实现社会多方的利益平衡？这些问题仍然悬而未决，有待我们进一步讨论……

感谢律师王默、实习生董美孝、林德鑫、陈佩龄对本文作出的贡献。

再论 AIGC 的可版权性——中美司法实践剖析与比较

宋海燕 李梓潼

引言

本文为前文《浅谈 AIGC 的可版权性——美国、欧盟、英国与中国之比较》的承接与更新。过去的一年间，伴随人工智能（AI）技术的飞速迭代，关于人工智能生成内容（AIGC）可版权性的讨论也在持续增加。在层出不穷的实践案例推动下，美国版权局、法院与中国法院对这一问题作出了最新回应，促使 AIGC 可版权性问题的司法进程逐步前进。

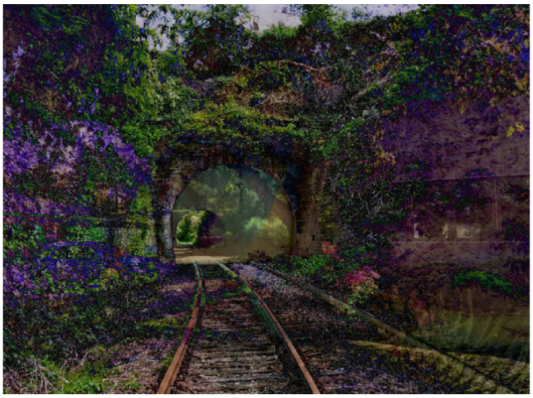
本文将围绕“AIGC 是否为受版权保护的作品（即可版权性）”这一问题，介绍美国当局对此的最新认定意见，并通过分析美国与中国的不同司法实践，探讨对 AIGC 可版权性问题的法律回应。

一、美国

自前文述及的Zarya of the Dawn案之后，美国当局又通过几个拒绝版权登记的案例，进一步实践、强化了其对 AIGC 可版权性的判断标准。下文将按时间顺序，介绍过去一年内美国当局对 AIGC 可版权性的司法回应。

（一）美国哥伦比亚特区地方法院：《A Recent Entrance to Paradise》

计算机科学家 Stephen Thaler 博士利用 AI 程序“Creativity Machine”生成了一幅名为《A Recent Entrance to Paradise》的二维图像，随后 Stephen Thaler 就该图像向美国版权局申请登记。



A Recent Entrance to Paradise

图 1 涉案图片《A Recent Entrance to Paradise》，来源：美国哥伦比亚特区地方法院判决书

Stephen Thaler 在向版权局提交的登记申请中写明：AI 程序“Creativity Machine”是作者，图像系由 AI 算法自主运行生成；申请者作为 AI 的所有人，依据普通法上的财产转移规则与版权法上的雇佣作品规则而获得该生成图像的版权。

对于 Stephen Thaler 的主张，美国版权局不予认同。在两次复议均被驳回后，Stephen Thaler 于 2022 年 6 月起诉美国版权局官员（Shira Perlmutter, Register of Copyrights and Director, U.S. Copyright Office）。2023 年 8 月 18 日，美国哥伦比亚特区地方法院发布判决，维持了美国版权局的决定[1]。判决中涉及 AIGC 可版权性的主要观点为：

- 人类作者身份是版权保护的基本要求。（“Human authorship is a bedrock requirement of copyright.”）
  - 根据美国《1976 年版权法》，原创作品应当由作者（author）或由作者授权完成，而版权法所称之作者仅指人类（human），非人类作者不是美国版权法的创作激励对象。
  - 尽管版权法旨在与时俱进，容纳各类新型表达形式，但对新作品的保护应当基于一个统一的理解，即人类创造力（human creativity）始终是可版权性的核心必要条件。
- 由于 AI 在生成图像时未产生有效的版权，故无须再考虑后续的版权转移问题。
- 法院承认，在 AI 与版权领域仍存在许多悬而未决的问题，包括人类贡献达到何种

[1] THALER v. PERLMUTTER (1:22-cv-01564), https://storage.courtlistener.com/recap/gov.uscourts.dcd.243956/gov.uscourts.dcd.243956.24.0_2.pdf. Last visited on April 7, 2024.

标准后能成为 AI 生成图像的作者、如何划定生成图像的受版权保护范围、如何评估生成图像的原创性、如何以版权制度激励 AI 创作等。目前而言，该判决仅是在特定案情下对版权局决定的维持，上述新问题仍有待回应。

本案明确了美国当局对 AIGC 可版权性的第一个立场，即“完全由 AI 生成的作品因缺乏人类作者身份而不受版权保护”。同时，根据法院判决，人类对 AI 程序是否有所有权并不影响法院对 AIGC 可版权性的判断。

（二）美国版权局：《Théâtre D'opéra Spatial》

游戏设计师 Jason Allen 使用 AI 绘图工具 Midjourney，创作了一幅名为《Théâtre D'opéra Spatial》的二维艺术图像。在美国科罗拉多州举办的新兴数字艺术家竞赛中，《Théâtre D'opéra Spatial》获得“数字艺术 / 数字修饰照片”类别一等奖。随后，Jason Allen 以该图像向美国版权局提交了登记申请，并以自己为作者。

图 2 涉案图片《Théâtre D'opéra Spatial》，来源：美国版权局审查委员会决定

由于该图像曾在公开比赛中获奖，版权局得知其包含 AI 生成内容，要求申请者提交创作过程说明。根据说明，Jason Allen 在生成过程中输入了至少 624 次文本提示得到图像的初始版本，并使用图像处理软件 Adobe Photoshop 消除部分瑕疵、新增视觉内容，最后通过图像处理工具 Gigapixel AI 提高了图像的分辨率和尺寸。

在综合考察案涉图像与两次复议申请后，美国版权局审查委员会（the U.S. Copyright Review Board）于 2023 年 9 月 5 日再次拒绝了《Théâtre D'opéra Spatial》的版权登记申请[2]。版权局审查委员会认为，该图像包含的 AI 生成材料超过了最低限度（more than de minimis），不得整体登记为作品。具体而言：

- **对于申请者使用 Photoshop 增加图片内容的行为**，审查委员会认为目前缺乏充

[2] Re: Second Request for Reconsideration for Refusal to Register Théâtre D’ opéra Spatial (Sep. 5, 2023), https://www.copyright.gov/rulings-filings/review-board/docs/Theatre-Dopera-Spatial.pdf. Last visited on April 7, 2024.

足的信息用以判断手动增加的内容是否达到版权保护的标准。如继续申请，需要补充有关说明。（根据审查委员会的观点，倘若申请者仅使用了 Photoshop 的锐化、色彩平衡、AI 填充等功能，则其手动增加的内容不能体现人类创造性。）

- **对于申请者使用 Gigapixel AI 提高像素的行为**，审查委员会认为该行为没有在图像中引入新的原创要素，因此不能体现人类创造性。
- **对于申请者输入至少 624 次文本提示的行为**，审查委员会认为该行为不足以使申请者成为图像的作者。尽管申请者输入了大量提示词，但最终的生成结果仍取决于 Midjourney 系统如何处理人类的提示。
- 审查委员会认为，AI 无法理解人类的语法和句子结构，不会将提示词理解为创造特定表达结果的具体指令，人类往往需要进行多次迭代方能选择理想的图像。在数百次迭代过程中，难以认为人类对 AI 生成结果有控制作用，发挥主要功能的依然是 AI 技术。
  - 这一观点与美国版权局此前发布的指南相符：当 AI 系统仅接收到人类用户的提示词，并依此生成复杂的文字、视频、音频时，“传统作者元素”（traditional elements of authorship）是由技术决定和执行的，而非人类用户。（《版权登记指南：包含人工智能生成材料的作品》[3]）
  - 此外，审查委员会承认部分提示词可能具有足够的人类创造性，可以单独作为文学作品受到版权保护。但这不意味着向 AI 提供文本提示就“实际形成”（actually form）了生成的图像，申请者的行为不对图像元素构成“创意控制”（creative control）。

本案显示了美国当局对 AIGC 可版权性的第二个立场，即“对于基于 AI 生成的作品，若其中人类参与创作的程度低于标准，则其整体不受版权保护”。

- 该标准可具体阐释为：作品表达的创造性必须由人类控制，人类艺术家应以足够有创意的方式选择或安排 AI 生成的材料，修改 AI 生成的材料以符合版权保护标准，使得 AI 生成的作品具备“作者创造性的想法（智力活动），（由作者）赋予表现形式”，则可以支持版权主张。（《版权登记指南：包含人工智能生成材料的作品》）

（三）美国版权局：《SURYAST》

律师兼艺术家 Ankit Sahni 使用 AI 工具“RAGHAV”[4]，以其拍摄的落日图片为底

[3] Copyright Registration Guidance for Works Containing Ali-Generated Material, https://www.govinfo.gov/content/pkg/FR-2023-03-16/pdf/2023-05321.pdf. Last visited on April 7, 2024.
[4] RAGHAV 是由 Ankit Sahni 资助研发的人工智能绘画程序，Sahni 为该人工智能程序的所有人。

稿，参照梵高的美术风格，生成了一幅名为《SURYAST》的二维艺术图像。随后，Ankit Sahni 以该图像向美国版权局提交了版权登记申请，以自己和 AI“RAGHAV”为合作作者[5]。

图 3 涉案图片《SURYAST》，来源：美国版权局审查委员会决定

申请者 Ankit Sahni 在向版权局提交的创作过程说明中表示：他拍摄了 AI 生成图像所参照的底稿 (the “base image”)，指示 AI 以梵高的《星夜》为图像的风格参照 (the “style image”)，并手动选定了风格强度变量。在考虑谁为作者时，Ankit Sahni 认为他与 AI 对生成图像的贡献是互不相同、各自独立的，因此申请与 AI 共同作为合作作者。

历经初次决定和二次复议后，2023 年 12 月 11 日美国版权局再次作出拒绝《SURYAST》版权登记的决定[6]。基于下述原因，版权局审查委员会认为《SURYAST》未包含足够的人类作者身份（human authorship）：

- **从法律背景上看**：《版权法》的保护对象是“固定在任何有形表达载体上的原创作品（works of authorship）”。在先前判例 Thaler v. Perlmutter 案中，法院也认定“原创作品”应为人类创作的作品。故此，版权局只保护人类作者的创作。
- **从作品认定的法律标准上看**：根据“RAGHAV”的生成过程，“RAGHAV”的作用是从底稿与风格图像中学到的特征生成新的图像，因此不能同摄像作品一样获得保护；尽管申请者声称其从无限排列的结果中选定了一张图像，但这仅体现出了他的想法——将照片变为《星夜》风格的图像，而这种思想不受版权法保护。
- **对于申请者选择风格强度变量的行为**，版权局认为这是一种不受保护的最低作者身份（de minimis authorship）。

[5] 除美国版权局外，Ankit Sahni 还向印度版权局、加拿大版权局提交了版权登记申请。
[6] Re: Second Request for Reconsideration for Refusal to Register SURYAST (Dec. 11, 2023), https://www.copyright.gov/rulings-filings/review-board/docs/SURYAST.pdf. Last visited on April 7, 2024.

对《SURYAST》版权登记申请的多次拒绝，显示了美国版权局对 AIGC 可版权性一如既往的审慎态度。

值得注意的是，同样面对人类作者使用 AI 生成图像是否受版权保护的问题，中美两国当局在相似案情中得出了相反结论，中国法院的判决理由将在下文阐述。

（四）美国产业界对 AIGC 可版权性的不同声音

对于美国版权局近于严苛的版权登记审查标准，美国产业界也提出了一些质疑的声音。

2023 年 8 月 30 日起，为应对日益增长的 AI 技术纠纷，美国版权局就“版权和 AI”向公众征求意见。在提交的回应评论中，美国电影协会（Motion Picture Association，以下简称“MPA”）呼吁版权局改进对 AIGC 的登记评估方法[7]。

与美国版权局相同，MPA 认为，完全由 AI 生成的作品不符合版权保护的条件。但与版权局相左的是，MPA 认为，人类使用 AI 生成的作品具有较为广泛的可版权性，此前版权局在《Zarya of the Dawn》和《Théâtre D'opéra Spatial》的登记审核中采用的标准过于严格。MPA 进一步指出，如果作品能够反映作者的创造性输入（creative input）和原创智力概念（original intellectual conceptions），那么该作品则可以受到版权保护。MPA 希望美国版权局在审核 AIGC 的可版权性时，将对人类作者身份的审查重心转移至“作品是否反映了人类的原创智力概念”这一更具灵活性的标准，而不要过于关注“输出结果的可预测性以及人类对其的控制程度”。

MPA 最后提出，在电影行业中，制作者已开始使用 AI 辅助进行后期制作，如色彩校正、细节优化、背景增强、添加特效等，由此产生的作品应当受版权法保护。

二、中国

2023 年 11 月 27 日，北京互联网法院对李某诉刘某著作权纠纷案作出生效判决，这是国内首例认定“AI 文生图”具有可版权性的案件。

（一）北京互联网法院对 AIGC 可版权性的认定意见

本案中，原告李某使用 AI 绘图模型 Stable Diffusion 生成了一幅人物图像，随后以《春风送来了温柔》为名发布于其社交平台[8]。

原告李某声称，其通过构思布局、输入约 150 个提示词、安排提示词的顺序、设定并不断修改参数、选定最终图像等方式创作了该图像，而被告刘某将其作为文章配图发布并截去水印，侵害了原告的署名权和信息网络传播权。

---

[7] Before the U.S. COPYRIGHT OFFICE LIBRARY OF CONGRESS Washington, D.C., Motion Picture Association, https://www.motionpictures.org/wp-content/uploads/2023/12/2023.12.06-MPA-Reply-to-CO-NOI-2.pdf. Last visited on April 7, 2024.

[8] 参见（2023）北京互联网法院“AI 文生图”著作权案判决书，(2023) 京 0491 民初 11279 号。

在判决书中，法院肯定该 AI 生成图像体现了原告李某的智力投入和个性化表达，具备“智力成果”和“独创性”要件，属于作品，受《著作权法》保护。同时，法院认为 AI 无法成为《著作权法》规定的作者，因此该作品的作者为原告李某。

（二）中美认定意见之比较

面对 AIGC 的可版权性问题，中国与美国当局存在一定的共识。两国法院在判决中均明确，AI 不得作为作品的作者，版权法只保护人类的作品。在判断人类在 AI 生成中的参与程度时，两国都采取了同样的判断思路，即通过分析人类在 AI 生成中的过程性行为，如输入提示词、编排内容、选择最终呈现图像等，并结合最终呈现的图像样态，判断生成图像是否能够反映出人类的原创性智力成果。

但在一致思路下，中美两国进行认定的宽严标准不尽相同，具体而言：在《Théâtre D'opéra Spatial》案中，尽管申请者输入了 624 个提示词，并通过软件对图像进行了后期处理，美国版权局仍认为该图像不符合受版权法保护的标准。在此，美国版权局采取了相对机械、具象性的标准，更加关注生成结果的可预测性与人类对生成结果的控制程度，并要求对图像中的“AI 生成成分”与“人类创作成分”作出严格精细的划分。而中国法院在“AI 文生图”案中，对 AIGC 的可版权性采取了更为灵活、意象性的标准，只要图像的生成过程与最终呈现能够体现人类的智力投入与表达，则可受版权法保护。

结语

现阶段，可能存在争议的 AIGC 类型有以下三种，按照人类在创作过程中的参与程度由低到高排序为：

- **完全或绝大部分由 AI 生成的内容**（AI-Generated Output, with NO or minimal human input）：由 AI 自主运行生成，人类未参与生成过程，或在生成过程中的干预程度极低。在各国司法实践中，均不承认该类型内容具有可版权性，不应当受版权法保护。
- **基于 AI 生成的内容**（AI-Based Output, with certain degree of human input）：人类部分参与了 AI 生成过程。该类型内容目前处于版权法保护的“灰色地带”，在各国司法实践中具有较高的争议性，需要根据具体案情进行个案分析。
- **AI 辅助生成的内容**（AI-Assisted Output, where AI is primarily used as a tool and human creations dominate the creation process）：AI 仅作为生成工具，人类的创意主导创作过程。此类型内容具有可版权性，应当受到版权法保护。

在此基础上，如承认 AIGC 的可版权性，则需要进一步对作品的作者归属作出回应。目前，在版权法领域，尚未有国家承认 AI 可以作为单独作者，且只有加拿大版权局、印度版权局登记了以 AI 为合作作者的作品[9]。参考专利法领域，美国专利商标局于 2024 年 2 月 13 日发布的《人工智能辅助发明的发明人指南》同样强调，专利保护应“专注于人类的贡献”，美国专利和专利申请的发明人和共同发明人必须是自然人[10]。可见，现有知识产权法规则仍以保护人类智慧为逻辑起点，坚持人类的创作主体地位。但 AI 技术的跃迁正使人类社会逐步接近“强人工智能”的未来，彼时对于 AI 身份的判断或面临更多法理与伦理争议，对此我们将保持关注。

感谢王默、赵怡冰对本文作出的贡献。

[9] 参见印度版权局登记作品查询页，https://copyright.gov.in/SearchRoc.aspx. Last visited on April 7, 2024；参见加拿大版权局数据库《SURYAST》登记信息页，https://www.ic.gc.ca/app/opic-cipo/cpyrghts/dtls.do?fileNum=1188619&type=1&lang=eng. Last visited on April 7, 2024.

[10] Inventorship Guidance for AI-Assisted Inventions, https://www.federalregister.gov/documents/2024/02/13/2024-02623/inventorship-guidance-for-ai-assisted-inventions. Last visited on April 7, 2024.

AI 案例评析

The First Digital Avatar Case in China

宋海燕 陈佩龄

I. Introduction

In April 2023, the Chinese Hangzhou Internet Court ruled on the first case regarding digital avatar in China1. The plaintiff, Mofa Company ("Mofa"), lodged a complaint alleging that the defendant, a network company located in Hangzhou (the "Defendant"), had engaged in copyright infringement and unfair competition. Ultimately, the court of first instance ordered the Defendant to bear the legal responsibility of eliminating the impact and compensating for the loss. Unsatisfied with the result, the Defendant appealed to the Intermediate People's Court of Hangzhou. At present, the case is awaiting commencement of court session by the court of second instance.

In this case, the court of first instance focused on the following issues: (1) whether digital avatar enjoys copyright and/or neighboring right protection; (2) whether the image of digital avatar and its related videos are copyrightable subject matter; (3) whether Mofa, the developer of the digital avatar, enjoys the neighboring right protection; and (4) whether the economic interests of digital avatars can be regulated and protected by the Copyright Law and Anti-Unfair Competition Law.

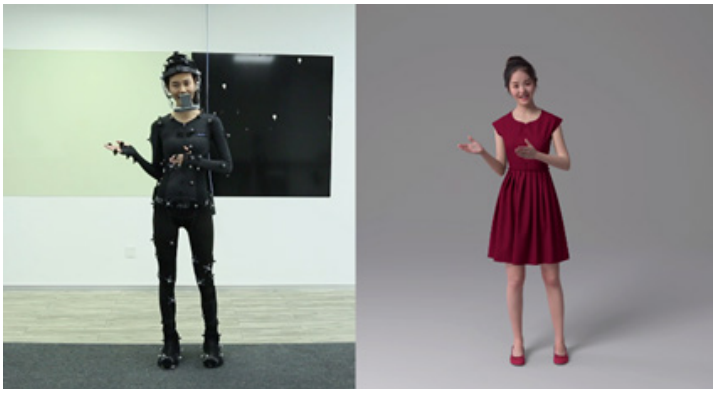
II. Background

Mofa is the developer and creator of the digital avatar "Ada", based on a number of artificial intelligence technologies, including AI performance animation technology, intelligent modeling and binding technology for hyperrealistic characters, intelligent animation and speech synthesis technology, among others.

Screenshot of Ada video by Mofa

[1] See Hangzhou Internet Court (2022) Zhe 0192 Min Chu 9983. According to Hangzhou Intermediate People's Court (2023) Zhe 01 Min Zhong 4722, the second-instance court dismissed Defendant's appeal and affirmed the judgment of the first-instance court on August 8, 2023.

In October 2019, Mofa released Ada through a public event. In the following few months, Mofa released two short videos through Bilibili, a popular Chinese social network platform, to promote Ada, using the motion capture pictures of a real actress named XU.



Screenshots of the video recordings by Mofa

Mofa claimed that the two videos distributed by the Defendant in July 2022 through its Douyin account, infringed the copyright of Mofa's earlier released videos, and thus the Defendant's behavior constituted copyright infringement, unfair competition and false advertising. The Defendant countered that Mofa did not enjoy the copyright for the infringed videos or the subject digital avatar.

The alleged infringing videos by the Defendant

III. Issues

The court identified the issues concerned as follows:

- Whether digital avatar enjoys copyright and/or neighboring right protection under the Copyright Law;
- Whether the image of Ada and its related videos are copyrightable subject matter, and whether Mofa enjoys the neighboring right protection for such images and videos;
- Whether the Defendant is liable for copyright infringement and unfair competition.

(I) Whether digital avatar enjoys copyright and/or neighboring right protection

With regard to the first issue whether digital avatar might enjoy copyright protection or neighboring rights protection, the court concluded that, under the current legal framework, Ada and other similar "weak AI" digital avatars do not enjoy either copyright or neighboring rights protection.

With regard to "copyright" protection, the court reasoned that a digital avatar, similar to Ada, being a specialized application of artificial intelligence technology and a fusion of several technological fields, embodies the interventions and decisions of developers and designers through pre-set algorithms, rules, computational capabilities and learning abilities. The subject digital avatar Ada is mostly driven and determined by human control, functioning as a form of weak artificial intelligence with a relatively limited scope of intellectual creation. Therefore, the court determined that Ada is a tool that supports authorial creativity, but does not qualify as an "author".

With regard to "neighboring rights", the court stated that the "performance" of Ada and similar weak AI digital avatars is, in essence, a digital projection of human performance. The digital avatar itself shall not be recognized as a performer under the current Chinese Copyright Law, and thus does not enjoy performers' rights. The court went further and held that, when a digital avatar participates in filming or acts as a character, it does not enjoy the copyright of audio-visual works or the neighboring right of video producers either.

(II) Whether the image of Ada and its related videos are copyrightable subject matter, and whether Mofa enjoys the neighboring right protection for such images and videos

The Hangzhou Internet Court ruled that the image of Ada and its related videos constitute copyrightable subject matter, and Mofa enjoys all rights related to Ada, including the right as producer of audio-visual works and as performer.

The court held that as a human-driven digital avatar, Ada's form of expression borrows from the physique of a real individual, while also embodying the developer's unique aesthetic choices and judgments regarding line, color and specific image design through the use of virtual beautification. Therefore, the image of the digital avatar Ada qualifies as an "artistic work", and the relevant videos qualify as "audio-visual works" respectively. As such, Mofa enjoys the copyrights for the images

and video productions related to Ada.

Furthermore, the court noted that the motion captures presented by Ada, including its monologues, dancing and other behaviors, are not self-generated. Instead, they closely mirror the actions of the real actress XU. Hence, the court concluded that XU meets the relevant criteria for "performers" as defined by the Chinese Copyright Law. Given the fact that XU is the employee of Mofa, and therefore, Mofa shall own the performing rights related to Ada.

(III) Whether the Defendant is liable for copyright infringement and unfair competition

Regarding the two infringing videos released by the Defendant, the court ruled that both videos were similar to Mofa's earlier released videos, and therefore, finding the Defendant liable for copyright infringement.

Additionally, the court also found that the Defendant exploited the digital avatar Ada's image and related videos on the Douyin platform for marketing and sales purposes. Specifically, the Defendant displayed the digital avatar Ada in video form on its Douyin account and posted product links associated with the digital avatar on its Douyin merchant page, thereby attempting to use Ada's image and related videos to promote courses produced by the Defendant. As such, the court held that the Defendant's behavior would affect consumers' rational decision-making, disrupt the order of market competition, and directly damage Mofa's business interests. As a result, the Defendant was found to constitute unfair competition.

Comments

In this first Chinese case regarding digital avatar, Hangzhou Internet Court addressed several important copyright issues related to digital avatars. What is interesting is that, the court rejected "authorship" status for Ada on the grounds that Ada is only "weak AI" and its performance is primarily determined by humans. When AI is becoming stronger and more intelligent every single moment, we look forward to seeing how the Chinese courts or legislators will respond to Super AI—will they follow the U.S. approach, i.e., rejecting authorship status to AI, not because they are not "strong" but rather they are not "humans"?

Thanks to Wang Mo for her contribution to this article.

生成式人工智能时代下：析美国联邦最高法院 Goldsmith 案中的合理使用新标准

宋海燕 尚文迪

引言

2023 年 5 月 18 日，美国联邦最高法院对安迪·沃霍尔视觉艺术基金会诉戈德史密斯一案（Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith，以下简称为“Goldsmith 案”）作出裁决，判决已故艺术家安迪·沃霍尔（Andy Warhol）根据已故歌手普林斯（Prince）的照片创作的一系列丝网印刷作品侵犯了摄影师林恩·戈德史密斯（Lynn Goldsmith）的版权[1]。

该判决在法律界和艺术界都引起了轩然大波，以至于几个月后，法律学者和艺术界人士仍在继续争论该判决的合理性。同时，在生成式人工智能（Generative AI）正成为社会讨论热点的背景下，有学者认为该判决可能会对与人工智能相关的版权诉讼产生深远影响[2]。本文将梳理该案件的诉讼过程及裁判要点，并简要分析该判决对生成式人工智能相关版权诉讼的影响。

一、背景介绍和诉讼程序

2017 年，安迪·沃霍尔视觉艺术基金会（Andy Warhol Foundation for the Visual Arts，以下简称为“AWF”）向纽约南区地方法院起诉摄影师戈德史密斯，请求法院作出其不侵犯被告的版权或沃霍尔的系列作品构成合理使用（Fair use）的宣告性判决。本案的背景概括如下：

1981 年，《新闻周刊》（Newsweek）聘请戈德史密斯拍摄歌手普林斯，并将该照片刊登在其关于普林斯的杂志报道中。1984 年，戈德史密斯授权《名利场》（Vanity Fair）一次性许可使用她拍摄的一张普林斯的照片，以作为“插图的艺术参考（Artist reference for an illustration）”。随后，受《名利场》的聘请，沃霍尔根据戈德史

---

[1] Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 143 S. Ct. 1258 (2023).

[2] Edward D. Lanquist, Jr, Dominic A. Rota, The Impact of the Supreme Court's Goldsmith Decision on Copyright Enforcement Against AI Tools, JDSUPRA, https://www.jdsupra.com/legalnews/the-impact-of-the-supreme-court-s-7317432/, last visited on August 16, 2023.

密斯的照片创作了一幅紫色的丝网印刷的普林斯肖像作品，该作品被刊登在《名利场》1984 年 11 月刊上。然而，在戈德史密斯不知情的情况下，沃霍尔又根据该照片另外创作了 15 幅普林斯系列作品（这 16 幅作品一起被称为“普林斯系列”），其中就包括本案中涉嫌侵权的作品《橙色普林斯》（Orange Prince）。

2016 年，《名利场》的母公司康泰纳仕（Condé Nast）为纪念当年去世的普林斯制作了一期特刊，在获得 AWF 的授权许可后，刊登了沃霍尔创作的这幅《橙色普林斯》作品。为此，康泰纳仕向 AWF 支付了 10,000 美元，但并未向戈德史密斯支付费用。戈德史密斯认为这一行为侵犯了自己的版权，并通知了 AWF。

随后，AWF 先行起诉了戈德史密斯，而被告戈德史密斯反诉原告 AWF 侵权。原告 AWF 认为，沃霍尔的作品并未侵犯被告的版权，其具有显著的转换性（Transformative-ness）从而构成对被告原始照片的合理使用。被告则辩称，沃霍尔未经许可使用其拍摄的照片构成了对其版权的侵犯。

Figure 1. A black and white portrait photograph of Prince taken in 1981 by Lynn Goldsmith.

图 1 戈德史密斯于 1981 年为普林斯拍摄的黑白肖像照，来源：美国联邦最高法院关于 Gold-smith 案的判决书，143 S. Ct. 1258 (2023)

Figure 2. A purple silkscreen portrait of Prince created in 1984 by Andy Warhol to illustrate an article in Vanity Fair.

图 2 沃霍尔于 1984 年为《名利场》杂志创作的紫色丝网印刷的普林斯肖像作品，来源：美国联邦最高法院关于 Goldsmith 案的判决书，143 S. Ct. 1258 (2023)

Figure 3. An orange silkscreen portrait of Prince on the cover of a special edition magazine published in 2016 by Condé Nast.

图 3 AWF 于 2016 年许可康泰纳仕在特刊上刊登的《橙色普林斯》，来源：美国联邦最高法院关于 Goldsmith 案的判决书，143 S. Ct. 1258 (2023)

本案的诉讼程序概括如下：

纽约南区地方法院认为，根据美国《版权法》第 107 条[3]中判断合理使用的四个要素，沃霍尔的该系列作品构成合理使用，并未侵权。其具体分析如下：（1）使用的目的和性

[3] 17 U. S. C. § 107.

质：该系列作品的部分被捐赠给博物馆，其盈利也被用于支持公众视觉艺术，商业性质被淡化；同时沃霍尔的作品具有转换性，存在区别于原始照片的显著特征，给原始照片赋予了新的表达，构成转换性使用；（2）版权作品的性质：虽然被告的原始照片属于创造性作品，且并未发表，但被告已授权《名利场》发表文章时作为普林斯的形象参考使用。然而当二次创作作品构成转换性使用时，第二项要素的参考重要性会被明显削弱。故综合而言，第二个要素对原被告双方均非有利理由；（3）使用占版权作品的数量和质量：尽管沃霍尔最初使用的是戈德史密斯所拍摄的普林斯照片中普林斯的头部和颈部线条，但他通过裁剪和压平戈德史密斯的照片，从而去掉或尽量减少了照片中对光线、对比度、阴影和其他表现力的使用，故沃霍尔创作时去除了原始照片中几乎所有可受保护的元素；（4）使用对版权作品潜在市场或价值的影响：因为沃霍尔和戈德史密斯的作品面对的是不同的市场，故前者的作品并不构成原始照片的市场替代品。由此，纽约南区地方法院认为合理使用的四个要素中三个要素均有利于原告，并于 2019 年作出有利于原告 AWF 的简易判决（Summary judgment）[4]。被告戈德史密斯对该判决不服，上诉至美国联邦第二巡回上诉法院。

美国联邦第二巡回上诉法院作出了截然相反的判决，其认为合理使用的四个要素均有利于被告，即沃霍尔的该系列作品不构成合理使用，侵犯了被告戈德史密斯的作品版权。其具体分析如下：（1）使用的目的和性质：上诉法院认为对该问题的判断关键在于“二次创作的作品对其原始材料的使用是否服务于根本不同和全新的艺术目的，从而使二次创作有别于创作时使用的原始材料”，而仅仅将一位艺术家的风格强加于原始作品并不足以满足这种区分标准。《橙色普林斯》与原始作品在广义上都是视觉艺术作品，在狭义上是同一个人的肖像画，具有相同的目的和功能；并且沃霍尔的作品对被告的原始照片并不构成转换性使用，因为尽管沃霍尔的作品对戈德史密斯的照片中某些元素进行了删减，但是其仍保留原始照片的基本元素；（2）版权作品的性质：戈德史密斯的原始照片具有创造性和未发表性，沃霍尔的作品是否构成转换性使用与该因素的分析无关，因此该因素有利于被告；（3）使用占版权作品的数量和质量：不论沃霍尔的作品对戈德史密斯的照片进行了何种改变，其仍保留了原始照片的基本元素；并且原告无法合理解释为何沃霍尔的系列作品必须使用戈德史密斯所拍摄的普林斯的肖像照片，而非从其他渠道获得的普林斯的肖像照片；（4）使用对版权作品潜在市场或价值的影响：尽管上诉法院基本认同纽约南区地方法院对原被告作品面对不同市场的判断，但其进一步指出 AWF 的商业许可侵犯了戈德史密斯受保护的市场，即许可被告的照片“用于出版物的编辑目的和其他艺术家创作

[4] Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 382 F. Supp. 3d 312 (S.D.N.Y. 2019).

衍生作品”。由此，上诉法院于 2021 年推翻地方法院的判决并发回重审[5]。原告 AWF 对该判决不服，随后将此案提交至美国联邦最高法院。

二、美国联邦最高法院审理的重点问题

美国联邦最高法院集中于审理判断合理使用的第一个要素，即“使用的目的和性质，包括此类使用是否具有商业性质或用于非营利教育目的”，仅考虑该要素是否有利于 AWF 对康泰纳仕的商业许可行为。

美国联邦最高法院多数意见认为，第一个要素有利于被告戈德史密斯。

美国联邦最高法院强调判断第一个合理使用要素应从客观出发：判决书中指出，“第一个合理使用因素考虑的是对版权作品的使用是否具有进一步的目的或不同的性质，这是一个程度问题，而不同的程度必须与使用的商业性质相平衡”。尽管在先前 Campbel v. Acuff-Rose Music，Inc.[6]（以下简称为“Campbel 案”）这一具有里程碑意义的判决中，美国联邦最高法院裁定 2 LIVE CREW 对 Roy Orbison 的《Oh，Pretty Woman》的模仿构成“转换性使用”而成立合理使用，其认为模仿虽然具有商业性质，但新用途的目的和特征与原始目的显著不同。然而，在本案中，美国联邦最高法院提高了 Campbel 案的标准，否定“任何为原始材料添加新的美学或新的表达方式的二次创作都必然具有转换性”的观点，认为法官不能从艺术家的角度判断作品具有转换性，不能仅仅取决于艺术家所表达或感知的意图，也不能过多考虑从作品中得出的意义，而应从客观视角衡量二创作品利用原创作的目的与性质，与原创作者的排他使用权间的利益平衡。

因此，当新作品的商业性质无可争议，并且新用途的目的与原始作品的目的基本相同时，需要有额外的理由来满足第一个合理使用因素。由此，美国联邦最高法院主要从使用的目的和性质两方面分析第一个要素有利于哪方：

（一）目的：AWF 对康泰纳仕的许可使用行为与戈德史密斯使用照片的行为目的相同

本案中，1984 年和 2016 年的出版物都是刊登普林斯的肖像并说明普林斯的故事，两次使用照片或作品的环境不是“独特和不同的”。因此，这两次使用行为具有基本相同的目的。该行为使得《橙色普林斯》构成对戈德史密斯原始照片的市场替代，这种相同的复制对原作品版权人而言是不公平的。

（二）性质：AWF 对康泰纳仕的许可使用行为具有商业性质

AWF 以 10,000 美元的价格将《橙色普林斯》授权给康泰纳仕，将照片刊登在杂志中

[5] Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 11 F.4th 26 (2d Cir. 2021).
[6] Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 114 S. Ct. 1164 (1994).

出版，该许可毫无疑问是商业性的。

综合来看，目的和性质两个因素——戈德史密斯的照片和 AWF 2016 年对《橙色普林斯》的许可具有实质上相同的目的，以及 AWF 对戈德史密斯照片的使用具有商业性质——都不构成合理使用。由于 AWF 没有进一步补充提供额外的令人信服的理由，法院认为该作品的用途并未通过合理使用的转换性测试。

最终，美国联邦最高法院于 2023 年 5 月 18 日以 7:2 多数通过裁决，裁定沃霍尔对戈德史密斯的照片的使用侵犯了其版权，不受合理使用保护[7]。

三、评论：对与生成式人工智能有关版权诉讼的启发

Goldsmith 案判决生效几个月以来，有关讨论和争议持续不断，赞成和批评兼而有之[8]。其中，引人关注的是，该判决在生成式人工智能不断发展的背景下，可能对与人工智能有关的版权诉讼产生连锁影响，而该影响不利于人工智能服务提供者利用合理使用制度进行抗辩[9]。

生成式人工智能技术通常通过爬取和学习互联网公开提供的大量资源而生成作品，由于其训练数据中包括受知识产权保护的作品及其他内容，其生成内容很可能与原作品基本相似，从而面临着侵害其版权的风险。在 Goldsmith 案判决作出之前，一些人工智能开发者通常依赖合理使用原则进行抗辩[10]；但在该判决之后，合理使用抗辩的证明难度大幅度提升。例如，在近期两起广受关注的 AI 案件——Getty Images 诉 Stability AI 案[11]，Andersen 等多位视觉艺术家诉 Stability AI 的集体诉讼案[12] 中，原告均诉称 Stability AI 未经许可复制和抓取了大量图像，其中包括原告受版权保护的图像，侵害其版权。Goldsmith 案判决为人工智能服务提供者的抗辩带来了困境：其将面临如何证明两张相似的照片被用于不同的目的，即新作品并未与原作品在市场上形成竞争或替代性关系；如果无法证明，其使用受版权保护的图像就可能构成侵权。对于原版权人来说，图库摄影师一般很容易证明新照片与原图库照片具有竞争关系，记者很容易展示人工智能生成的文本可以替换新闻文章；而对比到生成式人工智能技术而言，其举证难度显然更高。

---

[7] 该判决的少数意见提出异议，主要集中在集中在戈德史密斯的照片和沃霍尔的丝网印刷图像之间的艺术批评和艺术历史区别上，认为沃霍尔的作品无疑具有转换性。而限制合理使用的界限，很可能阻碍创作进步，损害创作自由。

[8] Artnet 评论员本·戴维斯（Ben Davis）肯定该判决，认为该判决否认“名人艺术家例外（celebrity-artist exception）”，参见 Ben Davis, Why Andy Warhol's 'Prince' Is Actually Bad, and the Warhol Foundation v. Goldsmith Decision Is Actually Good, Artnet, https://news.artnet.com/opinion/warhol-foundation-v-goldsmith-fair-use-2311801, 2023 年 8 月 7 日访问；
美国公民自由联盟艺术审查项目主任玛乔丽·海因斯（Marjorie Heins）反对该判决，认为该判决是“灾难性的错误（disastrously wrong）”，因为根据其适用的商业和非商业使用的区别，“艺术家、经销商、策展人、收藏家等艺术界人士都必须逐案猜测，一件一开始被视为合理使用的创意作品是否会因展示、销售或营销方式的不同而失去版权法的保护”，参见 Peter J. Karol, After Warhol, Artforum, https://www.artforum.com/slant/the-transformative-impact-of-warhol-v-goldsmith-90667, 2023 年 8 月 7 日访问。

[9] See Jonathan Bailey, What the Warhol Ruling May Mean for AI, Plagiarism Today, https://www.plagiarismtoday.com/2023/05/23/what-the-warhol-ruling-may-mean-for-ai/, last visited on August 7, 2023.

[10] See James Vincent, The scary truth about AI copyright is nobody knows what will happen next, The Verge, https://www.theverge.com/23444685/generative-ai-copyright-infringement-legal-fair-use-training-data, last visited on August 7, 2023.

[11] Getty Images (US) Inc. v. Stability AI Inc., No. 23-cv-135) (D. Del.).

[12] Andersen, et al. v. Stability AI Ltd., et al., Case No. 3:23-cv-00201-WHO (N.D. Cal.)

实践中，已经有作者援引 Goldsmith 案判决以对抗生成式人工智能服务提供者使用受版权保护的作品作为训练数据的行为：7 月 18 日，美国作者协会和 8,000 多名作者签署了一封致 OpenAI、Alphabet、Meta、Stability AI、IBM 和 Microsoft 的公开信，呼吁在训练 AI 时使用受版权保护的材料应征得作者同意、获得授权许可并给予公平补偿[13]。这封信中引用了 Goldsmith 一案，并指出“生成式人工智能的高度商业性与合理使用制度相悖”。

尽管 Goldsmith 案的判决为针对生成式人工智能的侵权诉讼提供了有利于原版权人的导向，但这并不意味着生成式人工智能产品一定构成侵权：一方面，判断合理使用的因素仍很复杂，美国联邦最高法院始终强调“转换性使用”的重要性，平台对其生成内容是否具有足够的“转换性”仍有充分的论述空间；另一方面，在该判决中，美国联邦最高法院强调要构成合理使用，新作品与原作品的目的需有明显不同，如果服务提供者能够明确回答将受版权保护的材料纳入训练数据的目的与原始作品的目的有本质区别，就有可能构建一条合理使用的路径。

结语

Goldsmith 案作为美国联邦最高法院暌违将近 30 年对版权法合理使用的判决，其本身具有显著的指导意义；而在生成式人工智能发展的时代背景下，该判决对生成式人工智能的未来产生了重大影响，并凸显了生成式人工智能的服务提供者和用户面临侵权诉讼的风险。目前，上述与生成式人工智能 Stability AI 有关的诉讼案件均未审结，Goldsmith 案判决究竟会对与人工智能版权相关诉讼产生何种影响，仍需我们持续关注。

感谢王默、赵怡冰对本文作出的贡献。

[13] Brian Fung, Thousands of authors demand payment from AI companies for use of copyrighted works, CNN, https://www.cnn.com/2023/07/19/tech/authors-demand-payment-ai/index.html, last visited on August 7, 2023.

China's First Case on Copyrightability of AI-Generated Picture

宋海燕 张永洁

I. Introduction

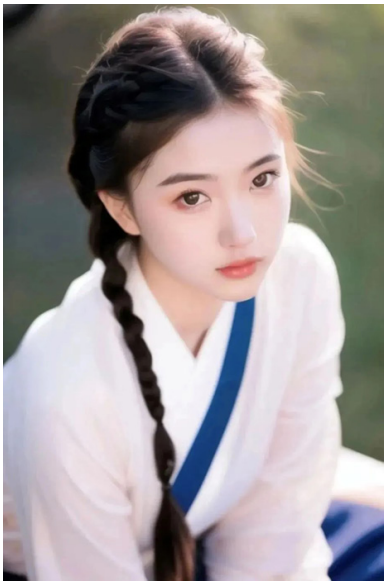
In November 2023, Beijing Internet Court rendered a verdict in China's first case concerning the copyrightability of AI-generated pictures LI v. LIU, which ruling triggered mixed reactions among the AI industry and the public[1]. The plaintiff, Mr. LI (the "**Plaintiff**"), filed a copyright lawsuit alleging that the defendant, Ms. LIU, a blogger on Baijiahao platform (the "**Defendant**"), violated his copyrights in the AI-generated picture, including his right of authorship and the right of dissemination via information networks. In its decision, Beijing Internet Court found that the AI-generated picture was a copyrightable work involving human authorship and that Defendant was liable for copyright infringement.

II. Background

On February 24, 2023, Plaintiff generated a few pictures using Stable Diffusion, a U.S.-based text-to-picture artificial intelligence ("**AI**") service. He labelled one of such pictures as "Spring Breeze Brings Tenderness—AI generated picture" (春风送来了温柔) and posted it on a popular Chinese lifestyle social media platform "Little Red Book" (Xiaohongshu). Defendant, a Chinese blogger, then published an article titled "Love in March, Among Peach Blossoms" (三月的爱情 , 在桃花里), using Plaintiff's AI-generated picture "Spring Breeze Brings Tenderness" as an illustration in her article, but removing Plaintiff's user ID and the watermark of Little Red Book from the picture. Plaintiff soon sued Defendant for copyright infringement, including violating his right of authorship and the right of dissemination via the internet.

---

[1] See Beijing Internet Court (2023) Jing 0491 Min Chu 11279.



Plaintiff's AI-generated picture "Spring Breeze Brings Tenderness"[2]

III. Issues

In this case, the court focused on the following issues: (1) whether the subject AI-generated picture constitutes a copyrightable work and therefore subject to Chinese copyright protection; (2) if yes, whether Plaintiff is the copyright owner of the subject AI-generated picture; and finally (3) whether Defendant should be held liable for copyright infringement.

(I) Whether the Subject AI-Generated Picture Constitutes Copyrightable Work.

On the first issue, Beijing Internet Court ruled that the subject AI-generated picture "Spring Breeze Brings Tenderness" constitutes a copyrightable work—a work of fine art, and therefore, subject to Chinese copyright protection.

The Beijing court started its analysis by listing out the criteria for a work to be protected under Chinese copyright law, including: (1) whether the work falls within the fields of literature, art and science; (2) whether it possesses originality; (3) whether it has a specific form of expression; and finally (4) whether it is intellectual creations (by humans).

With regard to the first and third criteria, the Beijing court held that because the subject picture is akin to commonly seen photographs and paintings, it has satisfied these two criteria.

With regard to the criteria of "intellectual creations", the court held that a (copyrightable) work should reflect the intellectual input/contributions of human beings. In this case, the court found that, Plaintiff has provided intellectual inputs throughout the subject picture-generation process, including: (1) choosing the preferred AI service provider (i.e., Stable Diffusion) among many other alternative AI picture service providers to render the picture style that Plaintiff prefers; (2) inputting

[2] Plaintiff's AI-generated picture "Spring Breeze Brings Tenderness", see Beijing Internet Court Civil Judgment (2023) Jing 0491 Min Chu 11279, p.11.

around 30 "Prompts" and over 120 "negative Prompts" to determine the output of the AI-generated picture; and (3) setting and re-setting various technical parameters to produce, choose and re-arrange the pictures that Plaintiff favors. As such, the court determined that the subject picture has reflected Plaintiff's intellectual input, thus meeting the criteria of "intellectual creations"[3].

With regard to the criteria of "originality", the Beijing court held that a copyrightable work should be independently created by its author and reflect the author's personalized expressions. The court went further, stating that "as to whether the use of AI to generate pictures could reflect the author's personalized expressions, it needs to be decided on a case-by-case basis"[4].The court held that, in this case, although Plaintiff did not physically draw the specific lines (using his own hands), Plaintiff designed the character styles and arranged the final layout and composition of the picture, by trying different prompt words, negative prompt words and various tech parameters. The judges seemed to be especially impressed with the fact that after Plaintiff obtained the first picture by entering around 150 prompts, negative prompt words and relevant parameters, Plaintiff continued to add more prompt words and keep changing the tech parameters until he received the final subject picture that he was happy with. The court thus concluded that, this entire process of adjustment and rearrangement reflected Plaintiff's aesthetic choices and personal judgment. Therefore, the court found that the subject picture is not merely a "mechanical intellectual creation" but possessing originality of the author[5].

The creation process of Plaintiff's subject picture[6]

[3] Beijing Internet Court Civil Judgment (2023) Jing 0491 Min Chu 11279, p.17.
[4] Beijing Internet Court Civil Judgment (2023) Jing 0491 Min Chu 11279, p.18.
[5] Beijing Internet Court Civil Judgment (2023) Jing 0491 Min Chu 11279, p.19.
[6] The process of Plaintiff generating the subject AI picture was described in great details at page 9, Beijing Internet Court Civil Judgment (2023) Jing 0491 Min Chu 11279.

(II) Whether Plaintiff is the Copyright Owner of the Subject Picture

The court first ruled out the possibility that an AI service itself could be considered as an author of a copyrightable work because an AI is not a human being.

Then, the court held that neither the developers/providers of AI services could be considered as the author in this case because such AI providers/developers neither had the intent to create the subject picture nor did they actually participate in the subject picture creation process. Further, based on the "CreativeML Open RATL++-M License" of Stable Diffusion posted on GitHub.com, the AI developers already waived their rights, if any, in the AIGC output—that they "do not claim rights to the output content"[7].As such, the court held that, because the subject picture was generated as a result of Plaintiff's intellectual input and reflected Plaintiff's personalized expressions, Plaintiff is the author of the subject picture.

(III) Whether Defendant Should Be Held Liable

The court finally found Defendant liable for infringing Plaintiff's copyright in the subject AI-generated picture, on the grounds that Defendant removed Plaintiff's user ID and the watermark of Little Red Book from Plaintiff's AI-generated picture and reposted it on social media platform without authorization.

Comments

Beijing Internet Court's decision in LI v. LIU seems to contradict with the recent U.S. decisions on the copyrightability of AIGC output (e.g., "Zarya of the Dawn"[8], "A Recent Entrance to Paradise"[9]and "Theatre D'opera Spatial"[10]), in which both the U.S. Copyright Office and the U.S. courts have denied copyright protection to AIGC outputs that lack human authorship. Yet the difference between the Chinese case LI v. LIU and the U.S. cases is not that Chinese courts believe that non-humans could be "authors" or that Chinese copyright law does not require "human authorship" in copyrightable works. Rather, in LI v. LIU, Beijing Internet Court seems to make a distinction between a straightforward AIGC output, where the human author simply takes and uses the AIGC output "as is" without any creative involvement and an AIGC output, where the human author keeps experimenting and adding various prompts, negative prompts and tech parameters until he receives the final satisfactory piece. In the later scenario, the Beijing court was actually deeming the subject work as "AI-assisted work", where Plaintiff has exercised aesthetic choices and personal judgment in the final representation of the work.

It should also be noted that in the court's hearing, Plaintiff has demonstrated that he could receive the exact same AI picture that he received earlier and claimed copyright for by inputting the exact same prompt instructions he selected (i.e., over 150 prompts, negative prompts and tech param-

---

[7] Beijing Internet Court Civil Judgment (2023) Jing 0491 Min Chu 11279, p.14. See also https://github.com/Stability-AI/stablediffusion/blob/main/LICENSE-MODEL. Last visited on December 6, 2023.
[8] https://www.copyright.gov/docs/zarya-of-the-dawn.pdf. Last visited on December 6, 2023.
[9] https://www.copyright.gov/rulings-filings/review-board/docs/a-recent-entrance-to-paradise.pdf. Last visited on December 6, 2023.
[10] https://www.copyright.gov/rulings-filings/review-board/docs/Theatre-Dopera-Spatial.pdf. Last visited on December 6, 2023.

eters). It would be interesting to speculate whether the Chinese judges would come to the same conclusion—recognizing the copyrightability of the subject AI picture if the AIGC output turns out to be unpredictable—producing various AI pictures each time. Would the Chinese judges change their rationale because the human authors do not have "control" in the AIGC output? Again, where to draw the line between an AIGC output that is not qualified for copyright protection and an AIGC output that is subject to copyright protection remains a challenge for the AI industry to watch in the years to come.

Thanks to Wang Mo and Zhao Yibing for their contributions to this article.

China's First Case on AIGC Output Infringement—Ultraman

宋海燕 王默

I. Introduction

On February 8, 2024, Guangzhou Internet Court rendered a verdict in China's first case concerning the infringement of AIGC outputs, finding the defendant, a text-to-image AIGC provider, liable for infringing the copyrights of the famous Ultraman IP.[1] The court also discussed China's first AI regulation—the Interim Measures for the Management of Generative Artificial Intelligence Services issued in July 2023 (2023 GAI Measures) in its decision, concluding that the defendant failed to exercise reasonable duty of care in generating its AIGC output, thus violating the 2023 GAI Measures.

II. Background

The Japanese company Tsuburaya Productions Co., Ltd. (Tsuburaya) is the copyright owner of the famous cartoon IP —Ultraman series. Tsuburaya granted to the plaintiff Shanghai Character License Administrative Co., Ltd. (SCLA) an exclusive license to the Ultraman series' works in China, including the right of reproduction, the right to prepare derivative works, and also the right to enforce. Tsuburaya has also registered its Ultraman series pictures at the Chinese Copyright Office.

The defendant [company A] provides text-picture AI generating services through its website Tab. In late December of 2023, the plaintiff found that by inputting prompts containing or related to "Ultraman", the defendant's website would generate identical or substantially similar pictures to its Ultraman series images. The plaintiff then sued the defendant for copyright infringement.

III. Issues

In this case, the court focused on the following two issues: (1) Whether the defendant infringed the plaintiff's copyrights, i.e., the right of reproduction, right to prepare derivative works, and right of dissemination via the Internet; and (2) If the defendant has constituted any copyright infringement, what civil liabilities it should bear.

---

[1] See Guangzhou Internet Court (2024) Yue 0192 Min Chu 113. (2024 粤 0192 初 113 号).

(I) Whether the defendant infringed the plaintiff's right of reproduction, right to prepare derivative works, and right of dissemination via the Internet.

On the first issue, Guangzhou Internet Court ruled yes to the first two infringement claims.

1. Whether the defendant infringed the plaintiff's right of reproduction

The court found that the subject Ultraman works are well-known in China and can be accessed from several major online streaming platforms such as iQiyi in China. In the absence of contradictory evidence, the defendant can be presumed to have access to the subject Ultraman works. Also, the court found that the AI pictures generated by Tab are substantially similar to the original expressions of the prior copyrightable works "Ultraman". As a matter of fact, during the AI-picture generating process, when the user input the prompt "Ultraman", the first picture generated by Tab was an identical Ultraman picture with that of Ultraman works. It was only after the user changed its prompts to "Ultraman with long hair", then a new picture of Ultraman with long hair would appear; and if the user input another prompt "Ultraman in cartoon style", he will then receive another picture of Ultraman in cartoon style. Therefore, the court found that the defendant infringed the plaintiff's right of reproduction.

2. Whether the defendant infringed the plaintiff's right to prepare derivative works

Here, the court found that the subject AI-generated pictures partially kept the original expressions of the "Ultraman Tiga Hybrid Imag", but also formed new features of their own, which constituted unauthorized derivative works to the prior copyrightable Ultraman works. Therefore, the court found that the defendant infringed the plaintiff's right to prepare derivative works.

The Comparison between Tsuburaya's Ultraman image and the AIGC Outputs[2]

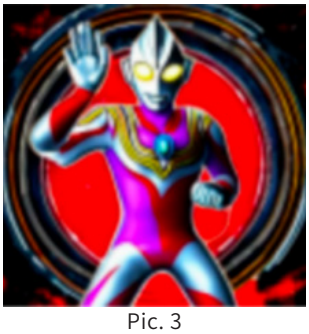
"Ultraman Tiga Hybrid Image" owned by Tsuburaya

[2] The comparison between Tsuburaya's Ultraman image and the AIGC outputs generated by the defendant's website Tab, see Guangzhou Internet Court (2024) Yue 0192 Min Chu 113, p.10-11.

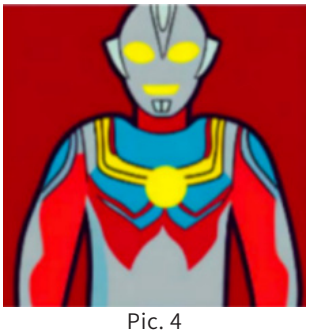
Pic. 1-8: Screenshots of the AIGC Outputs generated by Tab

Pic. 1

Pic. 2

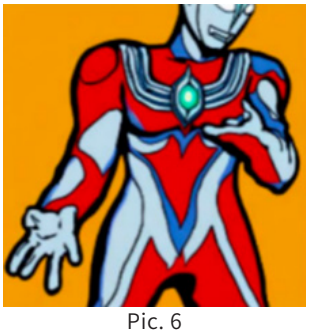


Pic. 3

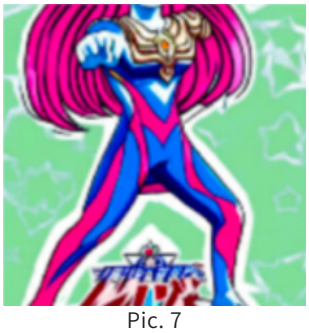


Pic. 4

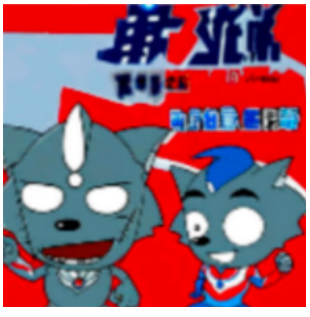
Pic. 5



Pic. 6



Pic. 7



Pic. 8

3. Whether the defendant infringed the plaintiff's right of dissemination via the Internet

The court reasoned that, since it already addressed the defendant's violation of the plaintiff's right of reproduction and right of preparing derivative works and ruled both in the plaintiff's favor, the court would not address this claim involving the same infringing act.

(II) If the defendant has constituted copyright infringement, what civil liabilities it should bear.

In terms of remedies, the court engaged in a detailed discussion on what specific measures that an AIGC provider should implement to avoid liability. It also quoted the 2023 GAI Measures and ruled that the defendant failed to exercise reasonable duty of care, thus violating Articles 4, 12, and 15 of the 2023 GAI Measures.

1. On ceasing the infringing act

The defendant argued that it did not develop the Tab AIGC model itself, but rather relied on a 3rd party vendor to generate such AIGC services by interfacing with the 3rd party application program, and therefore it should not be subject to the 2023 GAI Measures. The court however rejected the defendant's argument, quoting Article 22(2) of the 2023 GAI Measures, and ruled that "AIGC providers", by definition, also include those who provide AIGC services through interfacing with other application programs, thus finding the defendant qualified as an "AIGC provider" as defined under the 2023 GAI Measures.

Then the court acknowledged that the defendant had taken some measures, such as filtering certain keywords, to stop generating infringing pictures to some extent. However, during trial, one could still input other keywords related to Ultraman and generate pictures substantially similar to the prior copyrightable Ultraman images. Therefore, the court concluded that the defendant should take extra steps to the extent that its AIGC function no longer generates any pictures substantially similar to the prior copyrightable Ultraman works when its users input any Ultraman-related prompts.

In terms of the plaintiff's claim on training data, the court rejected its request for the defendant to delete the copyrightable Ultraman works from its training data base because the defendant itself did not train the AIGC model.

2. On whether the defendant should pay compensation for damages

The court found that the defendant's fault needs to be taken into consideration when determining its civil liability for paying compensation for damages. AIGC as tools can be used for both legal and illegal purposes. The court found the defendant as an AIGC provider failed to exercise reasonable duty of care as required by the 2023 GAI Measures in the following aspects:

- **Failure to implement a complaint reporting mechanism.** The court quoted Article 15 of the 2023 GAI Measures, finding that by the date of the trial, the defendant had not established any

complaint reporting mechanism on its website Tab for the copyright holders to enforce their copyrights.

- **Failure to remind users to respect IP in the ToU.** The court quoted Article 4(3) and 4(5) of the 2023 GAI Measures, finding that the defendant, as an AIGC provider, also failed to include relevant provisions in its Term of Use Agreement or in other ways to remind its users that they should not infringe the IP rights of others. The court emphasized that, AIGC users might not have a clear understanding of the infringement risks associated with AIGC outputs, and therefore AIGC providers have a duty to remind/educate users to refrain from using the AIGC services to infringe the IP rights of others.
- **Failure to mark AIGC outputs.** Pursuant to Article 12 of the 2023 GAI Measures and Article 17 of the Provisions on the Administration of Deep Synthesis of Internet-based Information Services (Deep Synthesis Provisions), AIGC providers have a duty to "mark" the AIGC outputs to distinguish them from human-authored works. The court noted that, AIGC providers' duty of marking AIGC outputs not only protects the public's right of information, but also protects the IP holders.

In summary, the court ruled that the defendant failed to exercise its duty of care and thus was subjectively at fault and should bear the corresponding liability to pay compensation of RMB10,000 for its infringement act.

Comments

The Ultraman case is the first Chinese case addressing the infringement issue associated with AIGC outputs, and it is also the first court decision interpreting China's first AI regulation—the 2023 GAI Measures.

In particular, the Guangzhou court had a detailed discussion with respect to AIGC providers' reasonable duty of care when generating AIGC outputs, including its obligations (1) to set up a complaint reporting mechanism; (2) to remind/educate users to respect IP rights of others; and (3) to mark the AIGC outputs to distinguish from human-authored works. The failure of AIGC providers to satisfy the aforesaid requirements could indicate its violation of the 2023 GAI Measures and also its copyright infringement liability.

It should also be noted that, the court did emphasize the need "not to overburden AIGC providers", where AIGC providers should take "proactive measures to fulfill reasonable and affordable duty of care", so as to leave space for the development of generative AI industry, which is still at its early stage. At the end of the decision, the court called for a creation of a Chinese-style AI governance system that is balanced, inclusive, and compatible with both innovation and protection.

Thanks to Zhao Yibing for her contribution to this article.

AI 合规管理

千帆竞发，百舸争流——AI 大模型在汽车行业应用合规风险管理

赵新华 王哲峰 单文钰 米华林

引言

随着 ChatGPT 走红，AI 大模型的热度与日俱增。不到一年的时间，国内已经进入了“百模大战”，10 亿参数规模以上的大模型就已经近百，20% 左右是通用大模型，80% 左右是垂直领域大模型[1]。各行各业都在思考和讨论如何借由 AI 大模型加速行业发展，为企业赋能。以北京为例，2023 年北京首批发布的行业大模型典型应用案例覆盖了医疗、电力、消费、金融、建筑、交通、汽车等多个行业[2]。

汽车行业作为科技创新的主力军之一，同时叠加汽车电动化与智能化两大趋势，AI 大模型“上车”成为众多车企的必选项。AI 大模型正在悄然改变汽车行业。本文拟通过梳理 AI 大模型在汽车行业的应用场景，分析 AI 大模型“上车”的重点法律问题并提供相关风险防范建议。

一、AI 大模型的概念和常见类别

（一）基本概念

AI 算法是一组通过计算机代码形式体现的计算规则，用于处理和分析数据。基于已有数据集运行 AI 算法后所得到的模型数据集与 AI 算法共同构成了 AI 模型，AI 模型可作为未来推理预测的基础和处理相关数据的参照。

AI 大模型通常是基于海量规模数据集运行相关算法完成预训练后的产物。以 GPT 大模型为例，从 GPT-1 到 GPT-3，模型的参数量从 1.17 亿个增长到了 1750 亿个，最新发布的 GPT-4 据悉已达到了万亿级别的参数量[3]。

（二）常见类别

1. 自然语言处理类大模型

自然语言处理（Natural Language Processing，缩写为 NLP）是一种使计算机能

[1] 36 氪，《大模型狂奔 300 天》：https://36kr.com/p/2469370022778758。
[2] 澎湃新闻，《北京发布首批 10 个行业大模型典型应用案例》：https://m.thepaper.cn/newsDetail_forward_23646784。
[3] Matthias Bastian, GPT-4 has more than a trillion parameters - Report：https://the-decoder.com/gpt-4-has-a-trillion-parameters/。

够认知、理解、生成人类语言的技术。常见的自然语言处理类大模型包括大语言模型（Large Language Model，缩写为 LLM），即包含数十亿以上参数的旨在理解和生成人类语言的大型语言模型。

ChatGPT 是基于大语言模型的典型产品，一方面 ChatGPT 能够认知、理解人类提出的问题和要求，另一方面 ChatGPT 能在一定程度上帮助人类完成回答问题、规划行程、购买产品、预约服务、创作、摘要、编程等工作。

2. 计算机视觉类大模型

计算机视觉（Computer Vision，缩写为 CV）是计算机和系统从图像、视频和其他视觉输入中获取有意义的信息，并根据该等信息采取行动或提供建议的技术。

MetaAI 于 2023 年 4 月在官网发布了基础模型 Segment Anything Model（SAM），该模型在包含超过 10 亿个掩码的多样化、高质量数据集（SA-1B）上进行训练，可以对图像中的各类对象进行分割，属于一种计算机视觉类大模型 [4]。由于图像分割是许多任务中的基础步骤，SAM 后续可能在自动驾驶、车牌识别、人脸识别等场景得到进一步应用。

3. 多模态大模型

多模态大模型指能够处理、分析文本、音频、图像、视频等多种输入类型并相应输出的大模型。相较于经过单一的文字类数据训练的自然语言类模型，多模态模型在训练阶段融合了多维度的数据，因此具备更高的通用性，可应用的场景也更多。

2021 年，OpenAI 发布了 Contrastive Language-Image Pre-Training（CLIP）模型，该模型是一种从文字覆盖到图像的多模态模型，可以在无监督预训练之后将文本和图像对应，从而基于文本对图片进行分类，而非只能依赖于图片标签 [5]。CLIP 可以广泛应用于图像检索、图像生成、视觉导航等场景中。此外，OpenAI 还推出了一个名为 DALL-E 的模型，它是在 CLIP 模型的基础上开发的一种生成模型，能够基于自然语言描述和 / 或图像生成图像，该模型也属于一种多模态大模型 [6]。

二、AI 大模型在汽车行业的应用场景

AI 大模型目前在汽车行业的应用场景主要包括智能座舱与自动驾驶：

（一）智能座舱

越来越多的车企开始考虑在车内引入 AI 大模型。针对当前车载语音系统智能化、个性化、情感化、交互性不足等问题，自然语言处理类大模型可以赋予车载语音系统以智能

[4] Meta AI 官网：https://ai.meta.com/research/publications/segment-anything/。
[5] OpenAI 官网：https://openai.com/research/clip。
[6] OpenAI 官网：https://openai.com/dall-e-3。

和情感，从而使车载语音系统能够处理完整对话并可以保持对前后文的理解，能够记录用户的喜好和习惯；同时，多模态大模型可以助力融合语音、视觉、手势、文字等多种交互方式，满足用户在不同场景下的不同使用习惯，从而赋予用户良好的人车交互体验。

（二）自动驾驶

自动驾驶系统一般分为感知、决策和执行三个环节。目前 AI 大模型主要运用于自动驾驶的感知环节，但未来 AI 大模型的作用有可能延伸到自动驾驶系统的决策和执行环节。此外，AI 大模型还可能会被应用于自动驾驶相关模型的预训练。

1. 感知环节

感知环节数据的收集一般通过摄像头、雷达等传感器完成。为了能够做出可靠的驾驶决策，对感知层收集到的多种数据进行充分融合并准确理解至关重要。

对此，多模态大模型有助于对传感器收集到的原始数据或从原始数据中提取的相关特征，在统一相关时间和空间后，映射到统一坐标系下进行前融合或特征融合，从而提升感知环节的精度。业内存在一种观点认为，相较于计算机视觉类大模型仅基于摄像头的纯视觉方案而言，基于摄像头和雷达的成熟的多模态大模型感知方案安全性可能更高，因为雷达为感知环节提供了安全冗余。

此外，在相关大模型的赋能下，自动驾驶车辆可在感知环节实时收集车辆周边的环境信息、理解交通规则并预测其他车辆的行为，从而可以减少甚至去除对于高精地图的依赖。大模型赋能为“脱图”提供了技术和理论上的可能性，“重感知，轻地图”可能成为行业发展的主流方向。目前不少自动驾驶厂商相继提出“脱图”时间表。

2. 端到端感知决策一体化

多模态大模型等 AI 大模型可能促成输入数据到输出控制仅通过一个 AI 大模型实现，实现“端到端”控制和感知决策一体化。端到端感知决策一体化方案的优势在于其使得自动驾驶成为一个整体，避免级联误差，更贴近人类的驾驶过程；其劣势在于该模式属于黑箱模型，出现问题时较难快速找到问题所在，需要依靠推测和实验进行调整，存在一定安全性隐忧，故目前尚未成为自动驾驶的主流方案[7]。尽管如此，目前仍有不少自动驾驶厂商在积极探索端到端感知决策一体化方案安全落地的可能性。

3. 模型预训练

在模型训练阶段，挖掘到有价值的数据后，需要对采集的数据进行标注，将未经处理的初级数据进行加工处理，转换为机器可识别的数据集，从而用于实现对模型的训练和迭代。

---

[7] 东方证券，《AI 大模型加速落地，汽车智能化迅速发展》第 1.2 部分，2023 年 6 月 27 日。

自动驾驶语境下的数据多为摄像头收集的图像、视频数据，对该等数据的标注涉及2D 至 4D 标注、车道线标注、语义分割等，复杂性高、工作量大，若全部由人工标注，可能存在效率低、成本高和准确性差等问题[8]。借助计算机视觉类大模型可实现以 AI 自动数据标注为主，人工复核为辅的进阶模式，有助于加速数据标注工作流程并大幅降低数据标注成本。

除使用真实场景数据外，自动驾驶相关模型的训练还可能使用仿真场景数据。仿真场景数据成本低，无需标注，且可构建边缘场景（如极端天气、长尾场景等），从而弥补大模型训练数据成本高、数量不足等问题。计算机视觉类大模型、多模态大模型等大模型可助力构建高仿真场景，缩小仿真数据与真实数据之间的差异，提高场景泛化能力，并提高仿真场景的针对性[9]。

三、AI 大模型“上车”时车企需要关注的主要法律问题

在 AI 大模型“上车”逐渐成为众多车企发展目标的同时，相应监管要求也正日益出台。相关企业（包括车企，技术研发企业、方案与零部件供应商等）需关注如何落实 AI 大模型“上车”相关合规要求，应对 AI 大模型这一新技术新应用带来的合规风险和挑战。

（一）分类分级监管

如果说数据是 AI 大模型的血肉，那么算法就是 AI 大模型的骨架。针对 AI 大模型的该两大核心要素，我国均提出了分类分级监管要求。

1. 数据分类分级

2021 年以来，《数据安全法》与《个人信息保护法》等数据保护法规相继出台，中国确立了数据分类分级监管的基本思路。

与 AI 大模型“上车”带来的应用革新伴生的是相关企业收集数据的种类愈发多元化，收集数据的数量呈指数级增长，对所收集数据的分析利用更加复杂。相应的，相关企业面临的数据安全风险更高，带来的数据分类与分级工作的压力也更大。

（1）智能座舱场景涉及的主要数据种类

如今，智能座舱已从触摸式中控大屏时代跃入人车交互时代。在新一代智能座舱中，通过基于多模态大模型的新一代人机交互技术与互动设计，车内人员可以仅通过语音或动作方式发送指令，甚至，可以仅通过“给一个眼神”来向智能汽车传达需求。

根据不同的人机交互技术，新一代智能座舱可能收集的数据主要可分为以下四类：

---

[8] 开源证券，《智能汽车系列深度（十）：自动驾驶算法篇——大模型助力，自动驾驶有望迎来奇点》第 2.5.1 部分，2023 年 6 月 29 日。
[9] 安信证券，《AI 大模型在自动驾驶中的应用》第 3.1 部分，2023 年 5 月 4 日。

- **视觉反馈数据**

  在搭载车内人员监控系统（DMS/OMS）的智能座舱内，车机系统可以通过摄像头追踪车内人员瞥向某一屏幕的视线，随即自动亮屏或唤醒车载助手；或者通过眼动追踪与面部状态监控，识别驾驶员是否分心或者疲劳驾驶，从而适时发出提醒。

  在智能座舱捕捉并分析车内人员的视线过程中，涉及收集车内人员眼睛位置、眼球反光点、眼球移动指标、视线停留时长等多种数据。

- **语音感知数据**

  通过 AI 语音与语义识别技术，智能座舱整体已经可以较好实现在自然环境中识别车内人员的指令，并作出正确响应。为了实现良好的语音交互效果，智能座舱需要收集车内语音，并对其进行语义分析，再根据语义分析结果对语音指令作出反馈。语音指令中可能包括家庭地址、公司地址、联系人姓名和电话、通讯录、住宿信息、行程等个人信息。智能座舱还可能收集车内人员的声纹信息，从而根据不同声纹判断声音的来源。

- **行为感知数据**

  新一代智能座舱可以通过摄像头收集车内人员的手势、面部朝向、身体位置、头部转动方向等数据，并根据该等数据分析车内人员的意图作出反馈。例如，对着车载屏幕挥手实现不同应用之间的切换。

- **其他常规数据**

  新一代智能座舱一般仍保留传统中控大屏，车内人员可通过中控大屏实现用户注册，查询信息，更改车内参数等功能，这个过程中可能涉及收集车内人员的姓名、电话、住址、住宿信息、行程等个人信息。

由此可见，智能座舱收集的车内数据中包含不少个人信息，其中还涉及声纹、面部识别特征、通讯录、住宿信息、行踪轨迹等敏感个人信息。若收集的个人信息数量较大，大规模的个人信息作为一个整体可能构成重要数据。

（2）自动驾驶场景涉及的主要数据种类

自动驾驶技术的关键在于赋予车辆以“眼睛”和“大脑”，促使车辆可以像驾驶员一样收集、识别和理解车辆周边的道路、信号灯、车道线、指示牌、其他车辆、行人、其他障碍物等信息，并进行相应的决策和执行环节。

培育 AI 大模型这一自动驾驶“大脑”的过程中，需要集采和测试车辆收集大量数据，

实际自动驾驶的过程中，亦需要通过摄像头、雷达等车辆的“眼睛”收集大量数据，可能收集的数据主要包括以下几类：

- **面向车外的摄像头数据**

  摄像头将收集车辆周边的道路、信号灯、车道线、指示牌、其他车辆、行人、其他障碍物等图像、视频格式的数据，其中可能涉及人脸信息和车牌信息。

- **雷达数据**

  激光雷达、毫米波雷达、超声波雷达等雷达可以收集车辆周围障碍物的尺寸、位置、形状及与障碍物的距离等数据。

- **从车辆本身获取的数据**

  包括车速、转向方位、转向角度、GPS 位置、剩余油量等从车辆本身获取的数据。

- **时间与位置数据**

  此类数据作为附加信息，反映了以上各类数据记录的时间与地理位置。

不同于智能座舱收集的多为个人信息，用于训练自动驾驶相关算法的数据主要为车辆周围的道路状况、建筑物、车辆密度、行人数量等非个人信息，仅包括少量个人信息如人脸信息、车牌号等。

自动驾驶过程中收集的部分非个人信息可能构成重要数据，如包含人脸信息或车牌信息等的车外视频或图像数据，以及某区域的车辆流量或人员流量等经济运行数据（自动驾驶技术往往需要海量数据用于训练和迭代 AI 大模型，通过对海量技术的综合分析，可能掌握某区域的车辆流量或人员流量等经济运行数据）[10]。

为了更好地应对数据分类分级的合规要求，相关企业首先可以盘点车内外收集的数据资产并标记不同数据的“类别”与“级别”，在标记数据“类别”与“级别”时，可以参考对数据分类分级的相关规定和指引，如《汽车数据安全管理若干规定（试行）》以及数据分类相关国家标准和行业指南；然后，企业可以根据标记实施不同管理措施，例如对不同类别、级别的数据采取访问权限控制等措施，对于特别敏感、重要的数据，例如 GPS 信息、通讯录信息、人脸信息、车牌信息等，可以不上传到云端仅在车内处理或者脱敏后再上传到云端处理。

2. 算法分类分级

除了对数据分类分级监管，中国也正在推进算法分类分级监管机制。

[10] 根据《汽车数据安全管理若干规定》第 3 条，车辆流量、人员流量等反映经济运行的数据，以及包含人脸信息或车牌信息等的车外视频或图像数据属于重要数据。

自 2023 年 8 月 15 日起施行的《生成式人工智能服务管理暂行办法》中明确提出，将“针对生成式人工智能技术特点及其在有关行业和领域的服务应用，完善与创新发展相适应的科学监管方式，制定相应的分类分级监管规则或者指引”[11]。这呼应了 2021 年国家网信办等九部委发布的《关于加强互联网信息服务算法综合治理的指导意见》中提出的要“坚持风险防控，推进算法分级分类安全管理，有效识别高风险类算法，实施精准治理”的要求[12]。

我国虽然尚未出台算法如何分类分级及实施监管的具体制度，但从我国目前对算法的监管思路（请见下文分析）可以看出，我国对风险相对较高的且具有舆论属性或者社会动员能力的算法采取了算法备案和安全评估等监管手段，该手段是一种对高风险类算法实施精准治理的专门措施。

参考以上监管思路，我们理解智能汽车场景下具备一定风险的算法包括可能具有影响意识形态、威胁人身安全、违背道德伦理等不利后果的算法。就智能座舱而言，AIGC 类相关算法根据车内人员提示输出的回复可能具有影响舆论、意识形态的风险，这要求相关企业就智能座舱中搭载的 AI 大模型输出的内容进行合规风险控制。就自动驾驶而言，相关算法可能影响自动驾驶感知环节信息识别和理解的准确性，一旦感知环节识别和理解的信息不准确，可能会对车内外人员的人身安全带来严重威胁，因此，开发者们需要在开发相关 AI 大模型的过程中采取适当措施保证 AI 大模型的可靠度和安全性。此外，在紧急场景下，如在车辆面临“有轨电车难题”时，AI 大模型的决策还可能会带来道德伦理方面的风险，开发者在设计自动驾驶决策相关算法时需考虑如何处理该等特殊场景。

欧洲也提出了一种算法分级监管的规则。2024 年 3 月 13 日，欧洲议会以 523 票赞成、46 票反对和 49 票弃权的表决结果通过了《人工智能法案》(Artificial Intelligent Act)（“AI 法案”）。按计划，AI 法案将于在官方公报上公布 20 天后生效，并在生效 24 个月后完全适用。AI 法案提出了一种基于风险的分级管理方法[13]。AI 法案参考功能、用途和影响将人工智能的应用分为四个风险级别，即不可接受的风险、高风险、有限风险和低风险。不同风险级别对应不同的监管要求，例如，对于高风险类人工智能，要求进行事前评估后才可投入市场[14]。AI 法案还列举了特定领域的高风险类人工智能系统，包括道路交通、供水、供气、供热和供电关键基础设施的管理和运营等八类[15]。后续中国在制定算法分类分级监管规则或者指引时，欧洲立法经验或可成为有益的参考。

---

[11] 《生成式人工智能服务管理暂行办法》第十六条。
[12] 《关于加强互联网信息服务算法综合治理的指导意见》第（二）条。
[13] Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts：https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206.
[14] AI 法案第 6 条。
[15] AI 法案附件 III。

（二）算法备案与安全评估

我国目前对大模型的主要监管手段包括算法备案和安全评估（实践中又称大模型上线备案）。对于具有舆论属性或者社会动员能力的推荐性算法，我国要求此类算法的互联网服务提供者、算法推荐服务提供者、深度合成服务提供者及生成式人工智能服务提供者履行算法备案手续[16]，并开展安全评估[17]。此外，对于深度合成类算法，若涉及生成或者编辑人脸、人声等生物识别信息的，或者涉及生成或者编辑可能涉及国家安全、国家形象、国家利益和社会公共利益的特殊物体、场景等非生物识别信息的，深度合成服务提供者和技术支持者也应当开展安全评估[18]。实践中，就算法备案而言，大模型产品的服务提供方和技术支持方均应办理；就安全评估而言，大模型产品的技术提供方一般应办理，而服务提供方是否需要办理安全评估需结合具体情况具体分析。

推荐性算法包括生成合成类、个性化推送类、排序精选类、检索过滤类、调度决策类等算法。针对智能座舱场景，其使用的 AIGC 相关算法属于生成合成类算法，通常属于具有舆论属性或者社会动员能力的算法，因此需要完成算法备案，并且视情况可能还需要完成安全评估。而自动驾驶场景下，其使用的算法不必然“具有舆论属性或社会动员能力”，因此相关企业可能不一定需要就该等算法进行算法备案和安全评估。

对于企业如何开展算法安全评估，我国尚未出台具体细则和指引。2023 年 8 月 6 日发布，并于 2024 年 3 月 1 日实施的《信息安全技术 机器学习算法安全评估规范》（GB/T 42888-2023）对算法推荐服务提出了一套详细的评估方法，具有一定参考价值。该标准提出，服务提供者可以从算法主体责任、信息服务、权益保护等方面开展评估工作，并详细列出了各项评估事项下具体应开展的工作以及评估要点，例如应当查阅哪些文件、审核哪些合规义务是否完成等[19]。此外，该标准也建议了对于生成合成类、个性化推送类、排序精选类、检索过滤类、和调度决策类五类算法的评估方法[20]。

目前，其他国家也正在建立算法影响评估制度。例如，2022 年 2 月，美国提出《2022 算法问责法案》（The Algorithm Accountability Act of 2022）（“美国算法问责法案”）[21]，也可在一定程度提供参考价值。美国算法问责法案提出了算法影响评估制度，其规定的评估内容包括：应评估算法应用过程中的隐私风险和保护措施、用户权利保障水平、对用户

---

[16]《互联网信息服务算法推荐管理规定》第 24 条；《互联网信息服务深度合成管理规定》第 19 条；《生成式人工智能服务管理暂行办法》第 17 条。

[17]《互联网信息服务算法推荐管理规定》第 27 条；《互联网信息服务深度合成管理规定》第 20 条；《生成式人工智能服务管理暂行办法》第 17 条；《具有舆论属性或社会动员能力的互联网信息服务安全评估规定》第 3 条；

[18]《互联网信息服务深度合成管理规定》第 15 条。

[19]《信息安全技术 机器学习算法安全评估规范》（GB/T 42888-2023）附录 A。

[20]《信息安全技术 机器学习算法安全评估规范》（GB/T 42888-2023）附录 B。

[21] Algorithmic Accountability Act of 2022, H.R. 6580, 117th Cong. (2022)：https://www.congress.gov/bill/117th-congress/house-bill/6580/text?r=2&s=1.

可能产生的风险及应对策略等 [22]。

（三）科技伦理审查

随着《科技伦理审查办法（试行）》公布，我国的科技伦理监管规则变得更加清晰。《科技伦理审查办法（试行）》主要规定了科技伦理审查的适用范围、责任主体、主要程序，以及各类主体的监督管理职责。

根据《科技伦理审查办法（试行）》，企业开展涉及 AI 的科学研究活动（包括智能座舱、自动驾驶相关 AI 大模型的研发与应用），较可能属于开展对“生命健康、生态环境、公共秩序、可持续发展等方面带来伦理风险挑战的科技活动”，从而需要开展科技伦理审查 [23]。

如果企业的人工智能科技研究内容涉及科技伦理敏感领域，该企业应设立科技伦理（审查）委员会 [24]。对于应用 AI 大模型开展自动驾驶研发活动的企业，由于自动驾驶技术本身的高度伦理敏感性，相关技术的研发企业有可能属于应当设立科技伦理（审查）委员会的义务主体。

此外，《科技伦理审查办法（试行）》还提出要建立需要开展专家复核的科技活动清单制度，对可能产生较大伦理风险挑战的新兴科技活动实施清单管理 [25]。清单上的企业通过本单位科技伦理（审查）委员会的初步审查后，需由本单位报请所在地方或相关行业主管部门开展专家复核。根据目前已公布的清单，智能座舱相关 AI 大模型研发可能属于“具有舆论社会动员能力和社会意识引导能力的算法模型、应用程序及系统的研发”，自动驾驶相关 AI 大模型研发可能属于“面向存在安全、人身健康风险等场景的具有高度自主能力的自动化决策系统的研发”，因此可能需要开展专家复核 [26]。

目前，对于不同类别的科技活动应当遵循何种审查标准，仍是有待讨论和解决的问题。一些国家和地区正尝试明确何为“符合科技伦理”的科技活动。

以自动驾驶为例，德国提出了“自动驾驶算法伦理准则二十条”（“伦理准则二十条”），旨在规定自动驾驶算法研发中应遵循的原则 [27]。自动驾驶算法开发活动中的核心伦理问题之一是，是否允许量化比较生命的价值。比如多人的生命是否比一个人的生命更富有价值，或者生命的价值是否能以年龄、性别、外表等因素衡量。伦理准则二十条对这种伦理困境提出的解决方案可概括为四点：首先，人的生命优先，动物与无生命的财产可以被牺牲；

---

[22] 美国算法问责法案第 4 部分。
[23] 《科技伦理审查办法（试行）》第 2 条。
[24] 《科技伦理审查办法（试行）》第 4 条。
[25] 《科技伦理审查办法（试行）》第 25 条。
[26] 《科技伦理审查办法（试行）》附件（需要开展伦理审查复核的科技活动清单）。
[27] https://www.bmvi.de/SharedDocs/EN/Documents/G/ethic-commission-report.pdf?__blob=publicationFile.

第二，生命与生命之间不可衡量；第三，不得基于人群特征预先设定方案；第四，面对真正的伦理困境，应当交由人类决策。尽管伦理准则二十条的上述倡议未直接明确何为“符合科技伦理”的科技活动，但该等原则性的解决方案对于判断何为“符合科技伦理”的科技活动提供了一定框架和参考。

就目前而言，企业如果希望开展科技伦理审查，除了依据有限的法规以外，也可考虑聘请外部专家协助企业设立科技伦理（审查）委员会，制定企业内部的科技伦理审查标准。对于可能需要上报开展专家复核的较大伦理风险的科技活动，企业可考虑尽早与相关监管部门沟通，并对行业内的科技伦理审查具体标准保持关注。

（四）数据语料的有序合法流动

数据是 AI 大模型的核心要素之一，数据质量和数据规模很大程度决定了 AI 大模型的整体效率和质量。AI 大模型开发企业目前使用的数据语料的主要来源有三个方面：一是点对点从第三方处购买；二是购买数据交易所内公开上架的产品；三是使用开源数据集等免费公开数据。

1. 点对点数据流动

从第三方点对点购买大量数据是获取数据语料的途径之一。然而，对于希望训练汽车行业 AI 大模型的开发者而言，购买训练所需数据并非易事。

如前文所述，用于迭代自动驾驶相关 AI 大模型的训练数据可能含有重要数据。根据《工业和信息化领域数据安全管理办法（试行）》，提供重要数据和核心数据的，应当与数据获取方签订数据安全协议明确双方的法律责任，对数据获取方数据安全保护能力进行核验，并需要采取必要的安全保护措施[28]，如校验技术、密码技术、安全传输通道或者安全传输协议等措施[29]。企业获取重要数据后，也需履行作为重要数据处理者的一系列合规义务，包括履行不同维度的申报、备案义务、采取更严格的安全保障措施、制定并执行重要数据处理活动全生命周期的管控制度和流程等。

对于应用于智能座舱的 AI 大模型，则更需要大量个人信息用于训练模型。我国个人信息保护制度对个人信息的共享活动提出了一系列要求。例如，《个人信息保护法》要求个人信息处理者将个人信息对外提供的，应当告知个人并取得同意，还需进行个人信息保护影响评估[30]。《信息安全技术—个人信息安全规范》（GB/T 35273-2020）则明确指出，在共享或转移个人信息时，应通过合同等方式规定数据接收方的责任和义务[31]。关于数据

---

[28] 《工业和信息化领域数据安全管理办法（试行）》第 18 条。
[29] 《工业和信息化领域数据安全管理办法（试行）》第 17 条。
[30] 《个人信息保护法》第 23、55 条。
[31] 《信息安全技术 - 个人信息安全规范》（GB/T 35273-2020）第 9.2 条 d) 项。

处理协议应当具备的主要内容与注意事项，请参考我们的文章《个人信息流动中的数据处理协议，你准备好了吗？》[32]。

2. 数据场内交易

随着我国大力推进数据交易所的设立，越来越多企业选择将优质数据产品上架至数据交易所进行交易，其中不乏以数据集、API、数据报告、数据模型为交付形态的相关产品[33]。

对于购买方，其获取数据的过程需同时遵守法律法规规定的数据保护合规义务以及数据交易所的交易规则。对于出售方，其上架的数据产品还需通过数据交易所对数据产品的合规评估。

如果相关企业考虑将自身拥有的数据集、算法、模型等上架数据交易所，可能需要提交关于数据来源、数据授权使用目的和范围、数据处理行为等方面的说明材料以及第三方服务机构出具的数据合规评估报告。该等数据合规评估报告一般包括对交易主体经营风险、交易标的的来源合法性、可交易性和可流通性等事项的评估结论，出售方企业需对此预留时间完成相应准备。

3. 使用开源数据集

在汽车行业，使用开源数据集训练 AI 大模型较为普遍。目前市面上已有不少高质量的开源数据集，帮助开发者“站在巨人的肩膀上”进行研发。

但是，开源并不代表毫无限制。开源数据集往往配套数据集许可协议，旨在规范他人对数据集的利用，保护作者的权益，以及促进数据的开放共享。数据集许可协议一般约定使用者是否应当保留原作者姓名、是否允许商用、是否允许基于商业目的传播、改编或者二次创作，是否要求基于原作的新作品也使用相同的许可协议等内容。企业使用开源数据集进行 AI 大模型科研活动的，除了应遵守法律法规规定的合规义务外，还需留意数据集许可协议中约定的义务。

结语

AI 大模型为汽车行业实现高度智能化带来了新生力量，同时也为企业带来新的合规风险敞口。现有的数据、算法、科技伦理等相关监管要求相对复杂，对企业的法律解读能力与合规实践水平提出了较高要求。在汽车行业拥抱 AI 大模型的时代，相关企业不仅应追求技术领先地位，同时也应提升合规经营水平，方能保持长期健康发展。

感谢实习生郭思雨和卢虹羽对本文作出的贡献。

---

[32] https://mp.weixin.qq.com/s?__biz=MzA4NDMzNjMyNQ==&mid=2653268368&idx=1&sn=0facbb737182b06e8d00c6f1f24aaf83&chksm=8439c23ab34e4b2cb6a44fb70cd002f32a0fbe1b850ba8f47629e32ed894fe61bb32874e6054&scene=21#wechat_redirect.

[33] 参考深圳数据交易所官网，https://www.szdex.com/。

大模型合规之现实初探

张逸瑞 张津豪 张一凡 周彤

引言

2023 年 7 月 31 日，苹果 APP Store 宣布对中国大陆区中大量提供 ChatGPT 类服务的应用进行集中下架。在面向应用开发者给出的回复中，苹果官方表示相关应用未依据中国大陆地区的法律要求取得许可证，故“需下架整改，整改完毕上架”。某种角度而言，该情况可以理解为《生成式人工智能服务管理暂行办法》（“《AIGC 暂行办法》”）施行在即引发的“连锁反应”。2023 年 8 月 15 日施行的《AIGC 暂行办法》是我国亦是全球针对生成式人工智能服务领域制定的首部法规，其中提出了对生成式人工智能服务的分类分级监管要求，明确了提供和使用生成式人工智能服务总体要求，一定程度上，其标志着我国生成式人工智能服务领域进入强监管和高合规标准的新阶段。

实际上，我国对生成式人工智能服务的合规监管的强化早已有迹可循。早在 2023 年年初，国家互联网信息办公室、工业和信息化部、公安部针对深度合成服务制定的《互联网信息服务深度合成管理规定》（“《深度合成管理规定》”）顺利施行，其明确了深度合成服务相关方的义务与主体责任，强化了对互联网信息服务深度合成领域的管理。《AIGC 暂行办法》将与《深度合成管理规定》一并为我国大模型领域构建更为完善的治理和监管框架。

本文将对我国监管体系项下的大模型领域的合规要素予以梳理，并重点关注现实环境下，梳理当前落地应用的大模型主要的合规义务。

一、什么是大模型？

（一）大模型——内含大量参数的深度学习模型

大模型，即 Foundation Models，通常是指具有大量参数和复杂结构的深度学习模型。这些模型的参数量较大，通常需要数十亿甚至上百亿个参数，相较于传统的较小规模模型，大模型具有更高的容量和表达能力。大模型可以通过训练大规模数据集，以实现更

准确的预测和更高的性能，并依据相关指令，完成各种目标任务。我们熟知的 OpenAI 的 ChatGPT 与 Google 的 PaLM 2 就是典型的语言类大模型：ChatGPT 以 Transformer 模型为基础，具有 1750 亿个参数；而 PaLM 2 具有超过 3400 亿个参数。

根据百度、华为等企业近期密集发声的情况来看，目前企业应用大模型主要体现为以下三种模式：一是自主构建基础大模型，但是考虑到训练大模型的成本和技术壁垒都非常高，因此只有少数企业会自建大模型。二是建立行业大模型，通常是了解行业 know-how 的企业，结合自身掌握的行业数据，用基础大模型精调出更贴合实际场景的垂类行业大模型。三是在基础大模型和行业大模型之上，开发 AI 应用，这也是目前大多数企业采取的模式[1]。

（二）以大模型为技术基石的生成式人工智能

生成式人工智能，是以大模型为技术基石、继专业生产内容（Professionally-Generated Content，PGC）、用户生成内容（User-Generated Content，UGC）之后的新型内容创作方式。在大模型的支撑下，早期生成式人工智能在文本生成领域以内容创作为主，后逐渐向音频生成、图像生成等领域推广，逐步在企业端和消费者端领域实现变现，并完成了在消费、产业、学术等诸多场景的落地和应用。目前，微软已将 ChatGPT 嵌入到微软各大系列产品，包括将 GPT-4 接入搜索引擎 New Bing 和 Edge 浏览器、推出集成 New Bing 和其他插件的 AI 助手平台 Copilot 以应用于 Office、协作软件 Teams 以及其他商业应用；同时，OpenAI 也正在着手打造基于语言类大模型的应用商店，打通所有接入 ChatGPT 的应用体系。

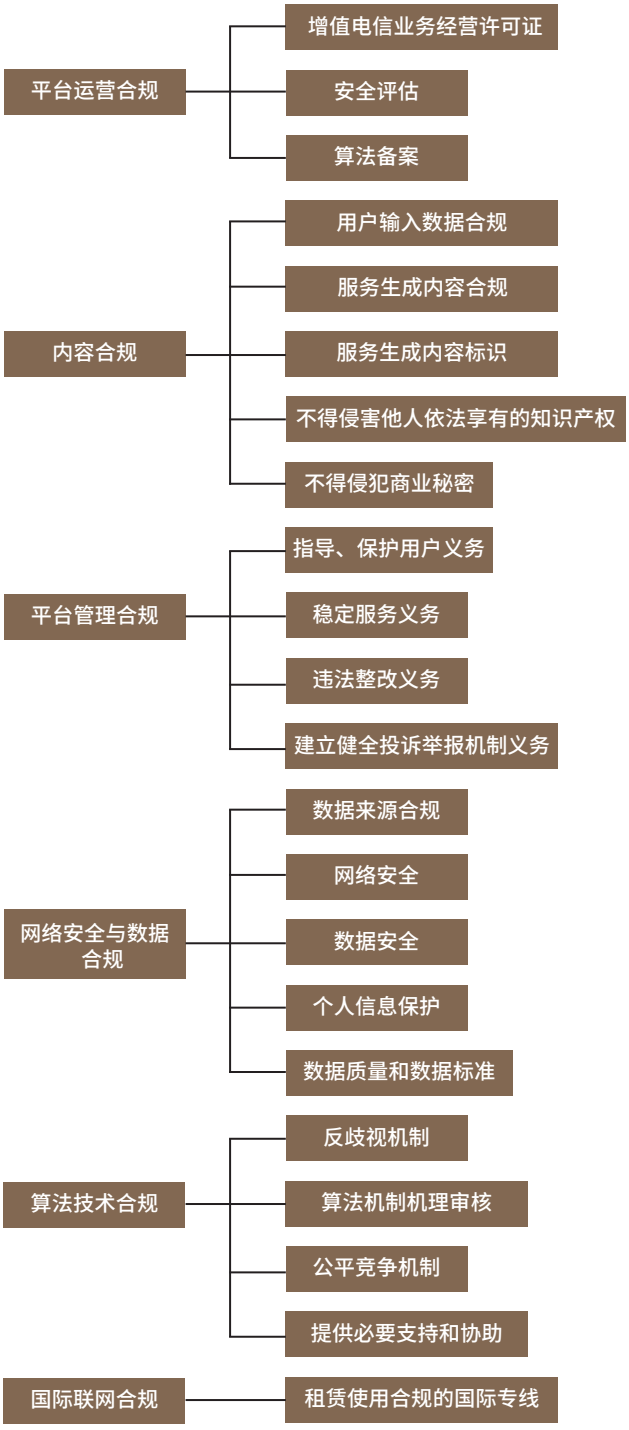
总体而言，大模型在自然语言处理、图像识别、语音识别等领域取得了显著的成果，带来了更精准和高效的机器学习和人工智能应用。但是，大模型的迅速推广应用引发了一系列隐患，如大模型服务被恶意利用开展违法犯罪活动、协助罪犯进行“AI”诈骗；又如部分高校师生利用大模型大量生成文章或者研究内容，在学术造假、学术不端的同时，也可能不知不觉侵犯了潜在权利人的知识产权；此外，还引发了虚假信息传播、数据和隐私信息泄露、偏见歧视等诸多问题。因此，大模型的推广应用，势必伴随着系统规范的大模型合规监管体系。

二、大模型合规要素

在我国当前的监管体系下，大模型合规要素主要涉及的范畴包括平台运营合规、内容合规、平台管理合规、网络安全与数据合规、算法技术合规、国际联网合规等方面，具体合规要素以及相应的法律法规依据详见下图：

---

[1] 参见《百度沈抖：文心大模型拥有中国最大的产业应用规模，已在十余个行业落地》，新浪财经，2023 年 7 月 6 日，链接：https://finance.sina.cn/2023-07-06/detail-imyztprp2163876.d.html，最后访问日期：2024 年 3 月 18 日。


平台运营合规
增值电信业务经营许可证
安全评估
算法备案
内容合规
用户输入数据合规
服务生成内容合规
服务生成内容标识
不得侵害他人依法享有的知识产权
不得侵犯商业秘密
平台管理合规
指导、保护用户义务
稳定服务义务
违法整改义务
建立健全投诉举报机制义务
网络安全与数据合规
数据来源合规
网络安全
数据安全
个人信息保护
数据质量和数据标准
算法技术合规
反歧视机制
算法机制机理审核
公平竞争机制
提供必要支持和协助
国际联网合规
租赁使用合规的国际专线

大模型合规要素一览

三、大模型合规相关概念解析

以下，我们首先对上述合规要素提及的“生成式人工智能技术”“深度合成技术”“算法推荐技术” “具有舆论属性或社会动员能力的互联网信息服务”等大模型合规的重要概念解析如下：

概念	定义
“生成式人工智能技术”	依据《AIGC 暂行办法》，“生成式人工智能技术”是指具有文本、图片、音频、视频等内容生成能力的模型及相关技术。
“深度合成技术”	依据《深度合成管理规定》，“深度合成技术”是指利用深度学习、虚拟现实等生成合成类算法制作文本、图像、音频、视频、虚拟场景等网络信息的技术，包括但不限于篇章生成、文本风格转换、问答对话等生成或者编辑文本内容的技术；人脸生成、人脸替换、人物属性编辑、人脸操控、姿态操控等生成或者编辑图像、视频内容中生物特征的技术；三维重建、数字仿真等生成或者编辑数字人物、虚拟场景的技术等。
“算法推荐技术”	依据《互联网信息服务算法推荐管理规定》（“**《算法推荐管理规定》**”），“算法推荐技术”是指利用生成合成类、个性化推送类、排序精选类、检索过滤类、调度决策类等算法技术向用户提供信息的技术。
“具有舆论属性或社会动员能力的互联网信息服务”	依据《具有舆论属性或社会动员能力的互联网信息服务安全评估规定》（“**《安全评估规定》**”），“具有舆论属性或社会动员能力的互联网信息服务”是指开办论坛、博客、微博客、聊天室、通讯群组、公众账号、短视频、网络直播、信息分享、小程序等信息服务或者附设相应功能以及开办提供公众舆论表达渠道或者具有发动社会公众从事特定活动能力的其他互联网信息服务。

需要说明的是，大模型的核心概念其实是“深度学习 + 自动生成”，而生成式人工智能技术、深度合成技术和算法推荐技术并非相互独立，三者相互配合运作才形成了完整的大模型结构。同时，在满足一定条件的情况下，基于生成式人工智能技术、深度合成技术和算法推荐技术所提供的服务会成为具有舆论属性或社会动员能力的互联网信息服务，而该等服务需要满足特殊的合规要求。

四、大模型合规义务承担主体

（一）大模型服务提供者

大模型服务提供者，即利用大模型技术提供服务的组织、个人。具体来讲，大模型服务提供者又分为以下两类：

- **平台运营方**

  平台运营方是指负责大模型的商业性开发，依据相关规定取得相应资质证照，承担相应义务与责任，提供大模型技术应用服务的组织、个人。在大部分情形下，平台运营方针对的是面向终端消费者的大模型应用场景，比如百度文心一格网站，抖音快手上面的一些 AI 特效功能。

- **技术支持方**

  技术支持方是指负责大模型的技术性开发的组织、个人。技术支持方是大模型的设计者、开发者和完成者，掌握着大模型背后的核心算法和运行规则，负责处理数据训练、生成内容标记、模型优化等技术性事项。在大部分情形下，技术支持方针对的是面向企业的大模型应用场景，通常以 API 形式为企业等提供大模型技术支持。

在《深度合成管理规定》中，合规主体分为“深度合成服务提供者”和“深度合成服务技术支持者”，分别对应上述“平台运营方”和“技术支持方”；而《AIGC 暂行办法》《算法推荐管理规定》等相关法律法规均未对“生成式人工智能服务提供者”“算法推荐服务提供者”进行进一步区分。尽管如此，根据该等规定项下“人工智能服务提供者”“算法推荐服务提供者”责任和义务相关的具体规定，“平台运营方”和“技术支持方”同样需要依据其提供的服务内容及类型承担不同的责任和义务。例如，负责模型训练的技术服务方应当确保训练数据的来源合法合规，而不参与模型训练、不涉及训练数据处理活动的平台运营方应当对技术支持方提供的模型进行必要的合规审查，要求技术支持方对训练数据来源的合法合规性进行陈述保证等，具体详见下文。

（二）什么是“向境内公众提供大模型服务”

根据《AIGC 暂行办法》，行业组织、企业、教育和科研机构、公共文化机构、有关专业机构等研发、应用生成式人工智能技术，未向境内公众提供生成式人工智能服务的，不适用本办法的规定（第 2 条）。也即，需要遵守相关大模型合规义务的主体，是指向境内公众提供了服务的大模型服务提供者。若上述主体未向境内公众提供服务的，则不适用大模型相关合规规定。

基于前述规定，实践中也出现了仅面向企业端提供大模型应用服务的大模型服务提供者是否可适用前述规定、豁免相关合规义务的讨论。我们理解，从该条款的目的来看，加强大模型的合规要求与监管要求旨在规范公共层面的数据流通、传播，避免重要、敏感信

息的泄露，以及防止违法、虚假信息和内容在社会层面广泛传播。因此，如果大模型服务提供者仅面向特定企业提供服务，且该企业仅在企业内部使用大模型服务，不会导致大模型服务成果向公众流通，则有可能并不适用相关合规义务。然而，若大模型服务提供者（“A 主体”）作为技术支持方自研大模型，向中国境内的另一作为平台运营方的大模型服务提供者（“B 主体”）提供大模型技术接口并收取技术服务费，接入了大模型技术接口的 B 主体进而面向中国境内的消费者提供大模型应用服务，我们倾向于认为 A 主体与 B 主体均需要履行相关的合规义务。

五、平台运营方与技术支持方的合规义务

（一）平台运营方的合规要求

1. 资质证照

为了保障大模型服务的合规发展，平台运营方在进入市场提供服务前，必须依照相关法律规定取得相应的资质证照。平台运营方作为互联网信息服务提供者，应当根据《互联网信息服务管理办法》和《中华人民共和国电信条例》，申请办理 B25 类信息服务业务的增值电信业务经营许可证（“ICP 证”）；同时，如平台运营方提供的服务具有舆论属性或者社会动员能力，平台运营方在向公众提供服务前，应当进行安全评估，并按照《算法推荐管理规定》履行算法备案手续。具体而言：

（1）增值电信业务经营许可证

根据《互联网信息服务管理办法》，互联网信息服务可分为经营性和非经营性两类。经营性互联网信息服务，是指通过互联网向上网用户有偿提供信息或者网页制作等服务活动。非经营性互联网信息服务是指通过互联网向上网用户无偿提供具有公开性、共享性信息的服务活动（第 3 条）。国家对经营性互联网信息服务实行许可制度；对非经营性互联网信息服务实行备案制度。未取得许可或者未履行备案手续的，不得从事互联网信息服务（第 4 条）。因此，针对经营性互联网信息服务，应取得经营许可证。许可证类型根据相应业务而决定，例如：从事经营性互联网信息服务，需取得 B25 类增值电信业务经营许可证（即 ICP 证）；从事在线数据处理与交易处理业务，需取得 B21 类增值电信业务许可证（即 EDI 证）。

结合大模型服务的特点，一方面，在平台运营方向用户提供大模型应用服务的情况下，平台运营方通过对训练数据和用户输入对话的采集和处理以及平台的建设，通过互联网向用户提供信息内容，通常情况下涉及为其他单位或个人用户发布文本、图片、音视频、应

用软件等提供平台服务，即信息发布平台和递送服务；值得注意的是，大模型服务提供的内容不是经检索与排序的原始信息，而是基于对用户对话的理解和训练数据的分析、编辑后生成的文本，大模型本身也参与了信息的生产过程，这与单纯的通过信息收集与检索、数据组织与存储、分类索引、整理排序等方式为用户提供网页信息、文本、图片、音视频等信息检索查询服务存在一定差异。另一方面，对于“经营性”和“非经营性”的判断，实践中，不宜简单以服务是否收费来判断有偿或是无偿，而往往需要充分考虑是否存在变相营利的情形，与科研、公益等非经营性活动有明显区分。因此，通常而言，大模型服务往往会涉及经营性互联网信息服务，平台运营方应当取得由国务院信息产业主管部门或者省、自治区、直辖市电信管理机构颁发的 ICP 证。

（2）算法备案

目前我国多部法律法规中均以《算法推荐管理规定》为基础，对于“算法备案”的要求予以明确，具体如下：

根据《算法推荐管理规定》，具有舆论属性或者社会动员能力的算法推荐服务提供者应当在提供服务之日起十个工作日内通过互联网信息服务算法备案系统填报服务提供者的名称、服务形式、应用领域、算法类型、算法自评估报告、拟公示内容等信息，履行备案手续。算法推荐服务提供者的备案信息发生变更的，应当在变更之日起十个工作日内办理变更手续。算法推荐服务提供者终止服务的，应当在终止服务之日起二十个工作日内办理注销备案手续，并作出妥善安排（第 24 条）。

根据《深度合成管理规定》，具有舆论属性或者社会动员能力的深度合成服务提供者，应当按照《互联网信息服务算法推荐管理规定》履行备案和变更、注销备案手续，同时，在完成备案后应当在其对外提供服务的网站、应用程序等的显著位置标明其备案编号并提供公示信息链接（第 19 条）。

根据《AIGC 暂行办法》，提供具有舆论属性或者社会动员能力的生成式人工智能服务的，应当按照《互联网信息服务算法推荐管理规定》履行算法备案和变更、注销备案手续（第 17 条）。

2023 年 6 月，国家互联网信息办公室发布境内深度合成服务算法备案清单，其中包括美团在线智能客服算法、快手短视频生成合成算法、百度文生图内容生成算法、百度 PLATO 大模型算法、天猫小蜜智能客服算法、菜鸟物流智能客服算法、讯飞星火认知大模型算法、腾讯云语音合成算法等。算法备案已经成为相关企业保证其合规、稳定发展不可或缺的重要手续。有实务人士指出，ChatGPT 本身未进行算法备案，这可能是相关应

用被集中下架的主要原因[2]。因此，我们理解平台运营方应履行算法备案手续以实现平台经营合规，避免后续在通过应用商店上架过程中遇到障碍。

（3）安全评估

目前我国多部法律法规中均涉及“安全评估”的要求，具体如下：

根据《安全评估规定》，互联网信息服务提供者开展安全评估，应当对信息服务和新技术新应用的合法性，落实法律、行政法规、部门规章和标准规定的安全措施的有效性，防控安全风险的有效性等情况进行全面评估（第 5 条），并且应该将评估报告通过全国互联网安全管理服务平台提交所在地地市级以上网信部门和公安机关（第 7 条）。互联网信息服务提供者在安全评估中发现存在安全隐患的，应当及时整改，直至消除相关安全隐患（第 6 条）。

根据《算法推荐管理规定》，具有舆论属性或者社会动员能力的算法推荐服务提供者应当按照国家有关规定开展安全评估（第 27 条）。

根据《深度合成管理规定》，深度合成服务提供者和技术支持者提供具有以下功能的模型、模板等工具的，应当依法自行或者委托专业机构开展安全评估：（一）生成或者编辑人脸、人声等生物识别信息的；（二）生成或者编辑可能涉及国家安全、国家形象、国家利益和社会公共利益的特殊物体、场景等非生物识别信息的。（第 15 条）。深度合成服务提供者开发上线具有舆论属性或者社会动员能力的新产品、新应用、新功能的，应当按照国家有关规定开展安全评估（第 20 条）。

根据《AIGC 暂行办法》，提供具有舆论属性或者社会动员能力的生成式人工智能服务的，应当按照国家有关规定开展安全评估（第 17 条）。

如前文所述，目前我国法律法规仅对“具有舆论属性或社会动员能力的互联网信息服务”予以界定，而对于何为具有舆论属性或社会动员能力的算法推荐服务、深度合成服务、生成式人工智能服务，我国法律法规并未给出定义。根据我们在过往项目中的经验，在实务中，对于何为“具有舆论属性或社会动员能力”的判断较为宽泛，几乎涵盖了所有具备信息共享功能的服务。因此，我们理解，一方面，大模型服务涉及“具有舆论属性或社会动员能力的互联网信息服务”的可能性较高，需按照《安全评估规定》通过全国互联网安全管理服务平台完成安全评估；另一方面，还需按照国家网信部门的要求，按照《AIGC 暂行办法》等法律法规的规定，满足针对大模型服务的特殊安全评估要求，包括主体安全保障、信息安全管理、用户安全、技术安全等等。

---

[2] 参见《苹果集中下架中国区 Chat GPT 相关产品，未进行算法备案与数据跨境不合规或为主因》，21 财经，2023 年 8 月 2 日，链接：https://m.21jingji.com/article/20230802/herald/37afec4f96019d56170767d2ccfe2259.html，最后访问日期：2024 年 3 月 18 日。

2. 内容合规

作为典型的互联网信息服务提供者，平台运营方需要承担我国法律对网络服务提供者设置的“监控义务”：一是审查义务，即在被明确告知违法信息存在之前，主动对其系统或网络中的信息的合法性进行审查；二是事后控制义务，即在知道违法信息的存在后及时采取删除、屏蔽等措施阻止侵权信息继续传播。除此之外，就用户输入数据与大模型服务生成内容（“服务生成内容”），平台运营方还面临着用户输入数据合规、服务生成内容合规和知识产权保护三方面的义务。

（1）用户输入数据合规

大模型的数据运用场景主要包括模型训练阶段对训练数据的使用以及模型使用阶段对输入数据的使用，且模型使用阶段收集的数据后续也可能成为新的训练数据。而平台运营方本身并不负责模型训练，故关于其数据合规义务的讨论，往往集中在模型使用阶段的输入数据。

平台运营方是典型的互联网服务提供者，需遵守《中华人民共和国网络安全法》（“《网络安全法》”）、《中华人民共和国数据安全法》（“《数据安全法》”）、《中华人民共和国个人信息保护法》（“《个人信息保护法》”）、《AIGC 暂行办法》等规定的网络安全、数据安全以及个人信息保护义务。关于平台运营方的数据合规相关义务，我们将在下文进行详细讨论。

此外，平台运营方作为深度合成服务提供者，需履行《深度合成管理规定》规定的用户输入数据审核义务，采取技术或者人工方式对用户的输入数据进行审核，识别违法和不良信息。

（2）服务生成内容合规

根据《AIGC 暂行办法》以及网络信息安全领域的监管要求，大模型平台运营方需要保证服务生成内容合规，承担对服务生成内容的审核义务，建立健全服务生成内容治理机制，依法设立辟谣机制、设立违法和不良信息识别特征库，积极承担信息内容管理主体责任，加强平台网络信息内容生态治理，培育积极健康、向上向善的网络文化；同时，当平台运营方发现违法内容的，应当及时采取停止生成、停止传输、消除等处置措施，并向有关主管部门报告。服务生成内容的具体合规要点详见下表：

序号	合规要点	具体内容
1	符合法律、行政法规，尊重社会公德、伦理道德	坚持社会主义核心价值观，不得生成煽动颠覆国家政权、推翻社会主义制度，危害国家安全和利益、损害国家形象，煽动分裂国家、破坏国家统一和社会稳定，宣扬恐怖主义、极端主义，宣扬民族仇恨、民族歧视，暴力、淫秽色情，以及虚假有害信息等法律、行政法规禁止的内容。
2	避免歧视	在算法设计、训练数据选择、模型生成和优化、提供服务等过程中，采取有效措施防止产生民族、信仰、国别、地域、性别、年龄、职业、健康等歧视。
3	尊重知识产权与公平竞争	尊重知识产权、商业道德，保守商业秘密，不得利用算法、数据、平台等优势，实施垄断和不正当竞争行为。
4	内容的真实性与准确性	基于服务类型特点，采取有效措施，提升生成式人工智能服务的透明度，提高生成内容的准确性和可靠性。
5	尊重他人合法权益	尊重他人合法权益，不得危害他人身心健康，不得侵害他人肖像权、名誉权、荣誉权、隐私权和个人信息权益。
6	服务生成内容标识义务	在生成或者编辑的信息内容的合理位置、区域进行显著标识，向公众提示深度合成情况，避免公众被混淆、误导。

针对平台运营方的服务生成内容标识义务，该规定主要针对的是目前服务生成内容难以被分辨，甚至出现技术被滥用、误用等问题，故标识的作用在于警示和提醒用户，确保用户明确知晓该内容是由大模型生成的，因此无法保证内容的真实性。大模型经过训练后，对一些概念具备了较为稳定的“认知”，围绕相关概念的生成内容往往表现出惊人的一致性。一旦模型在训练过程中引入偏见歧视等有害信息，在模型实际应用中很可能呈现负面的放大化效应，这是极为危险的。对此，有实务人士指出：“标识 AI 生成、深度合成的内容，是成本最低且有望从根本杜绝上述相关问题的方法。”[3]

（3）知识产权保护

如本书《ChatGPT 许可应用，知识产权和数据怎么看？》一文中所述，利用已有作品进行大模型训练的行为很难构成“合理使用”。因此，在服务生成内容生成过程中涉及与已有作品的接触且服务生成内容与已有作品存在实质性相似的情况下，服务生成内容可能涉及知识产权侵权。平台运营方作为网络服务提供者，应当尽到前述用户输入数据审核以及服务生成内容合规方面的义务，并履行《中华人民共和国民法典》第 1195 条规定的“通

[3] 参见《“生成式人工智能服务管理暂行办法”解读：明确“不适用”场景，充分“松绑”AI 发展》，央广网，2023 年 7 月 18 日，链接：http://m.cnr.cn/tech/20230718/t20230718_526333552.html，最后访问日期：2024 年 3 月 18 日。

知—删除”义务，否则可能因违反相应的注意义务而需承担共同侵权的责任。

3. 平台管理合规

根据《AIGC 暂行办法》等相关法律法规，平台运营方还需承担平台管理责任，具体要点如下：

序号	合规要点	具体内容
1	指导、保护用户义务	通过明确并公开其提供服务的适用人群、场合、用途，指导使用者科学理性认识和依法使用生成式人工智能技术，并且采取有效措施防范未成年人用户过度依赖或者沉迷生成式人工智能服务。
2	稳定服务义务	在其服务过程中，提供安全、稳定、持续的服务，保障用户正常使用。
3	违法整改义务	（1）发现违法内容的，应当及时采取停止生成、停止传输、消除等处置措施，采取模型优化训练等措施进行整改，并向有关主管部门报告； （2）发现使用者利用生成式人工智能服务从事违法活动的，应当依法依约采取警示、限制功能、暂停或者终止向其提供服务等处置措施，保存有关记录，并向有关主管部门报告。
4	建立健全投诉举报机制义务	建立健全投诉、举报机制，设置便捷的投诉、举报入口，公布处理流程和反馈时限，及时受理、处理公众投诉举报并反馈处理结果。

4. 网络安全与数据合规

对于平台运营方而言，在模型的使用阶段，其会收集各行业领域的不同类型的数据，因此，一方面，平台运营方需要保证对外提供的模型本身的合法合规，另一方面，面对收集和处理的海量数据，平台运营方还应当充分履行网络安全、数据安全以及个人信息保护相关义务。此外，

（1）模型数据来源合法性审查

虽然平台运营方本身不负责训练模型，但是作为直接面向消费者的生成式人工智能服务的提供者，平台运营方应当对模型的开发者即技术支持方开发提供的模型的数据来源合法性进行必要的审查，对技术支持方数据安全保护能力开展尽职调查。在平台运营方与技术支持方签署的相关技术服务合同中，平台运营方可以要求技术支持方对模型训练数据来源的合法合规性进行陈述保证，明确双方的权利义务，避免因技术支持方所提供的模型本

身的数据来源合法性问题影响平台运营方业务的持续开展。

（2）网络安全

《网络安全法》对作为网络运营者的企业提出的合规义务可以总结为两个方面：一方面，从网络运行安全的角度出发，要求网络运营者应当按照网络安全等级保护制度的要求，履行安全保护义务，保障网络免受干扰、破坏或者未经授权的访问，防止网络数据泄露或者被窃取、篡改。另一方面，从网络信息安全的角度出发，要求网络运营者应当对其收集的用户信息严格保密，并建立健全用户信息保护制度，并采取技术措施和其他必要措施，确保其收集的个人信息安全，防止信息泄露、毁损、丢失。根据《网络安全法》，只要是由运营软硬件设备组成的、按照一定的规则和程序对信息进行收集、存储、传输、交换、处理的信息系统的主体，均属于网络运营者。因此，平台运营方作为网络运营者也应当履行《网络安全法》项下的合规义务，在安全管理层面，平台运营方需在企业内部明确网络安全的责任，并通过完善的规章制度、操作流程为网络安全提供制度保障；在技术层面，平台运营方应当采取各种事前预防、事中响应、事后跟进的技术手段，应对网络攻击，从而降低网络安全的风险。

（3）数据安全

《数据安全法》从多方面规定了企业的数据安全保护义务，包括数据分类分级、安全管理制度、风险监测、风险评估等，面向消费者提供生成式人工智能服务的平台运营方作为《数据安全法》项下的数据安全合规主体，因此也应当履行《数据安全法》项下的合规义务，包括但不限于：对数据的重要程度、敏感程度等进行分级，并根据其重要程度、敏感程度的不同进行分级保护；建立健全全流程数据安全管理制度，组织开展数据安全教育培训，采取相应的技术措施和其他必要措施，保障数据安全；加强风险监测，发现数据安全缺陷、漏洞等风险时，应当立即采取补救措施等。

（4）个人信息保护

《个人信息保护法》规制个人信息全生命周期的保护和处理活动，要求企业应在个人信息的收集、存储、使用、加工、传输、提供、公开、删除等方面落实合规义务。面向消费者的生成式人工智能应用服务在个人信息保护方面与其他应用服务相比有很多相同之处，包括制定用户服务协议、隐私政策，明确处理用户数据的合法性基础。在此基础上，《AIGC 暂行办法》针对个人信息保护进一步规定，提供者对使用者的输入信息和使用记录应当依法履行保护义务，不得收集非必要个人信息，不得非法留存能够识别使用者身份的输入信息和使用记录，不得非法向他人提供使用者的输入信息和使用记录。提供者应当

依法及时受理和处理个人关于查阅、复制、更正、补充、删除其个人信息等的请求（第11条）。可以看出，个人信息保护已成为大模型合规的关注重点。

此外，个人信息的跨境传输问题也应当引起平台运营方的关注。根据《AIGC 暂行办法》，无论是中国境外的技术支持方直接面向中国境内公众提供生成式人工智能服务，还是平台运营方通过接入中国境外的 API 接口向中国境内公众提供生成式人工智能服务，均应当履行《AIGC 暂行办法》项下的合规要求。在此过程中，平台运营方很可能涉及将中国境内用户的个人信息传输至境外。在该等情形下，平台运营方还应当按照《个人信息保护法》《数据出境安全评估办法》《个人信息出境标准合同办法》等相关法律法规履行个人信息跨境传输相关的合规要求，并根据不同的场景选择合适的跨境传输方式。

（5）国际联网合规

根据《计算机信息网络国际联网管理暂行规定》及《工业和信息化部关于清理规范互联网网络接入服务市场的通知》，任何单位和个人不得自行建立或者使用其他信道进行国际联网，未经电信主管部门批准，个人、法人和其他组织不得自行建立或租用专线（含虚拟专用网络 VPN）等其他信道开展跨境经营活动，否则可能面临停止联网、警告、15000元以下的罚款及没收违法所得的行政责任。因此，平台运营方自行建立信道或租用未经电信主管部门批准建立的信道使用境外技术提供方提供的技术服务，将受到相应行政处罚。为了保证合规经营，避免不必要法律风险，平台运营方应该履行相应的申请手续，租赁使用合规的国际专线。

根据我国相关法律法规的规定，我国提供国际联网服务的经营者需要具有 A14-4 国际数据通信业务的基础电信业务经营许可证，目前仅有三大运营商，即电信、联通与移动具有该证照。部分电信运营企业可能会持有固定网国内数据传送业务（A24-1）或国内互联网虚拟专用网业务（B13）。尽管前述两项证照里都有 VPN 的字眼，但这两项证照不涉及 A14-4 国际数据通信业务，仅能在有限范围内提供 VPN 服务，不能提供跨境 VPN。因此，平台运营方应注意相关证照的具体范围，避免被证照名称中 VPN 的字眼所迷惑，确保供应商确有资格提供国际联网业务。

（二）技术支持方的合规要求

1. 资质证照

技术支持方作为算法推荐服务提供者、深度合成服务技术支持者以及生成式人工智能服务提供者，与平台运营方一样，需履行算法备案手续和安全评估义务。具体参见前序针对平台运营方的资质证照要求，在此不做赘述。值得注意的是，在实际备案和安全评估过

程中，技术支持方需填报的内容与平台运营方存在差异，例如，在算法备案的过程中，平台运营方需填报关联产品及功能信息，而技术支持方需填报技术服务方式，建议技术支持方予以关注。

2. 数据训练合规

数据训练是大模型技术存在的基础，是大模型应用的底层逻辑核心，数据是大模型最底层的“原料”，而数据训练是对“原料的使用”。因此，数据训练合规是满足服务生成内容合规、知识产权合规、个人信息合规等合规要素的重要前提。一直以来，数据训练合规都是大模型监管的重中之重。《AIGC 暂行办法》明确了生成式人工智能服务提供者在进行大模型训练时所应当履行的合规义务，其应当使用具有合法来源的数据和基础模型，不得侵害他人依法享有的知识产权，涉及个人信息的应当取得个人的同意或者符合法律、行政法规规定的其他情形。因此，在大模型数据训练环节，技术支持方首先应当确保训练数据来源的合法性，尤其应当关注训练数据中是否包含需要另行取得许可或授权的知识产权或个人信息等数据，对该问题的具体分析，可参见本书中《ChatGPT 许可应用，知识产权和数据怎么看？》一文。此外，与平台运营方一样，技术支持方在大模型训练环节同样也应当履行网络安全、数据安全和个人信息保护义务。值得注意的是，此次《AIGC 暂行办法》还对训练数据的质量和训练过程中的数据标注提出了更加明确的要求。

（1）数据质量要求

根据《AIGC 暂行办法》，生成式人工智能服务提供者应当依法开展预训练、优化训练等训练数据处理活动，遵守以下规定：……（四）采取有效措施提高训练数据质量，增强训练数据的真实性、准确性、客观性、多样性（第 7 条）。

模型的开发要求技术支持方必须借助大量的数据对模型进行训练，但是，与大量数据相比，良好的数据质量对于获得预期的最终结果至关重要。自生成式人工智能面世之初，其存在的“一本正经地胡说八道”现象便引起了人们的警惕。这种虚假信息的产生很可能会误导用户，加剧社会对共享信息的不信任。如何保障生成内容的真实性，既是产业界为进一步扩大生成式人工智能商用范围需要克服的技术难题，也是监管部门需要重点考量的问题[4]。而提高训练数据的质量，就是为了尽可能提高生成式人工智能的可靠性与可信度，进而有效堵住实际应用中的风险漏洞，避免生成式人工智能被错用、误用、滥用。

[4] 参见《专家解读｜推动生成式人工智能精细化治理》，中央网信网，2023 年 7 月 13 日，链接：http://www.cac.gov.cn/2023-07/13/c_1690898363806525.htm，最后访问时间：2024 年 3 月 18 日。

（2）数据标注

根据《AIGC 暂行办法》，在生成式人工智能技术研发过程中进行数据标注的，提供者应当制定符合本办法要求的清晰、具体、可操作的标注规则；开展数据标注质量评估，抽样核验标注内容的准确性；对标注人员进行必要培训，提升尊法守法意识，监督指导标注人员规范开展标注工作（第 8 条）。

数据标注是数据训练的关键环节。所谓数据标注，指的是对未经处理的语音、图片、文本、视频等原始数据进行加工处理，使其成为结构化数据让机器可识别的过程。因此，它决定着大模型最底层“原材料”的安全属性。但是，数据标注过程中，标注人员不可避免地会将个人意识投射至人工智能的算法逻辑中，而标注过程中的人为错误会导致数据质量变差，直接影响模型的性能和预测，因此制定清晰明确的标注规则、对标注人员进行培训均是提高人工智能生成内容的准确性和可靠性的必要措施。

（3）算法技术合规

除按规定履行算法备案手续，《算法推荐管理规定》《深度合成管理规定》《AIGC 暂行办法》等还为技术支持方设置了算法技术管理责任，有关算法技术管理责任的具体合规要点详见下表：

序号	合规要点	具体内容
1	反歧视机制	在算法设计、训练数据选择、模型生成和优化、提供服务等过程中，采取有效措施防止产生民族、信仰、国别、地域、性别、年龄、职业、健康等歧视。
2	算法机制机理审核	定期审核、评估、验证算法机制机理、模型、数据和应用结果；不得设置诱导用户沉迷、过度消费等违反法律法规或者违背伦理道德的算法模型。
3	公平竞争机制	不得利用“算法共谋”方式形成垄断，排除市场竞争，遵循反垄断、反不正当竞争相关法律规定。
4	提供必要支持和协助	有关主管部门依据职责对生成式人工智能服务开展监督检查，提供者应当依法予以配合，按要求对训练数据来源、规模、类型、标注规则、算法机制机理等予以说明，并提供必要的技术、数据等支持和协助。

结语

飞速发展的大模型给现代产业、教育、生活、娱乐、医疗领域带来了革命性发展。我们必须承认大模型已经成为现代社会进步与发展的必要工具。然而，大模型在大幅解放生产力的同时，相关的道德、伦理、法律等问题也备受关注。因此，对大模型的合规监管日益重要。基于对人工智能的规制不应限制技术而是防止其野蛮生长这一基本原则，世界各地探寻合理的监管与合规之策，而我国在初步形成大模型合规监管体系的基础之上，也将不断细化、深化该等监管体系。因此，包括平台运营方、技术提供方在内的各主体需及时关注相关合规法律动态，在拓展大模型的应用领域的同时，确保落实相应的合规要求。

感谢实习生张颖对本文作出的贡献。

大模型（Large Models）时代下资本市场赋能 AI 企业发展
——人工智能产业链企业境内 A 股上市
重点法律问题之实证分析

黄任重 张世源

前言

2024 年 2 月 15 日，人工智能文生视频大模型 Sora 正式对外发布，仅需简单的文字指令，Sora 就可快速生成一个长达 60 秒的高质量视频，视频的分辨率之高、内容之流畅均让人不由感叹。事实上，人工智能技术现已成为现代社会不可或缺的一部分，从文案创作、绘画海报到视频生成，从基础对话到智能服务，人工智能技术的高效便捷为人们的工作生活平添了许多色彩。自 2023 年以来，迎着 ChatGPT 所带来的热潮，百度、字节跳动、三六零、商汤科技、阿里巴巴、科大讯飞、云从科技、百川智能等科技企业相继推出其人工智能大模型产品，可服务于金融、工业制造、能源、政务、交通、教育等多个不同行业，应用于营销、创作、聊天、推理等多个场景，从而推动着人工智能产业步入大模型时代。

作为“新质生产力”的重要组成部分，人工智能现已成为当今国际竞争的新焦点，亦是经济发展的新引擎。“深化大数据、人工智能等研发应用，开展‘人工智能 +’行动，打造具有国际竞争力的数字产业集群”在《政府工作报告》中已被列为 2024 年政府工作任务之一。为进一步优化大模型时代下人工智能产业发展环境，提升产业创新能力和发展质量，我国各级政府已出台了相关的扶持政策。2023 年 5 月，北京市发布《北京市促进通用人工智能创新发展的若干措施》《北京市加快建设具有全球影响力的人工智能创新策源地实施方案（2023-2025 年）》，提出了“推动国产人工智能芯片实现突破”“加强自主开源深度学习框架研发攻关”“提升算力资源统筹供给能力”“加强公共数据开放共享”等主要任务，并探索和推动通用人工智能技术在政务服务、医疗、科学研究、金融、自动驾驶、城市治理等领域的示范应用。2023 年 5 月，深圳市发布《深圳市加快

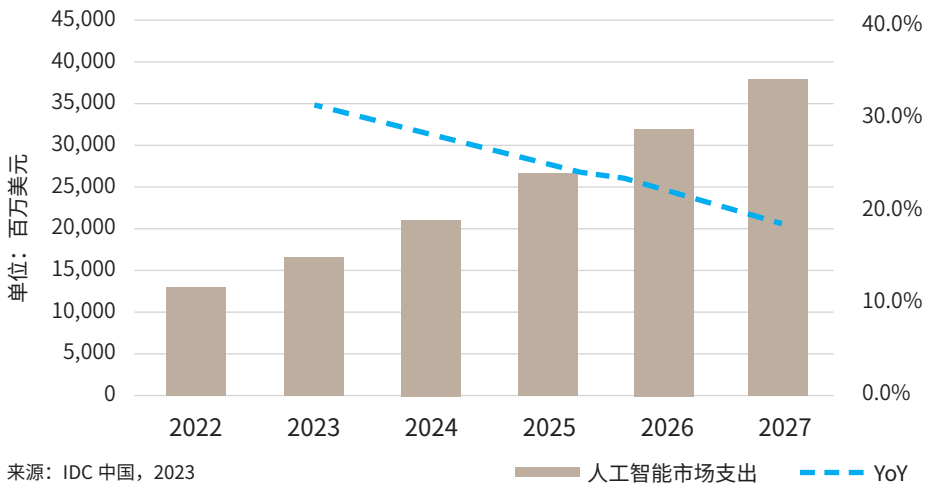
推动人工智能高质量发展高水平应用行动方案（2023—2024 年）》，提出了“强化智能算力集群供给”“增强关键核心技术与产品创新能力”“提升产业集聚水平”“打造全域全时场景应用”“强化数据和人才要素供给”及相关的具体要求。2023 年 11 月，上海市经济和信息化委员会等五部门联合制定并发布了《上海市推动人工智能大模型创新发展若干措施（2023-2025 年）》，提出了“实施大模型智能算力加速计划”“构建智能芯片软硬协同生态”“语料数据资源共建共享”“实施大模型示范应用推进计划”等重要措施。

一、人工智能行业发展及企业上市概况

（一）中国人工智能行业发展概况

根据工业和信息化部赛迪研究院发布的《2024 年我国人工智能产业发展形势展望》[1]及 IDC[2] 的相关研究分析报告 [3]，2022 年全球人工智能 IT 总投资规模为 1,288 亿美元，2023 年全球人工智能 IT 总投资规模预计将达到 1540 亿美元，同比增长 19.6%；2027 年预计增至 4,236 亿美元，五年复合增长率（CAGR）约为 26.9%；IDC 同时预计，2027 年中国 AI 投资规模有望达到 381 亿美元，全球占比约 9%。

中国人工智能市场支出预测，2022-2027



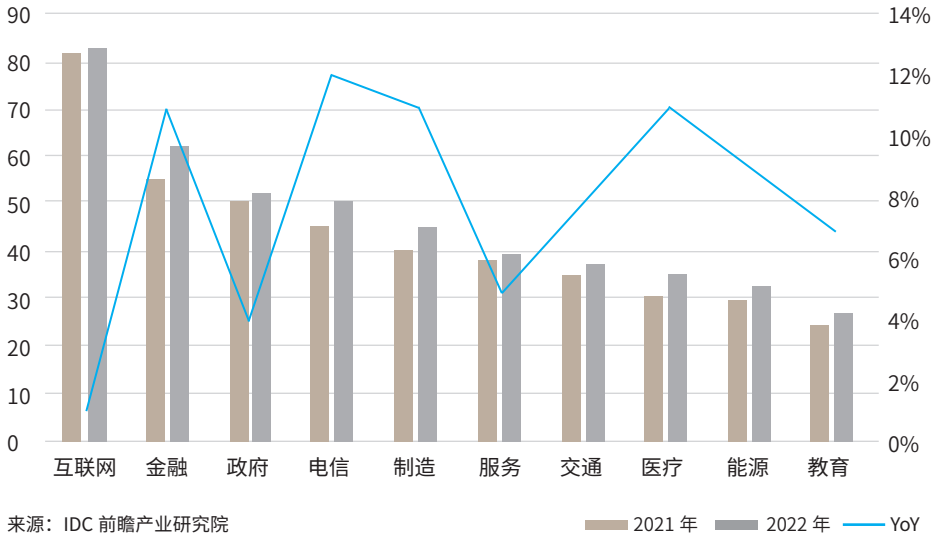
来源：IDC 中国，2023

根据 IDC 的统计，2022 年中国人工智能行业应用渗透度排名前五的行业依次为互联

[1] 摘自 http://www.csia-jpw.com/UserFiles/Article/file/638400559487735154798 3703.pdf，《2024 年我国人工智能产业发展形势展望》。
[2] IDC：International Data Corporation 国际数据公司。根据 IDC 官网介绍，IDC 是全球著名的信息技术、电信行业和消费科技咨询、顾问和活动服务专业提供商，成立于 1964 年，IDC 在全球拥有超过 1300 名分析师，为 110 多个国家的技术和行业发展机遇提供全球化、区域化和本地化的专业视角及服务。
[3] 摘自 https://www.idc.com/getdoc.jsp?containerId=prCHC51172823,《IDC 2027 年中国人工智能市场 IT 总投资规模预计超 380 亿美元》。

网、金融、政府、电信和制造[4]。

图表 :2021-2022 中国各行业人工智能渗透度



来源：IDC 前瞻产业研究院

根据工业和信息化部赛迪研究院发布的《2024 年我国人工智能产业发展形势展望》[5]，截至 2023 年 11 月，国产大模型有 188 个，其中通用大模型 27 个，目前已有超 20 个大模型获得备案，大多数已向全社会开放服务；基于 2200 家人工智能骨干企业的关系数据量化分析表明，我国人工智能已广泛赋能智慧金融、智慧医疗、智能制造、智慧能源等 19 个应用领域。

（二）人工智能产业链分类及人工智能企业上市概况

1. 人工智能产业链

人工智能产业链上游是基础层，主要指计算平台和数据中心提供数据和算力支持，包括 AI 芯片、云计算、传感器、数据类服务等技术；中游是技术层，指 AI 算法和技术，主要技术包括计算机视觉（图像识别与分析）、智能语音和自然语言处理技术（语音识别与合成）、机器学习与深度学习（分析决策及行动）等；下游是应用层，是将人工智能技术和具体场景结合研发的产品和服务，是人工智能产业的自然延伸，主要应用场景有医疗、交通、金融、家居、教育、安防等方面。人工智能产业链示意图如下：[6]

---

[4] 摘自 https://t.qianzhan.com/caijing/detail/231214-d1f521f1.html，《工信部赛迪研究院：2023 年中国生成式人工智能市场规模有望突破 10 万亿元【附 AIGC 行业发展前景预测】》。

[5] 摘自 http://www.csia-jpw.com/UserFiles/Article/file/638400559487735154798370 3.pdf，《2024 年我国人工智能产业发展形势展望》。

[6] 摘自《中科院成都信息技术股份有限公司 2023 年半年度报告》，公告于 2023 年 08 月 29 日，http://www.cninfo.com.cn/new/disclosure/detail?plate=szse&orgId=9900030367&stockCode=300678&announcementId=1217675032&announcementTime=2023-08-29.

2. 人工智能产业链企业上市概况

自 2019 年科创板开板以来，越来越多的人工智能产业链企业登陆境内资本市场；已上市公司亦充分利用资本市场带来的融资便利性，积极开展人工智能技术与应用产品的研发。截至 2023 年 12 月末，涉及人工智能产业链的境内 A 股上市公司已超过 400 家，业务领域上涵盖了人工智能产业链的各个环节，包括上游基础层的人工智能传感器、人工智能芯片的研发与制造，中游技术层及下游应用层的人工智能算法开发、人工智能软硬件产品的研发与制造、“智慧城市”“智慧金融”等行业解决方案供应商。

除境内 A 股外，港交所也是人工智能产业链企业走向资本市场的重要战场。2023 年下半年以来，人工智能领域头部企业第四范式、智能机器人企业优必选相继于香港联交所主板成功挂牌上市。此外，截至 2024 年 2 月末，智能可穿戴设备企业“出门问问”、企业级交互式人工智能解决方案企业声通科技等多家人工智能产业链企业已递交港交所上市的申请。

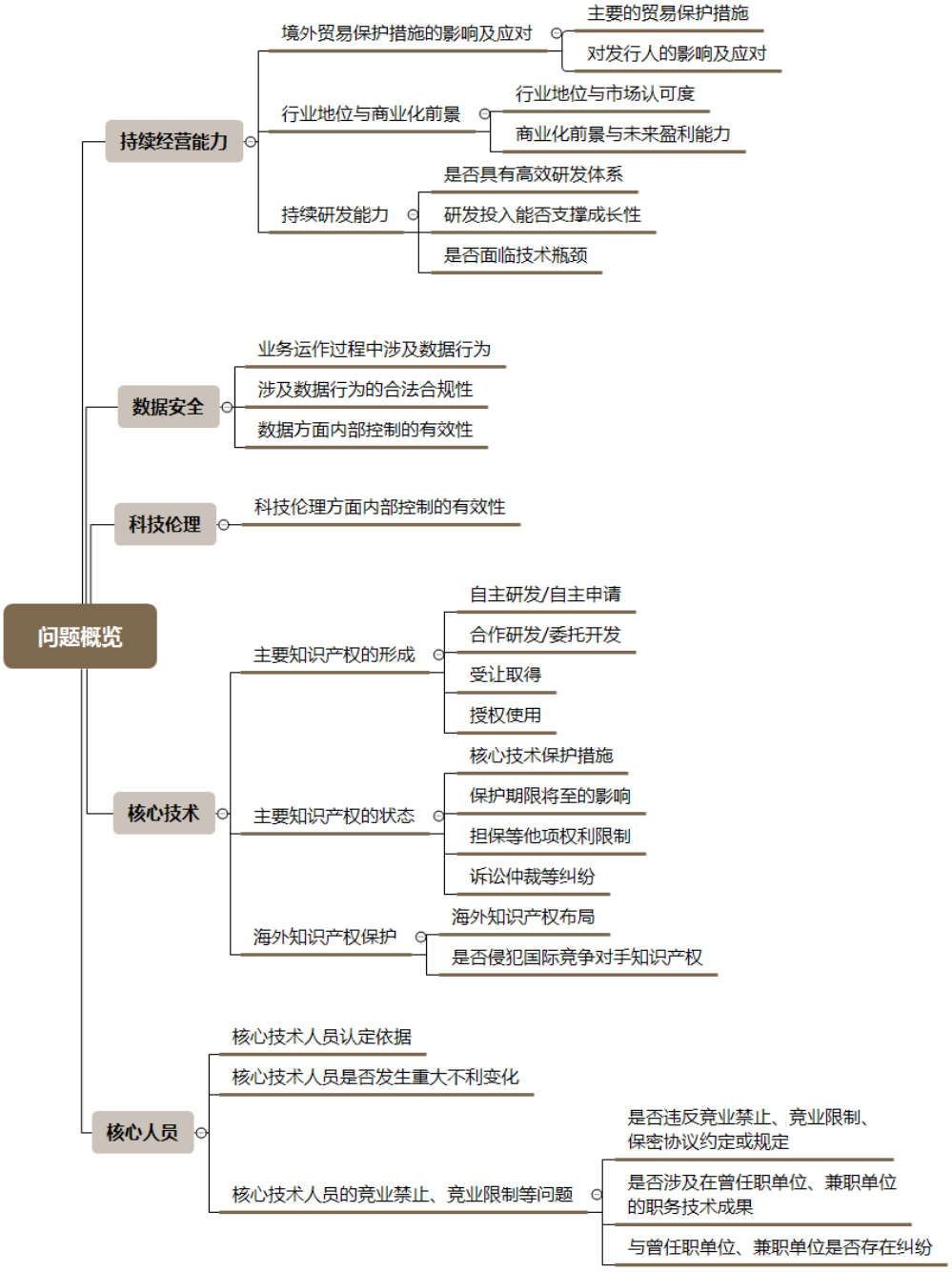
在资本市场的加持下，我国的人工智能行业正在飞速发展。

二、人工智能产业链企业境内 A 股上市重点法律问题之实证分析

本文将从“持续经营能力”“数据安全”“科技伦理”“核心技术”及“核心人员”

五个核心角度出发，对人工智能产业链企业在境内 A 股上市审核中最受关注的问题进行解析梳理，协助人工智能产业链企业直面上市时的挑战。

为方便阅读与理解，我们首先将后文将要涉及的问题整理相应的问题概览如下：



（一）持续经营能力

“具备持续经营能力”是境内 A 股上市的发行条件之一。《首次公开发行股票注册管理办法》明确，“发行人业务完整，具有直接面向市场独立持续经营的能力。”因此，人工智能产业链企业在上市审核过程中需要重点证明自己具有一定的行业地位、市场认可度、研发能力和商业化能力，且不存在经营环境已经或者将要发生重大变化等对持续经营有重大不利影响的事项。

1. 境外贸易保护措施的影响及应对

人工智能作为引领未来的战略性技术，已成为国际竞争的新焦点。世界主要发达国家把发展人工智能作为提升国家竞争力、维护国家安全的重大战略。近年来，以美国为代表的发达国家持续对我国多家人工智能产业链企业采取一系列贸易保护措施，主要包括以下：

序号	管制措施	具体影响
1	美国：列入《出口管制条例》（EAR）实体清单	从美国或其他国家进口美国原产的商品、技术或软件受到限制；进口美国管制物项价值占比超过一定比例的其他国家商品受到限制；进口利用美国原产技术或软件直接生产或利用美国原产技术或软件建设的工厂生产的产品受到限制
2	美国：列入“非 SDN- 中国军工复合体企业”（NS-CMIC）清单	限制美国投资者对清单上的主体投资
3	美国：对《美国出口管制条例》（EAR）进行针对性修订	限制中国等特定国家获得先进计算集成电路、开发和维护超级计算机以及制造先进半导体的能力，如阻止英伟达等公司向特定国家出口先进的 AI 芯片

由此可见，人工智能产业链企业可能在原材料供应、投融资等多方面受到境外贸易保护措施的不利影响，在上市审核的过程中将引发监管机构对其持续经营能力的质疑，需要拟上市企业充分解释不存在经营环境已经或者将要发生重大变化等对持续经营有重大不利影响的事项。

关注方面	具体要点
境外贸易保护措施的影响与应对	• 明确披露境外贸易保护措施对生产经营的具体限制，并在招股说明书中进行风险提示。 • 说明对原材料供应的具体影响，该等原材料是否构成产品的核心零部件，并说明保障核心原材料稳定性的具体措施及替代措施。 • 上述保障核心原材料稳定性的措施一般包括： – 在被列入“实体清单”之前，对生产经营所需的境外关键器件和产品进行了一定储备； – 积极与国内厂商开展合作，推进关键器件和产品的国产替代准备工作； – 在采购受管制产品时，需要供应商提供签署的《产品供应合规声明》，对没有签署或不能提供《合规声明》的供应商，将不会从其处采购任何受《美国出口管理条例》管制的“物项”； – 根据商务部于 2021 年 1 月发布的《阻断外国法律与措施不当域外适用办法》，如中国政府认定美国商务部将发行人列入实体清单存在不当域外适用情形的，则中国商务部可以发布不得承认、不得执行、不得遵守美国商务部关于将发行人列入实体清单的禁令。 • 说明境外销售的具体地区、产品，披露境外贸易保护措施对公司境外业务拓展的影响及应对措施。 – 解释不存在直接影响：仅限制被列入“实体清单”的企业获取涉美软硬件产品、技术，并未限制客户向发行人购买产品或服务，不影响客户正常购买发行人的产品或服务，客户不会因购买发行人提供的产品或服务而受到美国制裁； – 解释潜在的间接影响：可能会对发行人未来在人工智能前沿理论及学术研究和国际学术交流以及境外业务拓展产生一定不利影响。

2. 行业地位与商业化前景

拟上市企业为了强调自身的产品与技术优势，会在披露文件中重点渲染自身技术优势及行业内领先地位等。但这可不是随便说说这么简单，监管机构往往要求公司用具体依据来论证公司的行业地位与市场认可度，以说明企业并非经“粉饰包装”或“拼凑技术”的“伪科技”企业。对于人工智能产业链上游（即基础层）、中游（即技术层）的企业而言，目前普遍存在高研发投入、盈利欠佳、长期亏损的情况，在当前“以提高上市公司质量为导向，研究提高上市财务指标”的背景及“进一步从严审核未盈利企业，要求未盈利企业充分论证持续经营能力、披露预计实现盈利情况，就科创属性等逐单听取行业相关部门意见”的审核政策下，监管机构将重点考察其商业化前景及未来实现盈利的可能性。

关注方面	具体要点
行业地位与市场认可度	关于公司市场地位以及产品技术的领先性通常需要充分、具体的依据加以佐证，主要角度包括： • 市场与技术角度 – 所处细分行业市场空间、行业集中度和技术壁垒等情况； – 细分产品的市场占有率及其同行业可比公司对比情况； – 主要竞争对手情况，及不同产品与可比公司产品的技术优劣势； – 公司核心技术独特性、领先性及突破点，与最高技术水平的差距等； – 引用的技术指标是否完整、客观、全面。 • 权威第三方佐证 – 所获奖项的含金量，如来自权威机构评选及说明其他参与评选单位的情况等； – 主导或参与编制的相关行业标准等； – 权威机构的行业研报等文件中关于公司排名情况。
商业化前景与未来盈利能力	• 说明所在应用领域的各细分应用场景，及其目前发展阶段、商业化特点； • 解释报告期内公司持续亏损的原因； • 说明各细分应用场景下商业化的主要影响因素； • 说明在手订单情况、营收变化趋势以及未来业务发展方向； • 说明商业化拓展策略以及为实现盈利拟采取的措施。

3. 持续研发能力

持续研发能力的重要性对人工智能产业链企业长远发展的重要性不言而喻，监管机构会重点关注企业技术方案的长期发展战略，应对技术更迭和滞后的措施等。

关注方面	具体要点
是否具有高效研发体系	• 研发管理角度：介绍研发立项、设计、验证、试量等环节管理体系的完备性； • 人才培养角度：从人才招聘及内部培训等角度说明公司从人才角度保障研发实力； • 激励机制角度：比如通过股权激励、员工持股计划等激励措施发挥研发人员的积极性； • 知识产权管理：从知识产权管理与保护角度说明公司打造的知识产权体系与核心技术体系的成熟性。

关注方面	具体要点
研发投入能否支撑成长性	• 对于研发费用占比较低的企业，需要着重解释研发费用占比较低的原因及合理性，比如技术方向的特点及同行业公司对比情况等； • 基于公司业务特点，论述研发投入与营业收入增长的正向对应关系，以及未来研发投入的计划，说明公司研发投入能够支撑公司成长性与技术先进性。
是否面临技术瓶颈	• 充分说明技术研发进展情况，是否存在重大技术障碍、被其他新技术取代的风险以及依赖单一技术的风险； • 公司在特定时期未取得或取得较少专利可能会被认为遇到技术瓶颈，公司需要对背景原因进行合理解释，并且论证说明公司研发工作的有效开展情况及技术成果产出的持续稳定情况。

（二）数据安全

人工智能产业与数据的关系密不可分。对于人工智能产业链企业，在算法研发及相关人工智能系统搭建的阶段，可能通过训练数据的采集和运用以实现模型训练以及算法精度、效果的验证；在客户使用发行人产品的过程中，亦存在发行人采集或处理数据的可能。近年来《网络安全法》《数据安全法》《个人信息保护法》《互联网信息服务深度合成管理规定》《生成式人工智能服务管理暂行办法》等法律法规相继颁布与实施，监管机构在上市审核中亦充分关注拟上市企业对数据的采集、管理、运用等是否合法合规。

关注方面	具体要点
业务运作过程中涉及数据行为	• 说明发行人各个业务阶段及业务类别上是否涉及个人信息等数据的采集和运用； • 说明是否存在数据销售的行为； • 说明采集数据的具体来源（如公司自行采集、委托第三方供应商采集）以及各来源的区别、占比。
涉及数据行为的合法合规性	• 说明是否存在数据方面的侵权行为、诉讼纠纷或潜在诉讼纠纷、可能被处罚的情形； • 说明是否取得从事数据服务所需的全部资质、许可或备案； • 走访相关主管部门（网信部门、公安机关网络安全部门等）或获取其出具的合法合规证明； • 必要时可聘请律师事务所等第三方机构对数据合规及相关内部控制情况出具专业意见，并根据专业意见对数据管理不完善或不合规的情形进行针对性整改完善； • 必要时就法规政策变动对公司业务的影响进行风险提示。

关注方面	具体要点
数据方面内部控制的有效性	• 措施可以包括： – 制定数据安全与合规相关内部制度规则体系； – 设立专门的内部组织机构，全面负责数据安全工作； – 完善并提升相关的技术手段； – 对委托第三方供应商采集的数据，要求数据供应商签署含数据来源合法合规确认条款的协议或书面确认，后续持续跟踪其实际履行情况； – 对于发行人自行采集的数据，采集前对被采集网站是否限制自动化采集等情况做综合评估，并定期检查被采集网站的规则或规定是否变化。

（三）科技伦理

科技伦理方面内部控制的有效性一直以来是人工智能产业链企业在上市审核过程中的重点问询事项。2023 年 12 月 1 日起生效的《科技伦理审查办法（试行）》已经明确要求“从事生命科学、医学、人工智能等科技活动的单位，研究内容涉及科技伦理敏感领域的，应设立科技伦理（审查）委员会；单位科技伦理（审查）委员会无法胜任审查工作要求或者单位未设立科技伦理（审查）委员会以及无单位人员开展科技活动的，应书面委托其他满足要求的科技伦理（审查）委员会开展伦理审查。”因此，对于拟境内 A 股 IPO 的人工智能产业链企业，如研究内容涉及科技伦理敏感领域的，应结合最新规定的相关要求，对相关的内控制度建设予以完善并严格落实。

关注方面	具体措施
科技伦理方面内部控制的有效性	• 设立或书面委托符合《科技伦理审查办法（试行）》要求的科技伦理（审查）委员会，作为总体负责人工智能技术的伦理规范审查的内部组织机构： – 人数应不少于 7 人，设主任委员 1 人，副主任委员若干； – 委员任期不超过 5 年，可以连任； – 委员会成员应由具备相关科学技术背景的同行专家以及伦理、法律等相应专业背景的专家组成，并应当有不同性别和非本单位的委员，民族自治地方应有熟悉当地情况的委员； – 制定章程，建立健全审查、监督、保密管理、档案管理等制度规范、工作规程和利益冲突管理机制； – 科技伦理（审查）委员会作出的审查决定，应经到会委员的三分之二以上同意； – 应在设立科技伦理（审查）委员会后 30 日内，通过国家科技伦理管理信息登记平台进行登记。 • 建立伦理审查内部制度，并结合公司实际情况制定针对业务运作各个阶段（如分为立项 / 设计 / 开发 / 交付阶段）的程序性文件，以明确各个阶段具体的操作流程。 • 不断完善优化相关的技术手段。

（四）核心技术

人工智能产业链企业作为技术密集型企业，其核心技术的形成及稳定是境内 A 股上市过程中审核要点的重中之重。对于拟申报科创板 IPO 的人工智能产业链企业，更应当凸显其“硬科技”特色及科创属性要求，以论证其符合板块定位。在这里我们从主要知识产权的形成、主要知识产权的状态及海外发明专利三个角度展开，分别解析几个问题的关注要点和注意事项。

1. 主要知识产权的形成

人工智能产业链企业使用的知识产权主要来源包括自主研发、合作研发 / 委托开发、受让取得、授权使用等，监管机构对不同取得方式的关注各有侧重，但核心还是要确保发行人自主、独立、完整地拥有核心技术，能够依靠核心技术持续稳定经营。

取得方式	关注要点
自主研发 / 自主申请	• 自主研发的关注度相对较小，企业能够清晰说明自主研发及申请的具体过程，体现自身具备成熟研发管理体系即可； • 但需注意完善内部知识产权管理制度，明确员工所产生技术成果的所有权归属。
合作研发 / 委托开发	• 关注合作研发 / 委托开发的协议约定内容，特别关注协议中关于研发成果的归属是否有明确约定； • 涉及企业核心技术的合作研发 / 委托开发成果建议尽可能约定归属于发行人所有； • 关注合作研发 / 委托开发项目的研发成果在发行人主要产品及在核心技术中的运用情况； • 如涉及共有知识产权的，则需注意各方对于共有知识产权的占有份额以及共同知识产权的权利行使（使用、许可、转让、维权等）是否有明确约定，避免使用过程中存在潜在纠纷； • 关注对合作单位是否存在技术依赖，是否采取适当保密措施。
受让取得	• 关注受让取得知识产权的过程，如转让背景、协议履行情况及价格是否公允，确保发行人取得相关专利不存在潜在纠纷； • 确认受让取得的知识产权无他项权利安排，确保发行人技术资产完整性； • 转让方系科研院所、事业单位的，关注简易程序是否合法合规； • 关注受让取得的知识产权与发行人核心技术的关系，是否对第三方主体存在技术依赖，能否确保发行人核心技术独立性。
授权使用	• 授权的稳定性：关注授权协议的安排，确保企业能够长期稳定使用被授权的相关技术； • 授权价格的公允性：确保授权费用的构成及支付安排符合一般市场惯例和标准，不存在潜在利益输送； • 对授权技术的依赖性：确保使用授权技术的产品或服务在发行人主营业务收入中比重较小，避免产生关于技术依赖的怀疑。

2. 主要知识产权的状态

除了知识产权的形成方式外，监管机构也同样关注人工智能产业链企业对知识产权的有效保护及是否可以持续稳定地使用这些知识产权，确保核心技术能够得到有效的保护，使用过程中不存在纠纷或障碍。

关注方面	具体要点
核心技术保护措施	• 制度角度：建立完善内控制度，对涉及技术秘密事项的操作流程和保护措施进行详细规定； • 人事角度：如聘请内部知识产权专员对知识产权申请、维护进行监管，与员工签署保密协议等； • 合同角度：在销售合同中约定“保密信息”“知识产权赔偿”等保护性条款； • 技术角度：通过权限管理、外网隔离等技术处理措施防止技术泄密。
保护期限将至的影响	• 论证保护期将至或已到期的知识产权是否属于核心技术； • 相关知识产权对应产品的销售收入所占营业收入比重，保护期到期是否对企业持续经营产生重大不利影响； • 论证如果不能基于知识产权限制竞争对手或通过授权方式取得授权使用费对企业生产经营的影响。
担保等他项权利限制	• 尽可能确保核心技术、知识产权上不存在担保等权利限制。 • 如有他项权利限制，需说明相关背景、合理原因，并结合在发行人主要产品、核心技术中的运用情况及重要程度分析说明不会对经营产生重大不利影响。
诉讼仲裁等纠纷	• 避免或消除关于核心技术、知识产权的诉讼仲裁等纠纷； • 所持知识产权如涉诉的，需充分论证相关涉诉知识产权的重要性、对应产品收入的比重，说明相关诉讼纠纷对公司生产经营不存在重大不利影响、（如申请上海证券交易所科创板 IPO）不会导致发行人不符合专利数量相关的科创属性指标。

3. 海外知识产权保护

在国际竞争不断加剧、知识产权保护日趋严格的环境下，包括人工智能行业在内的技术型企业正在不断完善海外专利布局，为企业产品进入海外市场保驾护航，也借此积累专利实力，抗衡或制约竞争对手。因此，人工智能产业链企业在海外的知识产权保护也是上市过程中的重点关注事项。

关注方面	具体要点
海外知识产权布局	• 说明公司海外知识产权布局的情况，以确保海外市场销售的产品得到有效的知识产权保护，如有纠纷也不会影响境外销售。
是否侵犯国际竞争对手知识产权	• 说明海外市场的产品是否有侵犯竞争对手专利的风险； • 说明没有被当地知产管理部门认定侵犯第三方知识产权，没有侵犯第三方知识产权相关的诉讼仲裁等纠纷。

（五）核心人员

2023 年年末，OpenAI 创始人萨姆·阿尔特曼（Sam Altman）和 OpenAI 的解雇风波引起全世界的关注。其在五天内自被免除 OpenAI 的 CEO 和董事会主席职务，到微软公司宣布其加入微软并领导一个新的高级人工智能研究团队，再到重返 OpenAI 担任 CEO 的经历，足以证明核心骨干技术人员对于人工智能产业链企业具有的无可替代的重要性。作为核心技术研发的人力保障和重要因素，核心技术人员的构成、稳定性及历史任职经历均是重点关注的问题。

1. 核心技术人员认定依据

对于拟申请在上海证券交易所科创板上市的人工智能产业链企业，需要对其核心技术人员作出认定，并要求相关人员遵守核心技术人员相关的特殊规定（如锁定期要求）。关于核心技术人员的认定是较为基础的关注问题，但也不容轻视，监管机构可能会反向询问部分人员没有被认定为核心技术人员的原因。

关注方面	具体要点
核心技术人员的认定依据	可以参考使用的主要认定依据有： • 在公司就职的期限及资历背景，具有一定的学历和科研背景； • 取得的专业资质及获得奖项情况； • 在公司担任的职务，如在研发、设计等岗位上担任重要职务； • 任职期间取得的成果，主要产品研发过程中发挥的作用，如主导完成多项核心技术的研发，带领业务团队完成多项专利设计的申请、负责或参与行业主要技术标准的起草与制定。

2. 核心技术人员的变化

科创板的发行条件中明确要求发行人的“核心技术人员应当稳定且最近二年内没有发生重大不利变化”，发行人核心技术人员的稳定性是企业持续稳定经营及维持核心竞争力

的关键因素。

关注方面	具体要点
核心技术人员是否发生重大不利变化	• 说明最近 2 年内的变动人员情况，并计算变动人数所占人员合计总数的比例； • 核查原核心技术人员在任职期间承担的具体研发项目、进展情况、具体研发成果及交接情况，并分析其离职或无法正常参与生产经营是否会对发行人产生重大不利影响。

3. 核心技术人员的竞业禁止、竞业限制等问题

人工智能行业内高端技术人才流动较为常见，拟上市企业中核心技术人员往往具有曾于同行业企业、科研院所任职的背景，亦有部分核心技术人员为相关科研院所的兼职人员。核心技术人员于同行业单位任职的相关的竞业禁止、竞业限制等问题也是上市过程中的重点关注问题。

关注方面	具体要点
是否违反竞业禁止、竞业限制、保密协议约定或规定	• 明确披露相关人员任职履历情况，及与原任职单位、兼职单位签订竞业禁止 / 限制协议、保密协议的情况； • 说明相关竞业禁止 / 限制协议、保密协议所约定的期限、限制范围等内容，确保相关人员在发行人处任职不会违反签署协议相关约定； • 确认相关人员没有因为违反劳动合同、竞业禁止 / 限制协议、保密协议与原任职单位、兼职单位发生纠纷或存在潜在纠纷； • 核心技术人员为科研院所等事业单位兼职人员的，确认在发行人的任职符合事业单位人员兼职的相关规定。
是否涉及在曾任职单位、兼职单位的职务技术成果	• 核查明确相关人员在原单位、兼职单位任职所产生的技术成果归属情况； • 分析论述公司的产品、技术与相关人员于曾任职单位、兼职单位所参与研发的产品、技术存在差异； • 确认没有因为侵犯曾任职单位、兼职单位知识产权而发生的诉讼仲裁等争议纠纷或潜在争议纠纷。
与曾任职单位、兼职单位是否存在纠纷	• 没有纠纷自然是最好的，但也不排除上市过程中收到同行冤家的狙击，同行可能从民事诉讼、实名举报甚至报警立案等多维度进行打击，遇到这种情况拟上市公司只能见招拆招、刚柔并进； • 当然，创始团队或核心技术人员在从上家离职或分家时也要注意完整履行相关程序、保留好协议等证据，避免日后的潜在纠纷。

结语

自 ChatGPT 于 2022 年 11 月上线以来，在引起大众欢呼与热议的同时，也推动着人工智能产业步入大模型时代。2024 年 2 月 Sora 大模型的发布更是让人工智能领域迎来了一次重大突破。作为“十四五”数字经济发展规划中的战略性前瞻性领域及“新质生产力”的重要组成部分，人工智能产业在各级政府的扶持政策下蓬勃发展。自 2023 年 2 月起全面实行的股票发行注册制，拓宽了人工智能产业链企业的融资渠道，充分发挥了资本市场对新质生产力企业的支持作用。根据上文分析，对人工智能产业链企业而言，“持续经营能力”“数据安全”“科技伦理”“核心技术”及“核心人员”等五方面是境内 A 股上市过程中监管机构关注的重点问题。在如今“综合考虑二级市场承受能力，实施新股发行逆周期调节”的背景下，建议人工智能产业链企业在启动上市前，尽早在中介机构协助下，对企业问题进行全面的诊断，对企业优势进行挖掘与凸显，以可控的成本满足上市条件，及早利用资本的力量推动关键技术领域的突破，实现企业腾飞。

人工智能（AI）：科技伦理治理走起

张逸瑞 冯宝宝 张一凡

引言

随着人工智能深度学习能力的不断强化，人工智能技术大幅发展，ChatGPT、Gemini 等 AIGC 产品的陆续推出，人工智能技术开始被广泛地应用于人们日常生活的方方面面。不过，正如过去诸多科幻类作品所展示的那样，现实中，人工智能的发展带来的科技伦理问题也逐步显现。例如，当人工智能使用的训练数据本身存在虚假、错误信息时，或人工智能被用于炮制虚假、错误信息时，人工智能将助推虚假、错误信息的传播。再如，尽管人工智能技术研发的初衷是使人们得以从一些简单重复的工作中脱离出来进而从事更具有创造性的工作，人工智能自动生成的绘画、诗句、文章展现出的出乎人们意料的创造力，引发了社会对于人工智能取代人类的忧虑以及对就业市场的巨大冲击。在近期的好莱坞编剧罢工事件中，为了避免被 AI 取代，好莱坞编剧们提出了 AI 生成的文学材料不得被视为人类撰写的文学材料、禁止利用编剧的劳动成果来训练 AI 等一系列诉求。再如，人工智能系统可能包含强化或固化歧视或偏见的应用程序和结果，这将加剧社会中已有的歧视、偏见与成见。此外，人类为使用人工智能提供的服务，也将涉及向人工智能提供生理信息、行为偏好、兴趣偏好等个人隐私信息，如前述信息被不当收集和利用，人工智能将极有可能成为窥探个人隐私、侵扰个人生活的工具。

为了应对上述人工智能带来的伦理问题，联合国教育、科学及文化组织于 2021 年 11 月 23 日通过《人工智能伦理问题建议书》（Recommendation on the Ethics of Artificial Intelligence，“《建议书》”），提出了人工智能系统生命周期的所有行为者应尊重的价值观和原则以及落实前述价值观和原则的政策行动，建议会员国在自愿基础上适用《建议书》的各项规定，根据本国的宪法实践和治理结构并依照国际人权法在内的国际法采取适当步骤，包括进行必要的立法。目前，欧盟、美国和英国等国家和地区均已出台

了一系列监管规则，与此同时，我国科学技术部会同教育部、工业和信息化部、国家卫生健康委等十部门联合印发的《科技伦理审查办法（试行）》正式明确了我国的科技伦理审查体系。

一、国外 AI 伦理治理监管体系概述

目前欧盟、美国及英国均已出台了与人工智能相关的监管规则，科技伦理的治理是其中的重点。具体如下：

（一）欧盟

《人工智能法案》（Artificial Intelligence Act）作为欧盟人工智能监管体系的核心，在经历了一系列修正之后，目前将进入欧盟委员会、议会和成员国三方谈判协商的程序从而确定最终版本。《人工智能法案》是欧盟首部有关人工智能的综合性立法，其以人工智能的概念作为体系原点，以人工智能的风险分级管理作为制度抓手，以人工智能产业链上的不同责任主体作为规范对象，以对人工智能的合格评估以及问责机制作为治理工具，从人工监管、隐私、透明度、安全、非歧视、环境友好等全方位监管人工智能的开发和使用，详细规定了人工智能市场中各参与者的义务。

在伦理治理方面，《人工智能法案》强调，人工智能应该是一种以人为本的技术，不应该取代人类的自主性，也不应该导致个人自由的丧失，而应该主要服务于社会需求和共同利益，因此应提供保障措施，以确保开发和使用尊重欧盟价值观和《欧洲联盟基本权利宪章》（Charter of Fundamental Rights of the European Union）的道德嵌入式人工智能。对于 AI 系统的风险分级标准，《人工智能法案》将伦理风险作为考量因素，将下述类型的 AI 系统归为“存在不可接受风险的 AI 系统”，在欧盟成员国内将完全禁止该等 AI 系统投入市场或者交付使用：

- 采用潜意识技术或有目的的操纵或欺骗技术；
- 利用个人或社会群体的弱点（例如已知的人格特征或社会经济状况、年龄、身体精神能力）；
- 利用人的社会行为或人格特征进行社会评分；
- 在公众场所的“实时”（包括即时和短时延迟）远程生物识别系统。

此外，在评估 AI 系统是否属于“高风险 AI 系统”时，《人工智能法案》要求考量 AI 系统对《欧洲欧盟基本权利宪章》所保护的基本权利造成的不利影响的程度，该等基本权利包括：人的尊严、尊重私人和家庭生活、保护个人数据、言论和信息自由、集会

和结社自由以及不受歧视的权利、受教育权、消费者保护、工人权利、残疾人权利、性别平等、知识产权、获得有效补救和公平审判的权利、辩护权和无罪推定、良好管理的权利。“高风险 AI 系统”投放市场及交付使用均受到严格的管控并需履行评估及备案等一系列要求。

（二）美国

美国在联邦层面尚未通过一部完整且专门的针对 AI 系统的法案，而是试图通过调整政府机构的权利，在现有的立法框架及监管规则内对人工智能进行规制。在伦理治理方面，目前联邦层面的合规重点主要涉及反歧视、保护数据隐私等要求。例如：

1.《2022 年算法问责法案》（Algorithmic Accountability Act of 2022）

2022 年 2 月，美国众议院颁布了《2022 年算法问责法案》，要求使用自动化决策系统做出关键决策的企业研究并报告这些系统对消费者的影响，其内容包括是否会因为消费者的种族、性别、年龄等生成对消费者有偏见或歧视性的自动决策等。该法案形成了“评估报告—评估简报—公开信息”三层信息披露机制。此外，联邦贸易委员会还将建立可公开访问的信息存储库，公开发布关于自动化决策系统的有限信息。

2.《人工智能权利法案蓝图》（Blueprint for an AI Bill of Right）

2022 年 10 月，美国白宫科技政策办公室（OSTP）颁布了《人工智能权利法案蓝图》，提出了指导人工智能的设计、使用和部署的五项原则：技术的安全性和有效性、防止算法歧视、保护数据隐私、告知及解释义务以及人类参与决策，并在技术指南部分针对五项原则中的每一项均解释了原则的重要性、原则所指引的期望以及各级政府到各种规模的公司等多种组织为维护原则可以采取的具体的实施步骤、原则的实践案例。

3.《美国数据隐私和保护法案》（the American Data Privacy and Protection Act，“ADPPA”）

2022 年 6 月，美国参众两院共同发布了 ADPPA，ADPPA 规定，如人工智能所使用的数据集涵盖个人信息、数据与隐私，则构成“覆盖算法”；使用“覆盖算法”的大数据持有人，如果对个人或群体构成相应伤害风险，并单独或部分使用“覆盖算法”来收集、处理或传输覆盖数据，则应当根据 ADPPA 规定的评估标准进行隐私影响评估。另外，ADPPA 还对隐私政策的告知与退出机制、反偏见等内容做出了规定。ADPPA 规定，企业或代表企业的服务提供商需要告知个人有“选择退出”的权利，即拒绝企业对其个人数据的收集、处理或传输。

4.《关于通过联邦政府进一步促进种族平等和支持服务不足社区的行政命令》（Executive Order on Further Advancing Racial Equity and Support for Underserved Communities Through the Federal Government）

2023 年 2 月，拜登签署了《关于通过联邦政府进一步促进种族平等和支持服务不足社区的行政命令》，规定人工智能大模型应避免由于大量输入训练数据中存在的对种族、性别、年龄、文化和残疾等的偏见而导致训练结果输出内容中存在偏见。联邦政府在设计、开发、获取和使用人工智能和自动化系统时，各机构应在符合适用法律的前提下，防止、纠正歧视和促进公平，包括保护公众免受算法歧视。

（三）英国

2021 年 5 月，英国中央数字与数据办公室、人工智能办公室与内阁办公室联合发布了《自动决策系统的伦理、透明度与责任框架》（Ethics, Transparency and Accountability Framework for Automated Decision-Making，“ETAF”），对人工智能涉及的算法和自动化决策的伦理治理要求进行规定。ETAF 强调，算法和自动化决策在上线之前应该进行严格的、受控的和分阶段的测试。在整个原型和测试过程中，需要人类的专业知识和监督来确保技术上的弹性和安全，以及准确和可靠的系统。测试时，需要考虑自动化决策系统的准确性、安全性、可靠性、公平性和可解释性。ETAF 规定，企业必须对算法或自动决策系统做一个平等影响评估，使用高质量和多样化的数据集，发现和抵制所使用数据中明显的偏见和歧视。ETAF 指出，算法或计算机系统应该被设计为完全可以负责和可被审计的，算法和自动化的责任和问责制度应该明确。

二、我国与 AI 伦理治理相关的法律法规和行业规范概览

（一）法律法规

在我国，2017 年，国务院印发《新一代人工智能发展规划》并在该规划中提出了制定促进人工智能发展的法律法规和伦理规范的要求，之后《中华人民共和国科学技术进步法》《互联网信息服务算法推荐管理规定》《互联网信息服务深度合成管理规定》《生成式人工智能服务管理暂行办法》《网络安全标准实践指南—人工智能伦理安全风险防范指引》等一系列法律法规和相关规定相继对于 AI 伦理治理的要求予以规定。2022 年中共中央办公厅、国务院办公厅发布了《关于加强科技伦理治理的意见》，该意见是我国首个国家层面的、专门针对科技伦理治理的指导性文件，提出了科技伦理治理原则以及基本要求。2023 年 10 月发布的《科技伦理审查办法（试行）》对于科技伦理审查的基本程序、标准、条件等提出统一要求，标志着我国 AI 伦理治理监管体系建设进入了新阶段。我们在下表

中梳理了我国与 AI 伦理治理相关的法律法规及其主要内容：

序号	名称	位阶	生效日期	AI 伦理治理相关要求
1	《中华人民共和国科学技术进步法》	法律	2022.01.01	**要求科学技术研究开发机构、高等学校、企业事业单位等应当履行科技伦理管理主体责任按照国家有关规定建立健全科技伦理审查机制，**对科学技术活动开展科技伦理审查，禁止危害国家安全、损害社会公共利益、危害人体健康、违背科研诚信和科技伦理的科学技术研究开发和应用活动，否则相关单位、直接负责的主管人员和其他直接责任人员将受到行政处罚。
2	《新一代人工智能发展规划》	行政法规	2023.08.15	提出人工智能伦理规范和政策法规建设的战略目标，以及制定促进人工智能发展的法律法规和伦理规范的要求。
3	《互联网信息服务算法推荐管理规定》	部门规章	2022.03.01	**算法推荐服务提供者**应当落实算法安全主体责任，建立健全算法机制机理审核、科技伦理审查、用户注册、信息发布审核、数据安全和个人信息保护反电信网络诈骗、安全评估监测、安全事件应急处置等管理制度和技术措施，制定并公开算法推荐服务相关规则，配备与算法推荐服务规模相适应的专业人员和技术支撑；此外还应当定期审核、评估、验证算法机制机理、模型、数据和应用结果等，不得设置诱导用户沉迷、过度消费等违反法律法规或者违背伦理道德的算法模型。
4	《互联网信息服务深度合成管理规定》	部门规章	2023.01.10	**深度合成服务提供者**应当落实信息安全主体责任，建立健全用户注册、算法机制机理审核科技伦理审查、信息发布审核数据安全、个人信息保护、反电信网络诈骗、应急处置等管理制度，具有安全可控的技术保障措施。

序号	名称	位阶	生效日期	AI 伦理治理相关要求
5	《生成式人工智能服务管理暂行办法》	部门规章	2023.08.15	提供和使用生成式人工智能服务，应当遵守法律、行政法规尊重社会公德和伦理道德，遵守以下规定： （一）坚持社会主义核心价值观，不得生成煽动颠覆国家政权、推翻社会主义制度，危害国家安全和利益、损害国家形象，煽动分裂国家、破坏国家统一和社会稳定，宣扬恐怖主义、极端主义，宣扬民族仇恨、民族歧视，暴力、淫秽色情，以及虚假有害信息等法律、行政法规禁止的内容； （二）在算法设计、训练数据选择、模型生成和优化、提供服务等过程中，采取有效措施防止产生民族、信仰、国别、地域、性别、年龄、职业、健康等歧视； （三）尊重知识产权、商业道德，保守商业秘密，不得利用算法、数据、平台等优势，实施垄断和不正当竞争行为； （四）尊重他人合法权益，不得危害他人身心健康，不得侵害他人肖像权、名誉权、荣誉权、隐私权和个人信息权益； （五）基于服务类型特点，采取有效措施，提升生成式人工智能服务的透明度，提高生成内容的准确性和可靠性。
6	《科技伦理审查办法（试行）》	部门规章	2023.12.01	详见下文

序号	名称	位阶	生效日期	AI 伦理治理相关要求
7	《网络安全标准实践指南人工智能伦理安全风险防范指引》	其他规范性文件	2021.01.05	将 AI 伦理安全风险总结为以下五大方面： （1）失控性风险：如 AI 的行为与影响超出服务提供者预设、理解和可控的范围对社会价值等产生负面影响； （2）社会性风险：不合理使用 AI 而对社会价值等方面产生负面影响； （3）侵权性风险：AI 对人的基本权利，包括人身隐私、侵权性风险财产等造成侵害或产生负面影响； （4）歧视性风险：AI 对人类特定群体具有主观或客观偏见，影响公平公正、造成权利侵害或负面影响； （5）责任性风险：AI 相关各方行为失当、责任界定不清，对社会信任、社会价值等方面产生负面影响。
8	《关于加强科技伦理治理的意见》	其他规范性文件	2022.03.20	提出“科技伦理是开展科学研究、技术开发等科技活动需要遵循的价值理念和行为规范是促进科技事业健康发展的重要保障”，并明确了以下五大类科技伦理原则：增进人类福祉、尊重生命权利、坚持公平公正、合理控制风险和保持公开透明。

（二）行业规范

除前序法律法规和相关规定以外，在《新一代人工智能发展规划》等政策指引下，各机构、行业也积极响应，陆续发布了一系列 AI 伦理治理相关的行业规范，包括国家新一代人工智能治理专业委员会制定的《新一代人工智能治理原则——发展负责任的人工智能》《新一代人工智能伦理规范》、国家人工智能标准化总体组等制定的《人工智能伦理治理标准化指南》、同济大学上海市人工智能社会治理协同创新中心研究团队编制的《人工智能大模型伦理规范操作指引》、中国社会科学院国庆调研重大项目《我国人工智能伦理审查和监管制度建设状况调研》课题组编制的《人工智能法示范法 1.0（专家建议稿）》，提供了相关行业的 AI 伦理治理建议。下表梳理了我国与 AI 伦理治理相关的行业规范及其主要内容：

序号	行业规范	编制机构	发布时间	主要内容
1	《新一代人工智能治理原则——发展负责任的人工智能》	国家新一代人工智能治理专业委员会	2019.06	提出了人工智能治理的框架和行动指南，治理原则旨在更好协调人工智能发展与治理的关系，确保人工智能安全可控可靠，推动经济、社会及生态可持续发展，共建人类命运共同体。治理原则突出了发展负责任的人工智能这一主题，强调了和谐友好、公平公正、包容共享、尊重隐私、安全可控共担责任、开放协作、敏捷治理等八条原则。
2	《新一代人工智能伦理范》	国家新一代人工智能治理专业委员会	2021.09	旨在将伦理道德融入人工智能全生命周期，促进公平、公正和谐、安全，避免偏见、歧视《新隐私和信息泄露等问题。一代人工智能伦理规范》的适包研发、供应、使用等相关活动的自然人、法人和其他相关机构。在此基础上，《新一代人工智能伦理规范》明确了人工智能的基本伦理规范，包括增进人类福祉、促进公平公正保护隐私安全、确保可控可信.强化责任担当、提升伦理素养同时，《新一代人工智能伦理规范》提出了一系列人工智能应用管理规范、研发规范、供应规范和使用规范。
3	《人工智能伦理治理标准化指南》	国家人工智能标准化总体组	2023.03	共分为六章，以人工智能伦理治理标准体系的建立和具体标准研制为目标，重点围绕人工智能伦理概念和范畴、人工智能伦理准则、人工智能伦理风险分析、人工智能伦理治理的 T 知能价甲技术解决方安治理标准体系建设以及展望与建议等六个方面展开研究。
4	《人工智能大模型伦理规范操作指引》	同济大学等	2023.07	主要包括 AI 大模型全生命周期的技术与伦理要素、大模型的研发与应用的伦理原则、大模型技术研发的伦理实践指南三部分内容，提出了尊重人的自主权、保护个人隐私、保障公平公正、提高透明度和可解释性、负责任的创新等 5 项大模型伦理原则，以及公平性、透明性、隐私、安全性、责任、人类的监督与控制、可持续性等 7 项大模型伦理实践操作建议。

三、《科技伦理审查办法（试行）》要点

（一）适用范围

2023 年 10 月 8 日，科学技术部、教育部、工业和信息化部等多部门联合发布《科技伦理审查办法（试行）》（“《科技伦理审查办法》”），该办法于 2023 年 12 月 1 日起正式实施。该办法对于涉及以人为研究参与者的科技活动，包括利用人类生物样本、个人信息数据等的科技活动，或不直接涉及人或实验动物，但可能在生命健康、生态环境、公共秩序、可持续发展等方面带来伦理风险挑战的科技活动进行的科技伦理审查和监管做出了明确的规定，由此可见，该办法的适用范围几乎涵盖所有科技活动，包括人工智能相关的科技活动。

（二）审查主体

在审查主体方面，《科技伦理审查办法》明确要求从事生命科学、医学、人工智能等科技活动的单位，研究内容涉及科技伦理敏感领域的，应设立科技伦理（审查）委员会。其他有伦理审查需求的单位可根据实际情况设立科技伦理（审查）委员会。单位应在设立科技伦理（审查）委员会后 30 日内，通过国家科技伦理管理信息登记平台进行登记，登记内容包括科技伦理（审查）委员会组成、章程、工作制度等，相关内容发生变化时应及时更新，并在每年 3 月 31 日前，向国家科技伦理管理信息登记平台提交上一年度科技伦理（审查）委员会工作报告。科技伦理（审查）委员会的主要职责包括：

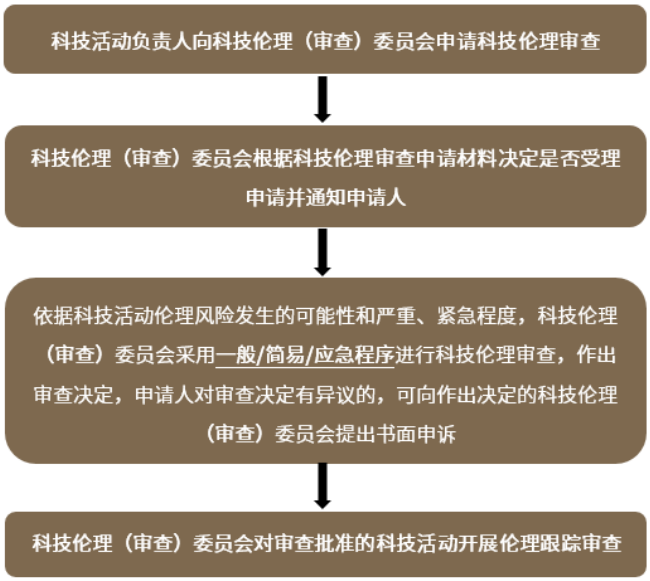
- 制定完善科技伦理（审查）委员会的管理制度和工作规范；
- 提供科技伦理咨询，指导科技人员对科技活动开展科技伦理风险评估；
- 开展科技伦理审查，按要求跟踪监督相关科技活动全过程；
- 对拟开展的科技活动是否属于《需要开展伦理审查复核的科技活动清单》范围作出判断；
- 组织开展对委员的科技伦理审查业务培训和科技人员的科技伦理知识培训；
- 受理并协助调查相关科技活动中涉及科技伦理问题的投诉举报；
- 按照主管部门要求进行登记、报告，配合地方、相关行业主管部门开展涉及科技伦理审查的相关工作。

在科技伦理（审查）委员会的人员组成方面，科技伦理（审查）委员会人数应不少于 7 人，设主任委员 1 人，副主任委员若干；委员任期不超过 5 年，可以连任。委员会成员应由具备相关科学技术背景的同行专家以及伦理、法律等相应专业背景的专家组

成，并应当有不同性别和非本单位的委员，民族自治地方应有熟悉当地情况的委员。由此可见，科技伦理（审查）委员会的设立具备一定门槛，根据《科技伦理审查办法》，如单位或个人涉及开展需进行科技伦理审查的科技活动，但单位未设立科技伦理（审查）委员会或个人无单位的，应书面委托其他单位的科技伦理（审查）委员会开展伦理审查工作。

（三）审查程序

根据《科技伦理审查办法》，科技伦理（审查）委员会开展科技伦理审查的流程如下：



科技伦理审查流程图示

根据《科技伦理审查办法》，针对纳入科技部发布的《需要开展伦理审查复核的科技活动清单》的科技活动，通过科技伦理（审查）委员会的科技审查后，除非国家实行行政审批等监管措施且将符合伦理要求作为审批条件、监管内容的，还需由科技活动承担单位（即开展科技活动的单位）报请所在地方或相关行业主管部门组织成立复核专家组开展专家复核，并由负责该等科技活动的科技审查的科技伦理（审查）委员会根据专家复核意见作出科技伦理审查决定；前序科技伦理（审查）委员会应在纳入清单管理的科技活动获得伦理审查批准后 30 日内，通过国家科技伦理管理信息登记平台进行登记，登记内容包括科技活动实施方案、伦理审查与复核情况等，相关内容发生变化时应及时更新，并在每年 3 月 31 日前向国家科技伦理管理信息登记平台提交上一年度纳入清单管理的科技活动实

施情况报告。根据科技部于 2023 年 10 月 8 日附随《科技伦理审查办法》发布的《需要开展伦理审查复核的科技活动清单》，“具有舆论社会动员能力和社会意识引导能力的算法模型、应用程序及系统的研发” “面向存在安全、人身健康风险等场景的具有高度自主能力的自动化决策系统的研发”均属于需要开展伦理审查复核的科技活动。因此，人工智能相关科技活动除通过科技伦理（审查）委员会的科技审查以外，极有可能还需在复核专家组进行伦理审查复核。

（四）审查内容

根据《科技伦理审查办法》，针对所有需要依法进行科技伦理审查的科技活动，科技伦理审查的具体内容包括：

- 拟开展的科技活动应符合增进人类福祉、尊重生命权利、坚持公平公正、合理控制风险、保持公开透明的科技伦理原则，参与科技活动的科技人员资质、研究基础及设施条件等符合相关要求；
- 拟开展的科技活动具有科学价值和社会价值，其研究目标的实现对增进人类福祉、实现社会可持续发展等具有积极作用。科技活动的风险受益合理，伦理风险控制方案及应急预案科学恰当、具有可操作性；
- 所制定的利益冲突申明和管理方案合理。

针对利用人类生物样本、个人信息数据等的科技活动在内的涉及以人作为研究参与者的科技活动，还需审查下述内容：

- 所制定的招募方案公平合理，生物样本的收集、储存、使用及处置合法合规，个人隐私数据、生物特征信息等信息处理符合个人信息保护的有关规定，对研究参与者的补偿、损伤治疗或赔偿等合法权益的保障方案合理，对脆弱人群给予特殊保护；
- 所提供的知情同意书内容完整、风险告知客观充分、表述清晰易懂，获取个人知情同意的方式和过程合规恰当。

针对涉及数据和算法的科技活动，还需审查下述内容：

- 数据的收集、存储、加工、使用等处理活动以及研究开发数据新技术等符合国家数据安全和个人信息保护等有关规定，数据安全风险监测及应急处理方案得当；
- 算法、模型和系统的设计、实现、应用等遵守公平、公正、透明、可靠、可控等原则，符合国家有关要求，伦理风险评估审核和应急处置方案合理，用户权益保护措施

全面得当。

四、国内大模型服务提供者的 AI 伦理治理责任

如本书《大模型合规之现实初探》一文中所述，向境内公众提供大模型服务的大模型服务提供者，包括平台运营方和技术支持方，均属于生成式人工智能服务提供者，从而均应当承担相应的 AI 伦理治理责任，具体而言：

- 在程序层面，在《科技伦理审查办法》正式生效以后，如大模型服务提供者涉及开展《科技伦理审查办法》适用范围内的科技活动，需自行设立科技伦理（审查）委员会或委托其他单位的科技伦理（审查）委员会对于所涉科技活动进行科技伦理审查，如涉及“具有舆论社会动员能力和社会意识引导能力的算法模型、应用程序及系统的研发”“面向存在安全、人身健康风险等场景的具有高度自主能力的自动化决策系统的研发”，还需报请所在地方或相关行业主管部门组织开展专家复核，亦即，科技伦理审查将成为算法备案以及安全评估以外大模型服务上线前的另一前置程序。
- 在实体层面，目前尚未出台法律法规对于不同类型的大模型服务提供者的 AI 伦理治理细则予以规定，尽管如此，一方面，如前文所述，目前《互联网信息服务算法推荐管理规定》《互联网信息服务深度合成管理规定》《生成式人工智能服务管理暂行办法》均要求大模型服务提供者履行伦理治理方面的合规义务；另一方面，与 AI 伦理治理相关的国家标准、团体标准以及行业标准均在制定过程中，其发布和施行指日可待。因此，大模型服务提供者可以参照前述《新一代人工智能伦理规范》等行业标准，按照技术主体（即从事人工智能相关的科学研究、技术开发、产品研制等研发活动的主体）、应用主体（即从事人工智能产品与服务相关的生产、运营、销售等人工供应活动和使用活动）进行自我定位，明确自身的 AI 伦理治理责任。

结语

人工智能技术的大幅进步提高了生产效率，为人类进行创造性智力活动提供了更加便捷的环境，同时，人工智能也对人类社会的伦理道德造成了挑战。《科技伦理审查办法》的颁布，标志着我国对人工智能技术的监管将逐步走向规范化，我们将持续在 AI 伦理审查和治理等诸多方面助力人工智能领域的企业向以人为本的人工智能时代迈进。

感谢实习生赵鹤翔对本文作出的贡献。

科技伦理（审查）委员会：如何设立？

张逸瑞 冯宝宝 张一凡 朱佳蔚

引言

如本书《人工智能（AI）：科技伦理治理走起》一文中所述，我国科学技术部会同教育部、工业和信息化部、国家卫生健康委等十部门联合印发的《科技伦理审查办法（试行）》（以下简称“《科技伦理审查办法》”）正式在法律法规层面明确了我国的科技伦理审查体系，该办法已于 2023 年 12 月 1 日开始施行。在人工智能领域，阿里巴巴早在 2022 年 9 月就已设立科技伦理治理委员会，其核心成员分别来自阿里研究院、达摩院、法务合规、阿里人工智能治理与可持续发展研究中心以及相关业务板块，同时，该委员会还引入了外部视角和监督，聘请七位来自科技、法律、公共管理、哲学等领域的专家组成独立的顾问委员会，为规则制定和伦理审查提供咨询建议[1]；而百度则于 2023 年 10 月正式成立科技伦理委员会，该委员会汇集了各领域专家力量，同时引入权威外部视角和监督，通过专项研讨和评估等方式，确保 AI 技术和应用符合法律标准和伦理规范[2]。算法模型的研发、应用企业均相继迈出了在实践层面落实科技伦理治理监管要求的脚步，从设立科技伦理（审查）委员会入手，逐步完善针对人工智能技术研发和应用的科技伦理治理体系。

以下，我们梳理了科技伦理（审查）委员会的设立和运行相关的问题，供行业内的企业参考：

一、什么单位需要设立科技伦理（审查）委员会？

根据《科技伦理审查办法》第四条，从事生命科学、医学、人工智能等科技活动的单位（包括高等学校、科研机构、医疗卫生机构、企业等），研究内容涉及科技伦理敏感领域的，应设立科技伦理（审查）委员会。其他有科技伦理审查需求的单位可根据实际情况设立科技伦理（审查）委员会。

根据上述规定，对于高等学校、科研机构、医疗卫生机构、企业等单位，同时满足下述两项时，需设立科技伦理（审查）委员会：（1）从事生命科学、医学、人工智能等科

[1] 详见下述链接：http://www.sh.chinanews.com.cn/kjjy/2022-09-02/102917.shtml，最后访问日期：2024 年 3 月 19 日。
[2] 详见下述链接：https://company.cnstock.com/company/scp_gsxw/202311/5154028.htm，最后访问日期：2024 年 3 月 19 日。

技活动；（2）研究内容涉及科技伦理敏感领域。然而，关于何为“科技伦理敏感领域”，我国目前的法律法规并未作出明确规定。

值得注意的是，根据《科技伦理审查办法》第五十三条，地方、相关行业主管部门可按照本办法规定，结合实际情况制定或修订本地方、本系统的科技伦理审查办法、细则等制度规范；科技类社会团体可制定本领域的科技伦理审查具体规范和指南。目前，我国部分地区的政策文件在《科技伦理审查办法》的基础上，进一步明确了应履行科技伦理（审查）委员会设立义务的单位范围，例如：根据《海南省科技伦理治理实施方案》[3]，“从事生命科学、医学、人工智能等重点领域科技活动或研究内容涉及科技伦理敏感领域的单位，应根据需要设立综合性或单一领域方向的科技伦理（审查）委员会”；根据广东省人民政府发布的《关于加强科技伦理治理的实施方案》[4]，“从事生命科学、医学、人工智能等重点领域科技活动单位，应设立科技伦理（审查）委员会”。

基于前序监管现状，至少对于生命科学、医学、人工智能的企业，可以考虑将组建科技伦理（审查）委员会的工作提上日程。

二、是否可以委托其他单位的科技伦理（审查）委员会开展伦理审查？

根据《科技伦理审查办法》第十三条，如满足下述任何一种情形，单位可以书面委托其他满足要求的科技伦理（审查）委员会开展伦理审查：（1）单位科技伦理（审查）委员会无法胜任审查工作要求；（2）单位未设立科技伦理（审查）委员会；（3）无单位人员开展科技活动的。

尽管《科技伦理审查办法》确立了委托审查机制，关于该等机制的适用范围以及具体的操作方式（例如单位资助非本单位的研究人员开展科技活动或以其他间接方式参与科技活动是否属于“无单位人员开展科技活动”；受委托开展科技伦理审查的科技伦理（审查）委员会是否需满足特殊要求；某一设立科技伦理（审查）委员会的单位是否可以接受多个关联公司的委托开展科技伦理审查），仍有待主管部门出台相关规定，我们将持续关注。

三、如何设立科技伦理（审查）委员会？

（一）设立条件

根据《科技伦理审查办法》第七条、第八条，组建科技伦理（审查）委员会至少需要符合下述条件：

- 在专业背景方面，委员会委员应当包含具备相关科学技术背景的同行专家，以及伦理、法律等相应专业背景的专家组成，应具备相应的科技伦理审查能力和水平，

[3] 2023.01.06 发布并实施。
[4] 2023.05.09 发布并实施。

科研诚信状况良好；

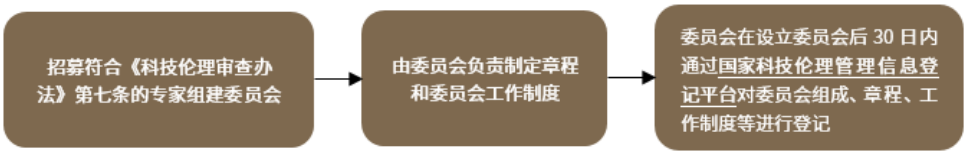
- 在人数方面，委员会委员的人数应不少于 7 人，其中应当包括主任委员 1 人，副主任委员若干；
- 在其他方面，委员会应当有不同性别和非本单位的委员，民族自治地方应有熟悉当地情况的委员。

除满足前述人员组成要求以外，为确保科技伦理（审查）委员会独立开展伦理审查工作，单位还应当为科技伦理（审查）委员会提供相应的工作条件，包括为委员会配备专（兼）职工作人员、办公场所、专门经费等等。

值得注意的是，为了促进科技伦理（审查）委员会建设标准的规范化和统一化，根据《科技伦理审查办法》第四十一条，国家将推动建立科技伦理（审查）委员会认证机制。即，未来，科技部可能会就科技伦理（审查）委员会认证机制发布详细的政策和指南，其中可能会进一步细化关于科技伦理（审查）委员会的设立要求。

（二）设立流程

根据《科技伦理审查办法》，我们将科技伦理（审查）委员会的设立流程归纳如下：



值得注意的是，根据《科技伦理审查办法》第五十四条，相关行业主管部门对本领域科技伦理（审查）委员会设立或科技伦理审查有特殊规定且符合本办法精神的，从其规定；本办法未作规定的，按照其他现有相关规定执行。此前，国家卫健委联合教育部、科技部和国家中医药局曾发布《涉及人的生命科学和医学研究伦理审查办法》[5] 对于涉及人的生命科学、医学的科技研究[6] 的伦理审查要求进行规定。结合前序《科技伦理审查办法》的规定，我们理解，对于涉及人的生命科学、医学的科技研究，《科技伦理审查办法》和《涉及人的生命科学和医学研究伦理审查办法》将同时适用，且后者可能将作为行业的特殊规定优先适用。就伦理审查委员会的设立流程而言，在《科技伦理审查办法》的基础之上，

---

[5] 2023.02.18 发布并实施。

[6] 根据《涉及人的生命科学和医学研究伦理审查办法》第三条，本办法所称涉及人的生命科学和医学研究是指以人为受试者或者使用人（统称研究参与者）的生物样本、信息数据（包括健康记录、行为等）开展的以下研究活动：（一）采用物理学、化学、生物学、中医药学等方法对人的生殖、生长、发育、衰老等进行研究的活动；（二）采用物理学、化学、生物学、中医药学、心理学等方法对人的生理、心理行为、病理现象、疾病病因和发病机制，以及疾病的预防、诊断、治疗和康复等进行研究的活动；（三）采用新技术或者新产品在人体上进行试验研究的活动；（四）采用流行病学、社会学、心理学等方法收集、记录、使用、报告或者储存有关人的涉及生命科学和医学问题的生物样本、信息数据（包括健康记录、行为等）等科学研究资料的活动。

《涉及人的生命科学和医学研究伦理审查办法》还要求从事涉及人的生命科学、医学的科技研究的机构在伦理审查委员会设立之日起 3 个月内向本机构的执业登记机关 / 上级主管部门进行备案，并在国家医学研究登记备案信息系统上传信息。因此，对于涉及人的生命科学、医学的科技研究的人工智能技术研发和应用单位（例如 AI 医学影像、AI 医疗机器人、临床决策支持系统（CDSS）的研发和应用单位）而言，在完成《科技伦理审查办法》项下的科技伦理（审查）委员会的设立流程以外，还应按照《涉及人的生命科学和医学研究伦理审查办法》的要求，进行相应的登记、备案。

四、科技伦理（审查）委员会有哪些职责?

根据《科技伦理审查办法》第五条，科技伦理（审查）委员会应当履行下述职责：

（一）制定完善科技伦理（审查）委员会的管理制度和工作规范

《科技伦理审查办法》第六条对于科技伦理（审查）委员会应当制定的管理制度和工作规范予以列举，其中包括：章程；审查、监督、保密管理、档案管理等制度规范；工作规程；利益冲突管理机制。

结合《科技伦理审查办法》第八条，我们理解前述管理制度和工作规范应至少涵盖针对委员会委员的下述要求：

- 遵纪守法：遵守我国宪法、法律、法规和科技伦理有关制度规范及所在科技伦理（审查）委员会的章程制度；
- 履行审查职责：按时参加科技伦理审查会议，独立公正发表审查意见；
- 保密：严格遵守保密规定，对科技伦理审查工作中接触、知悉的国家秘密、个人隐私、个人信息、技术秘密、未公开信息等，未经允许不得泄露或用于其他目的；
- 避免利益冲突：遵守利益冲突管理要求，并按规定回避；
- 参加培训：定期参加科技伦理审查业务培训；
- 其他：完成委员会安排的其他工作。

（二）提供科技伦理咨询，指导科技人员对科技活动开展科技伦理风险评估

在某些场景下，科技人员可能难以评估特定的科技活动的风险，从而无法判断是否需要针对该等科技活动进行科技伦理审查。此时，科技伦理（审查）委员会应当提供相关的咨询，指导科技人员对科技活动开展科技伦理风险评估。我们理解，该项职责是对常态科技活动中的复杂伦理问题提供咨询意见，相较于科技伦理审查职责而言，该项职责更类似一种铺垫性的、导入性的工作。为了减少重复性工作，科技伦理（审查）委员会可以汇总

在科技伦理咨询过程中科技人员关注的共性问题，并形成伦理审查申请指南、伦理审查指南等指南文件供科技人员参考。

（三）开展科技伦理审查，按要求跟踪监督相关科技活动全过程

《科技伦理审查办法》对于科技伦理（审查）委员会进行科技伦理审查的审查范围、审查程序以及审查内容等均进行了较为明确的规定[7]。值得注意的是，除在科技活动开展前对于科技活动进行事前的科技伦理审查以外，对于审查批准的科技活动，科技伦理（审查）委员会还应进行跟踪审查，跟踪审查间隔一般不超过 12 个月；科技伦理（审查）委员会有权根据跟踪审查需要，要求科技活动负责人提交相关材料，并在必要时作出暂停或终止科技活动等决定；跟踪审查的主要审查内容如下：

- 科技活动实施方案执行情况及调整情况；
- 科技伦理风险防控措施执行情况；
- 科技伦理风险的潜在变化及可能影响研究参与者权益和安全等情况；
- 其他需要跟踪审查的内容。

（四）对拟开展的科技活动是否属于《需要开展伦理审查复核的科技活动清单》范围作出判断

如我们在本书《人工智能（AI）：科技伦理治理走起》一文中所述，人工智能相关科技活动除需科技伦理（审查）委员会进行伦理审查以外，由于还可能落入《需要开展伦理审查复核的科技活动清单》范围，极有可能还需由地方或相关行业主管部门组织成立的复核专家组进行伦理审查复核；而科技伦理（审查）委员会应负责判断人工智能相关的科技活动是否纳入前述清单管理、是否需要由复核专家组进行伦理审查复核。

针对“具有舆论社会动员能力和社会意识引导能力的算法模型、应用程序及系统的研发”“面向存在安全、人身健康风险等场景的具有高度自主能力的自动化决策系统的研发”等纳入清单管理的科技活动类型，目前我国法律法规并未明确相应的判断标准，因此，现阶段，我们理解科技伦理（审查）委员会可以参考《信息技术 人工智能 风险管理能力评估》（T/CESA 1193—2022）等人工智能领域社会团体制定的科技伦理风险评估相关规范和指南以及人工智能伦理相关国际标准（详情可见国家人工智能标准化总体组等编制的《人工智能伦理治理标准化指南》[8] 所列出的人工智能伦理相关国际标准清单）判断人工智能领域科技活动是否纳入前述清单管理。

---

[7] 参见本书中《人工智能（AI）：科技伦理治理走起》一文。

[8] 链接如下：https://www.aipubservice.com/airesource/fs/%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD%E4%BC%A6%E7%90%86%E6%B2%BB%E7%90%86%E6%A0%87%E5%87%86%E5%8C%96%E6%8C%87%E5%8D%97.pdf，最后访问日期：2024 年 3 月 19 日。

（五）组织开展对委员的科技伦理审查业务培训和科技人员的科技伦理知识培训

一方面，科技伦理（审查）委员会应当围绕科技伦理审查业务、科技伦理审查知识组织开展对于委员的定期培训，确保委员能够准确把握科技伦理工作正确方向。另一方面，科技伦理（审查）委员会还应当对参与科技活动的科技人员开展科技伦理知识培训，培训的内容可以包括相关法律法规、部门规章、科技伦理审查相关知识等等。

（六）受理并协助调查相关科技活动中涉及科技伦理问题的投诉举报

根据《科技伦理审查办法》第四十六条，对科技活动中违反科技伦理规范、违背科技伦理要求的行为，任何单位或个人有权依法向科技活动承担单位或地方、相关行业主管部门投诉举报。若举报者向科技活动承担单位本身提起投诉举报，科技伦理（审查）委员会应当受理；若举报者地方、相关行业主管部门投诉举报，科技伦理（审查）委员会则应当协助调查。

（七）按照主管部门要求进行登记、报告，配合地方、相关行业主管部门开展涉及科技伦理审查的相关工作

我们将《科技伦理审查办法》规定的科技伦理（审查）委员会的登记和报告义务梳理如下：

序号	登记内容	时间要求
1	科技伦理（审查）委员会组成、章程、工作制度等	设立科技伦理（审查）委员会后 30 日内
2	纳入清单管理的科技活动实施方案、伦理审查与复核情况等	获得伦理审查批准后 30 日内
3	科技伦理（审查）委员会工作报告、纳入清单管理的科技活动实施情况报告	每年 3 月 31 日前

目前，各单位的科技伦理（审查）委员会可通过国家科技管理信息系统公共服务平台[9]进行前述登记和报告。

五、科技伦理（审查）委员会将承担哪些法律责任？

在法律责任方面，《科技伦理审查办法》第四十八条对于科技伦理（审查）委员会、委员的法律责任进行了原则性规定：科技伦理（审查）委员会、委员有下列行为的，由有管辖权的机构依据法律、行政法规和相关规定给予处罚或者处理，如造成财产损失或者其他损害的、构成犯罪的，还应依法承担相应民事、刑事责任：（1）弄虚作假，为科技活

[9] 平台链接如下：https://service.most.gov.cn/kjll/，最后访问日期：2024 年 3 月 19 日；相应操作说明链接如下：https://service.most.gov.cn/kjll/20240109/5433.html，最后访问日期：2024 年 3 月 19 日。

动承担单位获得科技伦理审查批准提供便利；（2）徇私舞弊、滥用职权或者玩忽职守的；（3）其他违反本办法规定的行为。

科技部此前发布的《科学技术活动违规行为处理暂行规定》[10] 对于科学技术活动咨询评审专家开展有关科学技术活动过程中出现的违规行为的处理进行规定。其中，“科学技术活动评审专家”是指为科学技术活动提供咨询、评审、评估、评价等意见的专业人员，我们理解科技伦理（审查）委员会委员落入前述“科学技术活动评审专家”的范畴。结合前述《科技伦理审查办法》关于“由有管辖权的机构依据法律、行政法规和相关规定给予处罚或者处理”的规定，我们理解在科技伦理（审查）委员会委员存在《科学技术活动违规行为处理暂行规定》规定的下述违规行为的情况下，可能受到当地科学技术行政部门（即科技厅）采取的警告、责令限期整改、约谈、一定范围内或公开通报批评、记入科研诚信严重失信行为数据库等处理措施：

- 采取弄虚作假等不正当手段获取咨询、评审、评估、评价、监督检查资格；
- 违反回避制度要求；
- 接受“打招呼”“走关系”等请托；
- 引导、游说其他专家或工作人员，影响咨询、评审、评估、评价、监督检查过程和结果；
- 索取、收受利益相关方财物或其他不正当利益；
- 出具明显不当的咨询、评审、评估、评价、监督检查意见；
- 泄漏咨询评审过程中需保密的申请人、专家名单、专家意见、评审结论等相关信息；
- 抄袭、剽窃咨询评审对象的科学技术成果；
- 违反国家科学技术活动保密相关规定。

六、针对科技伦理（审查）委员会需要准备哪些制度文件？

结合《科技伦理审查办法》的规定， 可以考虑从章程和工作制度两个模块构建科技伦理（审查）委员会的制度体系，具体而言：

（一）章程

章程是科技（审查）伦理委员会设立时最重要的文件之一，是规范委员会内部组织和活动的基本规则。一般而言，科技（审查）伦理委员会章程应当包括如下内容：（1）总则；（2）组织结构；（3）职责权限；（4）审查程序。在条款设计方面，可以参考一般公司

[10] 2020.07.17 发布，2020.09.01 实施。

章程的体例，结合科技伦理审查的特点（例如科技伦理审查依据的基本原则、根据科技伦理风险等级针对性适用不同的科技伦理审查程序等等），起草章程中的具体内容。

（二）科技（审查）伦理委员会工作制度

科技（审查）伦理委员会工作制度一般用于指导和约束委员会委员的具体行为，科技（审查）伦理委员会可以根据实际管理和工作需要，对于章程中的原则性要求进行细化。对于科技（审查）伦理委员会而言，基于《科技伦理审查办法》的要求，应当尤其关注保密以及利益冲突方面的制度建设，可考虑制定的工作制度包括：

- 保密和隐私保护制度；
- 利益冲突管理制度；
- 组织管理制度；
- 独立顾问聘任制度；
- 财务管理制度人员培训制度；
- 文档管理制度。

结语

《科技伦理审查办法》为科技伦理（审查）委员会的组建和运行提供了较为清晰完整的框架和指导。然而，我们也注意到，对于许多人工智能领域的企业而言，科技伦理（审查）委员会的设立并非易事；同时，科技伦理风险评估标准、人工智能伦理规范等科技伦理治理所依托的实体性规范仍有待有关行业的主管部门、社会团体进行研究和制定。在未来，我们将持续关注科技伦理审查相关的法律法规、行业标准和实践，我们也期待以担任委员 / 独立顾问、提供法律咨询、与监管部门对接等多种方式参与企业科技伦理（审查）委员会的组建和运行，为企业完善科技伦理治理体系保驾护航。

感谢实习生于子晗对本文作出的贡献。

AI 安全与合规：维护国家安全的新疆域

张逸瑞 景云峰

引言

2023 年 10 月 18 日，国家互联网信息办公室（“网信办”）发布《全球人工智能治理倡议》，重申各国应在人工智能治理中加强信息交流和技术合作，共同做好风险防范，形成具有广泛共识的人工智能治理框架和标准规范，不断提升人工智能技术的安全性、可靠性、可控性、公平性。2023 年 11 月 1 日，中国、美国、欧盟、英国在内的二十余个主要国家和地区在英国主办的首届人工智能安全全球峰会上共同签署了《布莱切利宣言》（The Bletchley Declaration），承诺以安全可靠、以人为本、可信赖及负责的方式设计、开发、部署并使用 AI。2023 年 11 月 8 日，习近平主席在 2023 年世界互联网大会乌镇峰会开幕式发表的视频致辞中指出，愿同各方携手落实《全球人工智能治理倡议》，促进人工智能安全发展。由此可见，人工智能已成为维护国家安全的新疆域，人工智能安全是我国政府一贯密切关注的重要议题。人工智能作为全球战略性与变革性信息技术，在对经济社会发展和人类文明进步产生深远影响的同时，也在网络安全、数据安全、算法安全、信息安全等领域引发新型国家安全风险。为防范上述风险，各国必将逐步健全和完善人工智能治理相关的法律法规。我们在此前与上海人工智能研究院、华为技术有限公司、上海昇思 AI 框架 & 大模型创新中心共同编制的《大模型合规白皮书》[1] 中，对于美国、欧盟以及我国在大模型及人工智能方面的法律监管现状进行了梳理，本文将重点围绕人工智能安全这一主题，为企业等制定和实施人工智能安全治理措施提供参考。

一、AI 安全与合规要点概览

前述《全球人工智能治理倡议》《布莱切利宣言》以及美国、欧盟、英国等主要国家和地区人工智能治理相关监管规则中均提出有关人工智能安全的要求。其中，我国《全球人工智能治理倡议》以及美国总统拜登于 2023 年 10 月签署的《关于安全、可靠和值得信赖的人工智能的行政命令》特别提出了针对国家安全风险的防范措施：

---

[1] 详见下述链接：https://mp.weixin.qq.com/s?__biz=MzA4NDMzNjMyNQ==&mid=2653343872&idx=2&sn=8cedd21fddf2eb991c6e44395acd6891&chksm=843abb2ab34d323cd6ab1b4fc17135956955f8d7037aad3a9c0c99f7e88d29d0a47dc858bacc&scene=21#wechat_redirect，最后访问日期：2024 年 3 月 19 日。

- 根据《全球人工智能治理倡议》，面向他国提供人工智能产品和服务时，应尊重他国主权，严格遵守他国法律，接受他国法律管辖；反对利用人工智能技术优势操纵舆论、传播虚假信息，干涉他国内政、社会制度及社会秩序，危害他国主权；此外，发展人工智能应坚持“智能向善”的宗旨，遵守适用的国际法，符合和平、发展、公平、正义、民主、自由的全人类共同价值，共同防范和打击恐怖主义、极端势力和跨国有组织犯罪集团对人工智能技术的恶用滥用，各国尤其是大国对在军事领域研发和使用人工智能技术应该采取慎重负责的态度。
- 根据《关于安全、可靠和值得信赖的人工智能的行政命令》（Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence），开发任何对国家安全、国家经济安全或国家公共健康和安全构成严重风险的基础模型的公司在训练模型时必须通知联邦政府，并且必须共享所有红队安全测试的结果；美国国家标准与技术研究院、国土安全部、能源部将负责一系列标准、工具和测试，帮助确保人工智能系统安全、可靠和可信；国家安全委员会和白宫办公厅主任将制定一份国家安全备忘录，指导有关人工智能和安全的进一步行动，以确保美国军方和情报界在执行任务时安全、合乎道德和有效地使用人工智能，并将指导采取行动，打击对手在军事上使用人工智能。

我们理解人工智能安全涉及国家安全的多个领域，如军事安全、政治安全、社会安全、经济安全、信息安全、算法安全、数据安全、网络安全等，具体而言：

（一）军事安全与政治安全

AI 在军事作战、社会舆情等领域的应用可能会影响一国的军事安全和政治安全。一方面，AI 可能通过基于收集用户画像进行深度引导并传播政治主张，对不同的政治信仰、国家、种族、团体进行有失公平的区别对待，在具有复杂历史背景的问题上与开发者立场保持一致等途径影响公众政治意识形态，间接威胁国家的军事与政治安全。另一方面，AI 可用于构建新型军事打击力量，例如提供最新的战场态势和战术建议、智能分析情报、协助武器智能化，从而提升战斗能力，直接威胁国家安全。

（二）社会安全

AI 的运用与发展对社会安全有着直观的影响。AI 的发展滋生了多种新型违法犯罪，例如利用生成式 AI 定制个性化诈骗话术行使诈骗敲诈、利用 AI 合成虚假的私人视频或色情视频进行侮辱诽谤、利用 AI“刷单”“造粉”制造冲突、操控舆论以制造或推动网络暴力等。此外，高度自治的 AI 系统，如无人驾驶汽车、医疗机器人等，一旦出现数据泄露、

网络连通性差等问题可能会直接危害人类身体健康甚至生命安全。同时，受到训练数据或决策算法有偏见、AI 产品缺乏道德约束、与 AI 相关的安全事件追责难等因素的影响，AI 产品可能会对社会现有伦理道德体系造成强烈冲击。

我国《全球人工智能治理倡议》、美国《关于安全、可靠和值得信赖的人工智能的行政命令》以及欧盟《人工智能法案》（Artificial Intelligence Act，“AI 法案”）均强调了 AI 在社会安全方面引发的风险：

- 《全球人工智能治理倡议》提出，坚持公平性和非歧视性原则，避免在数据获取、算法设计、技术开发、产品研发与应用过程中，产生针对不同或特定民族、信仰、国别、性别等偏见和歧视。同时，坚持伦理先行，建立并完善人工智能伦理准则、规范及问责机制，形成人工智能伦理指南，建立科技伦理审查和监管制度，明确人工智能相关主体的责任和权力边界，充分尊重并保障各群体合法权益，及时回应国内和国际相关伦理关切。
- 《关于安全、可靠和值得信赖的人工智能的行政命令》提出，美国将基于《人工智能权利法案蓝图》以及以一系列行政命令，指示各机构打击算法歧视，同时执行现有授权以保护人们的权利和安全，确保人工智能促进公平和公民权利；卫生与公众服务部还将制定一项安全计划，接收涉及人工智能的伤害或不安全医疗行为的报告，并采取行动予以补救。
- AI 法案提出，采用潜意识技术或有目的的操纵或欺骗技术、利用个人或社会群体的弱点（例如已知的人格特征或社会经济状况、年龄、身体精神能力）、利用人的社会行为或人格特征进行社会评分的 AI 系统均属于存在不可接受风险的 AI 系统，严厉禁止使用。针对高风险 AI 系统，AI 系统部署方必须进行基本权利影响评估，并向国家当局通报评估结果；此外，高风险 AI 系统必须在技术上具有稳健性，以确保技术适用于其目的，并且虚假的正面 / 负面结果不会对受保护群体（例如种族或民族起源、性别、年龄等）产生不成比例的影响，同时需要使用足够具有代表性的数据集进行训练和测试，以最小化模型中潜在的不公平偏见风险，并确保可以通过适当的偏见检测、纠正和其他缓解措施来解决这些问题。

（三）经济安全

AI 可能会给国家经济安全带来诸多风险，例如结构性失业、行业寡头垄断等等。在结构性失业方面，在 AI 与传统行业融合的过程中，AI 从替代人类的手足和体力发展到替代人类的大脑，使得重复体力劳动者、简单脑力从业者、咨询分析等知识型行业甚至艺术

行业等都可能面临下岗威胁，可能会导致结构性失业。在行业寡头垄断方面，AI 产品对算法、算力和数据的高度依赖，可能导致科技企业的寡头垄断，例如，目前生成式 AI 平台和应用系统大部分由国外互联网巨头及其参股或控股公司研发而成，这些企业可以根据自身积累的数据优势、建立“数据壁垒”，可能导致 AI 初创企业入局难。

我国《全球人工智能治理倡议》、美国《关于安全、可靠和值得信赖的人工智能的行政命令》也关注到了 AI 对经济安全造成的风险：

- 《全球人工智能治理倡议》提出，发展人工智能应坚持相互尊重、平等互利的原则，各国无论大小、强弱，无论社会制度如何，都有平等发展和利用人工智能的权利。鼓励全球共同推动人工智能健康发展，共享人工智能知识成果，开源人工智能技术。反对以意识形态划线或构建排他性集团，恶意阻挠他国人工智能发展。反对利用技术垄断和单边强制措施制造发展壁垒，恶意阻断全球人工智能供应链。
- 《关于安全、可靠和值得信赖的人工智能的行政命令》提出，为了降低人工智能带来的工作场所监控增加、偏见和工作流失等风险，支持工人集体谈判的能力，并投资于所有人都能获得的劳动力培训和发展，美国将制定原则和最佳实践，通过解决工作转移、劳动标准、工作场所公平、健康和安全以及数据收集等问题，减少人工智能对工人的伤害，最大限度地提高人工智能对工人的益处，这些原则和最佳实践也将为工人提供指导，防止雇主对工人的补偿不足、对求职申请的评估不公或影响工人的组织能力，从而使工人受益；此外，美国还将促进公平、开放和有竞争力的人工智能生态系统，为小型开发者和企业家提供获得技术援助和资源的机会，帮助小型企业将人工智能突破商业化。

（四）信息安全

AI 信息安全主要包括 AI 技术应用于信息传播以及 AI 产品输出的信息内容安全问题。在信息传播方面，以融合了 AI 技术的智能推荐为例，智能推荐能够根据用户的浏览记录、交易信息等数据对用户兴趣爱好、行为习惯进行分析与预测，并根据用户偏好推荐信息内容，不法分子可能会借助智能推荐将虚假信息、涉黄涉恐、违规言论、钓鱼邮寄等不良信息内容精准地投放给易攻击目标人群，增加了不良信息传播的针对性、有效性和隐蔽性。在输出内容方面，AI 技术可能被用来制作虚假信息内容，用以实施诈骗等不法活动，例如通过 AI 合成能够以假乱真的声音、图像，基于二维图片合成三维模型并根据声音片段修改视频内人物表情和嘴部动作，从而生成口型一致的视频合成内容实施诈骗。

《布莱切利宣言》、美国《关于安全、可靠和值得信赖的人工智能的行政命令》以及

欧盟均关注到了利用 AI 技术生成欺骗性内容、虚假信息的风险：

- 《布莱切利宣言》提出，人工智能系统操纵内容或生成欺骗性内容的能力可能会带来不可预见的风险，前沿人工智能系统可能放大虚假信息等风险问题；
- 《关于安全、可靠和值得信赖的人工智能的行政命令》提出，美国将通过建立检测人工智能生成内容和认证官方内容的标准和最佳实践，保护美国人免受人工智能带来的欺诈和欺骗；
- 为打击虚假信息，2023 年 6 月，欧盟委员会副主席乔罗娃（Vera Jourova）向谷歌、抖音国际版、微软、Facebook 和 Instagram 母公司 Meta 等超过 40 家科技企业要求，检测人工智能（AI）生成的图片、视频和文本，向用户提供明确的标记[2]；同时，《AI 法案》中亦要求生成深度合成内容的 AI 系统使用者需要对外告知该等内容是由 AI 生成或操纵的，而并非真实内容。

（五）算法安全

AI 算法安全涉及算法设计或实施与预期不符、算法潜藏偏见与歧视、算法黑箱可解释性差等问题。算法设计与实施中可能存在的错误设计目标函数、计算成本过高的目标函数导致无法实际运行或者选用算法模型表达能力与实际情况不匹配等问题会导致算法无法实现设计者的预设目标，最终导致决策偏离甚至出现伤害性结果。由于算法的设计目的、模型选择、数据使用等是设计者和开发者的主观选择，且训练数据作为社会现实的反映本身具有歧视性，训练得出的算法模型也会天然潜藏歧视和偏见。尽管 AI 本身擅长决策，但由于公司或个人主张商业秘密或私人财产、公众无法理解决策算法源代码、决策算法复杂度高等原因导致的 AI 算法黑箱或不透明，使得相关监督与审查陷入困境。

为确保算法的安全性、解决算法黑箱问题，欧盟《AI 法案》要求针对高风险的 AI 系统在系统全生命周期落实透明度要求，并且要求 AI 系统提供方向下游 AI 系统部署方、分销商等披露特定信息。

（六）数据安全

AI 带来的数据安全问题包括数据泄露、数据跨境传输等，其中数据泄露问题包括内部泄露和其挖掘分析能力带来的个人信息或隐私暴露。一方面，AI 在与用户交互过程中可能收集许多私密或敏感数据，这些数据也会被 AI 公司用于进一步训练模型，但在训练后的模型中很难删除相关数据。如果这些数据没有得到充分的保护，就可能被不法分子获取，导致个人信息或隐私数据泄露的风险增加。另一方面，AI 可基于其采集到无数个看

[2] https://www.stcn.com/article/detail/884259.html，最后访问日期：2024 年 3 月 19 日。

似不相关的数据片段，通过深度挖掘分析得到更多与用户个人信息或隐私相关的信息，导致现行的数据匿名化等安全保护措施无效。同时，根据部分 AI 产品的运作原理，用户在对话框进行交互时，相关问答数据可能会被传输到位于境外的产品开发公司，其间数据的跨境流动可能会引发数据跨境安全问题。

我国、美国以及欧盟等国家和地区均将数据安全视为 AI 安全方面最为重要的问题之一：

- 我国《全球人工智能治理倡议》提出逐步建立健全法律和规章制度，保障人工智能研发和应用中的个人隐私与数据安全，反对窃取、篡改、泄露和其他非法收集利用个人信息的行为；
- 美国《关于安全、可靠和值得信赖的人工智能的行政命令》呼吁国会通过两党数据隐私立法以保护所有美国人的隐私，要求联邦优先支持加快开发和使用保护隐私的技术，提出将制定评估隐私保护技术有效性的指导方针；
- 欧盟《AI 法案》要求 AI 系统的提供方以及部署方均按照《通用数据保护条例》的要求履行数据保护的义务，其中，AI 系统的提供方需对 AI 系统进行数据保护影响评估并发布摘要，以及提供输入数据或所使用的数据集的任何其他相关信息的说明。

（七）网络安全

AI 学习框架和组件存在的安全漏洞可能会引发系统安全问题。目前，国内 AI 产品和应用的研发往往基于国内外科技巨头发布的 AI 学习框架和组件展开，该等框架和组件很有可能未经严格测试管理和安全认证从而可能存在漏洞和后门等安全风险，一旦被攻击者恶意利用，就有可能会危及 AI 产品和应用的完整性和可用性。同时，AI 技术也可能被用来提升网络攻击效率与破坏能力。AI 产品自动生成代码的能力使得黑客更为便捷地设计、编写和执行恶意代码与指令以生成网络攻击工具，且攻击方将利用 AI 更快、更准地发现漏洞，发起更隐秘的攻击。

为应对 AI 模型被广泛使用而带来的网络风险，欧盟《AI 法案》要求具有系统性风险的 AI 模型（即使用总计计算能力超过 10^25 FLOPs 训练的通用人工智能模型）提供方履行进行风险评估、减轻风险、报告严重事件、进行尖端测试和模型评估、确保网络安全等一系列合规义务。此外，根据欧盟的现行要求，前述具有系统性风险的 AI 模型的标准可能会根据技术进展进行更新，并在特定情况下基于其他标准（例如用户数量或模型的自治

程度）判定 AI 模型为具有系统性风险的 AI 模型。

二、当前中国新国家安全格局下的 AI 安全实践

结合上述主要国家和地区在 AI 安全方面的关注重点以及 AI 安全在中国的实践，我们理解企业构建 AI 安全体系时，至少应该将下述要点纳入考量：

（一）国家安全角度的整体要求

根据《中华人民共和国国家安全法》[3] 对于国家安全的定义，国家安全是指国家政权、主权、统一和领土完整、人民福祉、经济社会可持续发展和国家其他重大利益相对处于没有危险和不受内外威胁的状态，以及保障持续安全状态的能力。这就意味着，国家安全的目标是追求一个相对稳定的低风险状态，而国家安全法的立法目的是实现针对国家安全的冲击和威胁的治理。

如前文所述，作为具有颠覆性意义的通用目的技术（General Purpose Technology），AI 技术的发展将对国家社会经济生活产生重大影响，因此，应当从国家安全角度对 AI 技术进行重点规制。结合前述 AI 安全的要点，以及《中华人民共和国国家安全法》等法律法规，我们理解，AI 产品的全生命周期都应当遵守《中华人民共和国国家安全法》在政治安全、军事安全、经济安全、文化安全、社会安全、科技安全、网络安全等国家安全领域的原则性要求。倘若 AI 产品涉及《中华人民共和国国家安全法》第五十九条规定影响或者可能影响国家安全的外商投资、特定物项和关键技术、网络信息技术产品和服务、涉及国家安全事项的建设项目，还需按照我国国家安全审查和监管的制度和机制进行相应的国家安全审查。

（二）科技伦理安全

如前文所述，社会安全是 AI 安全的要点之一，而伦理安全是实现社会安全的基础。在我国，2017 年，国务院印发《新一代人工智能发展规划》并在该规划中提出了制定促进人工智能发展的法律法规和伦理规范的要求，之后《中华人民共和国科学技术进步法》《互联网信息服务算法推荐管理规定》《互联网信息服务深度合成管理规定》《生成式人工智能服务管理暂行办法》《网络安全标准实践指南—人工智能伦理安全风险防范指引》等一系列法律法规和相关规定相继对于 AI 伦理治理的要求予以规定。2022 年中共中央办公厅、国务院办公厅发布了《关于加强科技伦理治理的意见》，该意见是我国首个国家层面的、专门针对科技伦理治理的指导性文件，提出了科技伦理治理原则以及基本要求。2023 年 10 月发布的《科技伦理审查办法（试行）》对于科技伦理审查的基本程序、标准、条件等提出统一要求，标志着我国 AI 伦理治理监管体系建设进入了新阶段。我们在下表中梳理了我国与 AI 伦理治理相关的法律法规及其主要内容：

[3] 2015 年 7 月 1 日发布并实施。

序号	名称	生效日期	AI 伦理治理相关要求
1	《中华人民共和国科学技术进步法》	2022.01.01	**要求科学技术研究开发机构、高等学校企业事业单位等应当履行科技伦理管理主体责任，按照国家有关规定建立健全科技伦理审查机制**，对科学技术活动开展科技伦理审查，禁止危害国家安全、损害社会公共利益、危害人体健康、违背科研诚信和科技伦理的科学技术研究开发和应用活动，否则相关单位、直接负责的主管人员和其他直接责任人员将受到行政处罚。
2	《新一代人工智能发展规划》	2023.08.15	提出人工智能伦理规范和政策法规建设的战略目标，以及制定促进人工智能发展的法律法规和伦理规范的要求。
3	《互联网信息服务算法推荐管理规定》	2022.03.01	**算法推荐服务提供者**应当落实算法安全主体责任，建立健全算法机制机理审核、科技伦理审查、用户注册、信息发布审核数据安全和个人信息保护、反电信网络诈骗、安全评估监测、安全事件应急处置等管理制度和技术措施，制定并公开算法推荐服务相关规则，配备与算法推荐服务规模相适应的专业人员和技术支撑；此外，还应当定期审核、评估、验证算法机制机理、模型、数据和应用结果等，不得设置诱导用户沉迷、过度消费等违反法律法规或者违背伦理道德的算法模型。
4	《互联网信息服务深度合成管理规定》	2023.01.10	**深度合成服务提供者**应当落实信息安全主体责任，建立健全用户注册、算法机制机理审核、科技伦理审查、信息发布审核、数据安全、个人信息保护、反电信网络诈骗、应急处置等管理制度，具有安全可控的技术保障措施。
5	《科技伦理审查办理法（试行）》	2023.12.01	对于涉及**以人为研究参与者的科技活动，包括利用人类生物样本、个人信息数据等的科技活动，或不直接涉及人或实验动物但可能在生命健康、生态环境、公共秩序可持续发展等方面带来伦理风险挑战的科技活动进行的科技伦理审查和监管作出**了明确的规定。

序号	名称	生效日期	AI 伦理治理相关要求
6	《网络安全标准实践指南一人工智能伦理安全风险防范指引》	2021.01.05	将AI伦理安全风险总结为以下五大方面:（1）失控性风险：如 AI 的行为与影响超出服务提供者预设、理解和可控的范围，对社会价值等产生负面影响；（2）社会性风险：不合理使用 AI 而对社会价值等方面产生负面影响；（3）侵权性风险：AI 对人的基本权利，包括人身、隐私、侵权性风险财产等造成侵害或产生负面影响；（4）歧视性风险 :AI 对人类特定群体具有主观或客观偏见，影响公平公正、造成权利侵害或负面影响；（5）责任性风险: AI 相关各方行为失当、责任界定不清，对社会信任社会价值等方面产生负面影响。
7	《关于加强科技伦理治理的意见》	2022.03.20	提出“科技伦理是开展科学研究、技术开发等科技活动需要遵循的价值理念和行为规范，是促进科技事业健康发展的重要保障”，并明确了以下五大类科技伦理原则 : 增进人类福祉、尊重生命权利、坚持公平公正、合理控制风险和保持公开透明。

基于上述法律法规和相关规定的要求，AI 领域的企业应当承担相应 AI 伦理治理责任，具体而言：

- 在程序层面，如 AI 领域的企业涉及开展《科技伦理审查办法（试行）》适用范围内的科技活动，需自行设立科技伦理（审查）委员会或委托其他单位的科技伦理（审查）委员会对于所涉科技活动进行科技伦理审查，如涉及“具有舆论社会动员能力和社会意识引导能力的算法模型、应用程序及系统的研发”“面向存在安全、人身健康风险等场景的具有高度自主能力的自动化决策系统的研发”，还需报请所在地方或相关行业主管部门组织开展专家复核，亦即，对于特定领域，科技伦理审查将成为算法备案以及安全评估以外 AI 产品上线前的另一前置程序。
- 在实体层面，目前尚未出台法律法规对于 AI 伦理治理细则予以规定，尽管如此，一方面，目前《互联网信息服务算法推荐管理规定》《互联网信息服务深度合成管理规定》《生成式人工智能服务管理暂行办法》均要求大模型服务提供者履行伦理治理方面的合规义务；另一方面，《新一代人工智能伦理规范》[4]《网络安全标准实践指南—人工智能伦理安全风险防范指引》[5] 等与 AI 伦理治理相关的国家标准均

[4] 国家新一代人工智能治理专业委员会制定，2021 年 9 月 25 日发布并实施。
[5] 全国信息安全标准化技术委员会秘书处制定，2021 年 1 月 5 日发布并实施。

已发布，其他相关国家标准、团体标准以及行业标准亦均在制定过程中。因此，AI领域的企业可以参照前述国家标准，进行自我定位，明确自身的AI伦理治理责任。

（三）算法安全

网信办等九部门于2021年联合发布的《关于加强互联网信息服务算法综合治理的指导意见》以算法安全可信、高质量、创新性发展为导向，对健全算法安全治理机制、构建算法安全监管体系提出建议，包括加强算法治理规范、积极开展算法安全评估。为了解决算法设计或实施与预期不符、算法潜藏偏见与歧视、算法黑箱可解释性差等算法安全问题，《互联网信息服务算法推荐管理规定》（“《算法推荐管理规定》”）、《互联网信息服务深度合成管理规定》（“《深度合成管理规定》”）、《生成式人工智能服务管理暂行办法》[6]（“《AIGC暂行办法》”）、《具有舆论属性或社会动员能力的互联网信息服务安全评估规定》[7]（“《安全评估规定》”）等法律法规对算法备案、安全评估、算法公开、算法管理等进行规定，以实现算法的透明性、公平性、可控性，具体而言：

- 在算法备案方面，《算法推荐管理规定》《深度合成管理规定》《AIGC暂行办法》都对算法推荐服务、深度合成服务、生成式人工智能服务提出了算法备案要求。根据《算法推荐管理规定》，具有舆论属性或者社会动员能力的算法推荐服务提供者应当在提供服务之日起十个工作日内通过互联网信息服务算法备案系统填报服务提供者的名称、服务形式、应用领域、算法类型、算法自评估报告、拟公示内容等信息，履行备案手续。
- 在安全评估方面，《算法推荐管理规定》《深度合成管理规定》《AIGC暂行办法》要求对于具有舆论属性或社会动员能力的算法推荐服务、深度合成服务、生成式人工智能服务按照《安全评估规定》通过全国互联网安全管理服务平台完成安全评估。根据《安全评估规定》，下述类型的互联网信息服务提供者需按《安全评估规定》自行进行安全评估：（i）开办论坛、博客、微博客、聊天室、通讯群组、公众账号、短视频、网络直播、信息分享、小程序等信息服务或者附设相应功能；（ii）开办提供公众舆论表达渠道或者具有发动社会公众从事特定活动能力的其他互联网信息服务。在此基础上，《安全评估规定》规定了互联网信息服务提供者应自行进行安全评估的具体情形。除进行自行安全评估的义务以外，《安全评估规定》还要求前述互联网信息服务提供者应履行消除安全隐患、形成安全评估报告、提交安全评估报告等各项义务。除前述安全评估以外，《AIGC暂行办法》还特别要求深度合成服务提供者和技术支持者提供具有以下功能的模型、模板等工具的，

[6] 2023年7月10日发布，2023年8月15日实施。
[7] 2018年11月15日发布，2018年11月30日实施。

应当依法自行或者委托专业机构开展安全评估：（一）生成或者编辑人脸、人声等生物识别信息的；（二）生成或者编辑可能涉及国家安全、国家形象、国家利益和社会公共利益的特殊物体、场景等非生物识别信息的。

- 在算法公开方面，《算法推荐管理规定》要求算法推荐服务提供者优化检索、排序、选择、推送、展示等规则的透明度和可解释性，以显著方式告知用户其提供算法推荐服务的情况，并以适当方式公示算法推荐服务的基本原理、目的意图和主要运行机制等。《AIGC 暂行办法》要求生成式人工智能服务提供者应当基于服务类型特点，采取有效措施，提升生成式人工智能服务的透明度，提高生成内容的准确性和可靠性；明确并公开其服务的适用人群、场合、用途；如有关主管部门依据职责对生成式人工智能服务开展监督检查，生成式人工智能服务提供者应当依法予以配合，按要求对训练数据来源、规模、类型、标注规则、算法机制机理等予以说明。
- 在算法管理方面，《算法推荐管理规定》要求算法推荐服务提供者应当定期审核、评估、验证算法机制机理、模型、数据和应用结果等，不得设置诱导用户沉迷、过度消费等违反法律法规或者违背伦理道德的算法模型。《AIGC 暂行办法》要求生成式人工智能服务不得生成民族歧视内容，且在算法设计、训练数据选择、模型生成和优化、提供服务等过程中，应采取有效措施防止产生民族、信仰、国别、地域、性别、年龄、职业、健康等歧视。

基于上述法律法规和相关规定的要求，我们理解针对 AI 产品，需要履行算法备案、安全评估等监管手续，此外还需遵守前述算法公开、算法管理等要求。值得注意的是，在算法备案方面，在选择“生成合成类（深度合成）算法”这一算法类型进行算法备案时需要区分备案主体身份（“深度合成服务技术支持者”或“深度合成服务提供者”），即 AI 产品的服务提供方和技术支持方需要作为不同的备案主体对同一算法进行备案，二者在算法备案项下的义务相互独立而不可互相替代。在安全评估方面，除需要对 AI 产品按照《安全评估规定》通过全国互联网安全管理服务平台完成安全评估以外，还需进行新技术新应用安全评估（“双新评估”），而关于双新评估的具体流程以及要求仍有待监管部门进一步公开。

（四）信息安全

我国法律法规和相关规定主要规定了内容标识以及内容治理等方面的要求，以防范深度合成技术、生成式人工智能服务以及其他新技术新应用对信息的生成和传播造成的风险，

具体而言：

- 在内容标识方面，《网络音视频信息服务管理规定》[8]《深度合成管理规定》《AIGC 暂行办法》对使用深度合成服务、生成式人工智能服务、基于深度学习 / 虚拟现实等新技术新应用生成的内容标识要求进行规定，其中，《深度合成管理规定》规定了内容标识的具体要求。根据《深度合成管理规定》，深度合成服务提供者应根据深度合成服务的类型对使用其服务生成或者编辑的信息内容进行标识：（1）使用一般深度合成服务生成或者编辑的信息内容，应当添加不影响用户使用的标识；（2）使用具有生成或者显著改变信息内容功能的深度合成服务[9]生成或者编辑的信息内容，且可能导致公众混淆或者误认的，应当在生成或者编辑的信息内容的合理位置、区域进行显著标识。此外，为贯彻落实《AIGC 暂行办法》有关内容标识的要求，指导生成式人工智能服务提供者等有关单位做好内容标识工作，全国信息安全标准化技术委员会秘书编制了《网络安全标准实践指南——生成式人工智能服务内容标识方法》（TC260-PG-20233A）并于 2023 年 8 月发布，该指南围绕文本、图片、音频、视频四类生成内容给出了具体的内容标识方法，包括添加显示区域标识、提示文字标识、隐藏水印标识、文件元数据标识、提示语音标识。
- 内容治理方面，《网络音视频信息服务管理规定》要求网络音视频信息服务提供者和网络音视频信息服务使用者不得利用基于深度学习、虚拟现实等的新技术新应用制作、发布、传播虚假新闻信息；网络音视频信息服务提供者应当建立健全辟谣机制，发现网络音视频信息服务使用者利用基于深度学习、虚拟现实等的虚假图像、音视频生成技术制作、发布、传播谣言的，应当及时采取相应的辟谣措施，并将相关信息报网信、文化和旅游、广播电视等部门备案。《深度合成管理规定》要求深度合成服务提供者应当建立健全违法和不良信息识别特征库，并对发现的违法和不良信息依法采取处置措施、记录网络日志并及时向网信部门和有关主管部门报告。《AIGC 暂行办法》要求生成式人工智能服务提供者应当依法承担网络信息内容生产者责任，履行网络信息安全义务；根据《网络信息内容生态治理规定》[10]，网络信息内容生产者是指制作、复制、发布网络信息内容的组织或者个人；网络信息内容生产者责任包括不得制作、复制、发布违法信息，且应当采取措施，防范和抵制制作、复制、发布不良信息。

---

[8] 2019 年 11 月 18 日发布，2020 年 1 月 1 日实施。

[9] 根据《深度合成管理规定》第十七条，具有生成或者显著改变信息内容功能的服务包括：（一）智能对话、智能写作等模拟自然人进行文本的生成或者编辑服务；（二）合成人声、仿声等语音生成或者显著改变个人身份特征的编辑服务；（三）人脸生成、人脸替换、人脸操控、姿态操控等人物图像、视频生成或者显著改变个人身份特征的编辑服务；（四）沉浸式拟真场景等生成或者编辑服务；（五）其他具有生成或者显著改变信息内容功能的服务。

[10] 2019 年 12 月 15 日发布，2020 年 3 月 1 日实施。

根据前述法律法规和相关要求，针对利用 AI 技术生成的内容，服务提供者应根据 AI 技术类型以及内容类型进行不同形式的标识，以与其他内容进行区分。此外，为了防范利用 AI 技术生成的违法和不良信息的传播，服务提供者应建立健全违法和不良信息识别特征库，避免生成违法和不良信息，并对发现的违法和不良信息依法采取处置措施、记录网络日志并及时向网信部门和有关主管部门报告。为了防范利用 AI 技术生成的虚假信息的传播，服务提供者不得生成虚假新闻内容，并应建立健全辟谣机制，对发现的谣言及时采取相应的辟谣措施并将相关信息报网信、文化和旅游、广播电视等部门备案。

（五）网络安全

我国网络安全法律体系以《中华人民共和国网络安全法》[11] 为基础，并通过《网络安全审查办法》[12]《网络产品安全漏洞管理规定》[13] 等配套法规对于《中华人民共和国网络安全法》中的原则性规定予以细化，具体而言：

- 《中华人民共和国网络安全法》规定了网络运营者应当履行的网络安全保护义务。网络运营者，是指由计算机或者其他信息终端及相关设备组成的按照一定的规则和程序对信息进行收集、存储、传输、交换、处理的系统的所有者、管理者和网络服务提供者。网络运营者的网络安全保护义务涉及网络运行安全、网络信息安全等两个方面。从网络运行安全的角度出发，网络运营者应当按照网络安全等级保护制度的要求，履行下列安全保护义务，保障网络免受干扰、破坏或者未经授权的访问，防止网络数据泄露或者被窃取、篡改：制定内部安全管理制度和操作规程，采取防范危害网络安全行为、监测记录网络安全事件等的技术措施，制定网络安全事件应急预案应对危害网络安全的事件，并在发生相应事件时向有关主管部门报告。从网络信息安全的角度出发，网络运营者应当设立用户信息保护制度，并采取技术措施和其他必要措施确保其收集的个人信息安全，防止信息泄露、毁损、丢失。
- 《网络安全审查办法》要求关键信息基础设施运营者采购网络产品和服务，网络平台运营者开展数据处理活动，影响或者可能影响国家安全的，应当进行网络安全审查。此外，掌握超过 100 万用户个人信息的网络平台运营者赴国外上市，必须向网络安全审查办公室申报网络安全审查。网络安全审查将重点评估相关对象或者情形的以下国家安全风险因素：

  （一）产品和服务使用后带来的关键信息基础设施被非法控制、遭受干扰或者破坏的风险；

[11] 2016 年 11 月 7 日发布并实施。
[12] 2021 年 12 月 28 日发布，2022 年 2 月 15 日实施。
[13] 2021 年 7 月 12 日发布，2021 年 9 月 1 日实施。

（二）产品和服务供应中断对关键信息基础设施业务连续性的危害；

（三）产品和服务的安全性、开放性、透明性、来源的多样性，供应渠道的可靠性以及因为政治、外交、贸易等因素导致供应中断的风险；

（四）产品和服务提供者遵守中国法律、行政法规、部门规章情况；

（五）核心数据、重要数据或者大量个人信息被窃取、泄露、毁损以及非法利用、非法出境的风险；

（六）上市存在关键信息基础设施、核心数据、重要数据或者大量个人信息被外国政府影响、控制、恶意利用的风险，以及网络信息安全风险；

（七）其他可能危害关键信息基础设施安全、网络安全和数据安全的因素。

- 《网络产品安全漏洞管理规定》对于网络产品提供者、网络运营者和网络产品安全漏洞收集平台的漏洞发行、报告、修补和发布行为进行规定。其中，网络产品提供者是指中华人民共和国境内的网络产品（含硬件、软件）提供者。我们对于《网络产品安全漏洞管理规定》规定的网络产品提供者和网络运营者分别应承担的网络产品安全漏洞管理义务梳理如下：

主体类型	网络产品安全漏洞管理义务
网络产品提供者	• 建立健全网络产品安全漏洞信息接收渠道并保持畅通留存网络产品安全漏洞信息接收日志不少于 6 个月； • 发现或者获知所提供网络产品存在安全漏洞后，应当立即采取措施并组织对安全漏洞进行验证，评估安全漏洞的危害程度和影响范围；对属于其上游产品或者组件存在的安全漏洞，应当立即通知相关产品提供者应当在 2 日内向工业和信息化部网络安全威胁和漏洞信息共享平台报送相关漏洞信息。报送内容应当包括存在网络产品安全漏洞的产品名称、型号、版本以及漏洞的技术特点、危害和影响范围等； • 应当及时组织对网络产品安全漏洞进行修补，对于需要产品用户（含下游厂商）采取软件、固件升级等措施的，应当及时将网络产品安全漏洞风险及修补方式告知可能受影响的产品用户，并提供必要的技术支持。
网络运营者	• 建立健全网络产品安全漏洞信息接收渠道并保持畅通留存网络产品安全漏洞信息接收日志不少于 6 个月； • 发现或者获知其网络、信息系统及其设备存在安全漏洞后，应当立即采取措施，及时对安全漏洞进行验证并完成修补。

综上所述，AI 产品的服务提供方以及技术支持方作为网络运营者，应当履行《中华人民共和国网络安全法》规定的一般网络安全保护义务，包括建立网络安全等级保护制度，并按照网络安全等级保护制度的要求，履行安全保护义务，并履行个人信息保护义务（详见下文数据安全部分的分析）。值得注意的是，作为实施网络安全等级保护的程序性要求，AI 产品的服务提供方以及技术支持方应当完成对象系统的定级、安全建设与整改、测评、备案等相关法定程序。特别地，AI 产品的服务提供方以及技术支持方作为网络产品提供者、网络运营者，还应履行《网络产品安全漏洞管理规定》规定的网络产品安全漏洞管理义务。此外，如 AI 产品的服务提供方、技术支持方自身涉及触发网络安全审查的情形（例如开展数据处理活动影响或者可能影响国家安全的、掌握超过 100 万用户个人信息的网络平台运营者赴国外上市），应自查是否存在网络安全审查重点评估的国家安全风险因素，并且加强对于上游网络产品和服务提供者的管理；如 AI 产品的服务提供方、技术支持方涉及向网络安全审查对象（例如关键基础设施运营者）提供网络产品或服务的，应当配合网络安全审查对象履行相应的合规义务。

（六）数据安全

数据（含个人信息，下同）是 AI 的重要生产资料之一。数据安全（含个人信息安全与保护，下同）将直接影响 AI 的发展安全。

随着《中华人民共和国网络安全法》《中华人民共和国数据安全法》[14]、《中华人民共和国个人信息保护法》[15] 以及其他数据安全相关的法律法规、国家标准的陆续出台，我国已基本形成了数据从收集、存储、使用、加工、传输、提供、公开、删除等全生命周期的合规体系。

就 AI 治理而言，我国多部法律法规和相关规定重申了数据安全的基本要求，并结合 AI 的特点提出了有针对性的规范：

（1）《算法推荐管理规定》要求算法推荐服务提供者应当落实数据安全管理制度和技术措施 [16]；

（2）《深度合成管理规定》重申了前述规定，并要求深度合成服务提供者和技术支持者应当加强训练数据管理，采取必要措施保障训练数据安全，以及训练数据包含个人信息的，应当遵守个人信息保护的有关规定 [17]；

（3）《AIGC 暂行办法》同样强调提供生成式人工智能服务不得侵害他人隐私权和个

[14] 2021 年 6 月 10 日发布，2021 年 9 月 1 日实施。
[15] 2021 年 8 月 20 日发布，2021 年 11 月 1 日实施。
[16] 《互联网信息服务算法推荐管理规定》第 7 条。
[17] 《互联网信息服务深度合成管理规定》第 14 条。

人信息权益[18]，以及生成式人工智能服务提供者应当依法承担个人信息处理者责任、履行个人信息保护义务[19]。此外，关于 AI 预训练、优化训练等数据处理活动，《AIGC 暂行办法》强调，生成式人工智能服务提供者应使用具有合法来源的数据和基础模型；采取有效措施提高训练数据质量，增强训练数据的真实性、准确性、客观性、多样性；涉及处理个人信息的，应当获得个人同意或具备其他合法性基础[20]；以及

（4）根据《科技伦理审查办法（试行）》，AI 科技活动是否符合数据安全的要求是科技伦理审查内容之一。例如，对于涉及以人为研究参与者的 AI 科技活动，需审查个人隐私数据、生物特征信息等信息处理是否符合个人信息保护的有关规定；对于涉及数据和算法的 AI 科技活动，需审查数据的收集、存储、加工、使用等处理活动以及研究开发数据新技术等是否符合国家数据安全等有关规定，以及数据安全风险监测及应急处理方案是否得当[21]。

结合前述法律法规及 AI 产品的生命周期，需从 AI 模型训练、应用和优化三个阶段分别关注各阶段的数据安全重点，以及日常数据安全体系的搭建和管理：

（1）AI 模型训练阶段：模型训练阶段主要涉及数据采集、数据清洗及数据标注等活动。从数据安全的角度，该阶段需重点关注训练数据来源合法合规：训练数据中包含个人信息的，应当取得相关个人信息主体的知情同意或具备其他合法性基础，或者可在数据清洗阶段对相关个人信息进行脱敏处理；针对非个人信息，需注意训练数据中是否可能包含国家秘密、核心数据或重要数据等可能关系国家安全、经济运行、社会稳定、公共健康和安全等的数据，若包含该等数据则需在数据清洗阶段对其进行脱敏处理或采取其他必要的措施。

（2）AI 模型应用阶段：模型应用阶段主要涉及用户使用基于 AI 模型构建的 AI 产品输入内容，AI 产品相应生成内容（主要针对生成式 AI）。从数据安全的角度，该阶段需重点关注用户输入数据来源合法合规和 AI 生成内容的数据泄露风险，以及若 AI 产品的服务器在境外，还需关注数据跨境传输的风险及相关合规要求（包括取得相关个人信息主体的知情单独同意、个人信息保护影响评估及视情况完成安全评估、备案或认证等）。

（3）AI 模型优化阶段：模型优化阶段主要涉及将 AI 模型应用阶段采集的数据作为训

[18] 《生成式人工智能服务管理暂行办法》第 4 条。
[19] 《生成式人工智能服务管理暂行办法》第 9 条、第 11 条。
[20] 《科技伦理审查办法 (试行)》第 15 条。
[21] 2020 年 10 月 17 日发布，2020 年 12 月 1 日实施。

练数据进一步优化 AI 模型。从数据安全的角度，该阶段需重点关注就使用用户输入数据进一步优化模型时注意用户输入数据中是否包含涉及国家安全、经济运行、社会稳定、公共健康和安全等的敏感数据，并在涉及个人信息时确保合规。

（4）日常数据安全体系的搭建和管理：需注意落实数据安全管理制度和技术措施，依法承担数据保护义务。此外，在针对 AI 科技活动开展科技伦理审查时，需关注该等科技活动是否符合数据安全的要求。

（七）技术出口安全

一般来讲，我国根据出口技术是否具有特殊属性，分别制定了两套不同的出口管控法律制度：

（1）管制技术：国家对两用物项、军品、核以及其他与维护国家安全和利益、履行防扩散等国际义务相关的货物、技术、服务等物项（统称“管制物项”）的出口活动专门制定了出口管制法律法规。其中，与企业经营活动关系较为密切的就是两用技术。涉及两用技术出口管制相关法律制度主要包括《中华人民共和国出口管制法》[22] 以及《两用物项和技术进出口许可证管理目录》[23]（下称“两用物项目录”）等。另外，商务部自 2022 年 4 月 22 日发布的《两用物项出口管制条例（征求意见稿）》预计即将成法落地，建议企业予以关注；

（2）民用技术：针对一般民用技术的出口管控法规主要包括《中华人民共和国技术进出口管理条例》[24]、《禁止出口限制出口技术管理办法》[25] 以及《中国禁止出口限制出口技术目录》[26] 等。

根据初步筛查结果，未见 2024 年度两用物项目录中列出人工智能相关的物项或技术。与此同时，经查询《禁止出口限制出口技术目录》，在该目录第十五类“计算机服务业”和第十六类“软件业”的限制出口技术项下新增的多项与人工智能相关的技术，如语音合成技术、语音识别技术、交互理解技术、印刷体扫描识别技术、手写体扫描识别技术、拍照识别技术、中英文作文批改技术，特别是基于数据分析的个性化信息推送服务技术等。此外，还新增了多项与网络安全相关的技术，如密码芯片设计和实现技术、量子密码技术以及数据库系统安全增强技术等。此外，目前与人工智能相关的技术没有被列入禁止出口的技术目录，因此，人工智能相关的民用技术目前属于自由出口的技术或者限制出口的技术。由于《禁止出口限制出口技术目录》中表述较为宽泛，涉及人工智能技术出口的企业

[22] 2020 年 10 月 17 日发布，2020 年 12 月 1 日实施。
[23] 2023 年 12 月 29 日发，2024 年 1 月 1 日实施。
[24] 2020 年 11 月 29 日发布并实施。
[25] 2009 年 4 月 20 日发布，2009 年 5 月 20 日实施。
[26] 2023 年 12 月 21 日发布并实施。

应当就具体的拟出口技术对照《禁止出口限制出口技术目录》，分析其是否属于限制出口的技术。如企业对于拟出口技术的分类存疑，可以向国家出口管制管理部门提出咨询，由出口管制管理部门给出确定的意见。

对于自由出口的技术，企业应当按照相关法律法规要求，向国务院外经贸主管部门办理登记；对于属于限制出口的人工智能技术，企业应当按照法律要求的时间期限，办理《中华人民共和国技术出口许可证》。如果限制出口的技术的出口经营者在没有取得前述出口许可的情况下擅自出口相关技术，可能被判定为刑法下的走私罪、非法经营罪等，从而被追究刑事责任；尚不够刑事处罚的，将根据海关法的规定进行处罚，或由国务院外经贸主管部门给予警告、罚款、撤销外贸易经营许可等行政处罚。

综合上述法律规定，我们结合实务中的经验，对拟出口人工智能技术的企业提出以下建议：

（1）全面梳理跨境技术合作、投资并购、培训交流、内部技术转移等可能涉及技术出境的风险场景，将技术受控情况纳入境外投资前的立项研究和分析，在审批流程中加入关于技术出境情况和受控情况判断的环节。

（2）由于《禁止出口限制出口技术目录》中对于人工智能相关的技术表述较为宽泛，需要企业将拟向境外传输的人工智能技术与前述目录中的技术描述进行对比分析，判断拟向境外传输的技术是否属于限制出口的技术。

（3）对于经过前述评估分析认为拟出口的技术可能属于涉密技术或限制出口的技术，应当按照法律规定，在法律要求的时间节点及时向有关部门进行申报，申请必要的出口许可。

（4）由于人工智能技术的发展速度较快，对国家安全的影响在不断变化，未来国家可能会在人工智能技术的受控情况方面进行政策或法律的变更或调整，企业应密切关注相关出口管制限制以及中国反制政策的变化，动态调整企业对于受限技术的出口管理。

三、AI 安全前瞻性合规风险预测

除前述现行有效的法律法规和相关规定项下的合规要点以外，随着我国对于 AI 产品研发、设计、部署、运行等各阶段进行约束的法律法规和国家标准体系逐步完善，我们对于 AI 安全前瞻性合规风险预测如下：

- 在生成式人工智能领域，全国信息安全标准化技术委员会于 2024 年 2 月 29 日发

布《生成式人工智能服务安全基本要求》（“《AIGC 安全基本要求》”），以《中华人民共和国网络安全法》《网络信息内容生态治理规定》《AIGC 暂行办法》《信息安全技术 个人信息安全规范》（GB/T35273）、《网络安全标准实践指南——生成式人工智能服务内容识别方法》等法律法规和相关规定以及国家标准为基础，对于生成式人工智能服务的基本安全要求予以规定，包括语料（即训练数据）安全、模型安全等方面的具体要求、人工智能服务提供者应遵循的相应安全措施以及安全评估的程序和内容。值得注意的是，《AIGC 安全基本要求》明确要求生成式人工智能服务提供者在进行备案时，应按照《AIGC 安全基本要求》进行安全评估，并提交评估报告；在安全评估的程序和内容方面，《AIGC 安全基本要求》规定了语料安全、生成内容安全和问题拒答的评估方法，要求安全评估的内容应覆盖《AIGC 安全基本要求》第 5 章至第 8 章中的所有条款，且每个条款应形成单独的评估结论，并与相关证明、支撑材料形成最终的评估报告。在《AIGC 安全基本要求》发布后，其中有关安全评估的要求将弥补我国目前在“双新评估”具体流程和要求方面的立法空白，为 AI 领域的企业提高 AI 产品的安全水平以及相关主管部门评价 AI 产品的安全水平提供参考。

- 在军事国防安全领域，2017 年，国务院印发了《新一代人工智能发展规划》，指出人工智能在国防建设领域得到广泛应用。未来，人工智能将有力提升情报、侦察、通讯、后勤等军事装备的自主性与智能化，并广泛应用于信息情报搜集和分析、战略设计、实施精准打击等，但与此同时，人工智能可能会触发各国人工智能军备竞赛，使未来战争的精准度和破坏力大大提升，增加了新型军事安全威胁。
- 在航空安全领域，国务院和中央军事委员会发布的《无人驾驶航空器飞行管理暂行条例》（2024 年 1 月 1 日实施）第五条规定，“国家鼓励无人驾驶航空器科研创新及其成果的推广应用，促进无人驾驶航空器与大数据、人工智能等新技术融合创新”。人工智能技术在航空领域的应用可能会对航空安全管理以及应急事件处置提出新的课题和挑战。此外，人工智能技术在火箭以及卫星行业的应用将大大提升卫星的通信能力以及数据收集和处理能力，而依据卫星收集的精确数据可以针对特定行业，如气象、水利等，进行深度分析，此类分析结论可能对国家安全产生潜在的影响。

四、AI 合规体系及搭建指引

综合上述内容，当前我国的实践中，AI 领域的企业应当考虑搭建相应的 AI 安全与合

规体系从而对于 AI 的安全与合规问题形成系统性保障，例如：

（一）内部合规管理体系及制度

为满足我国法律法规和相关规定以及国家标准项下关于 AI 安全的要求，建议 AI 领域企业结合其所开展的活动以及可能涉及的 AI 安全领域，建立相应的内部合规管理体系及制度。其中，就资质证照而言，除了增值电信业务经营许可证 /ICP 备案、公安联网备案、算法备案、安全评估等基础性资质外，AI 领域企业还应根据自身的业务模式综合判断是否需要办理特殊业务领域的资质证照，例如网络文化经营许可证、网络出版许可证、信息网络传播视听节目许可证。除此之外，AI 领域企业还应结合自身的业务定位建立相应的内控制度，例如科技伦理审查制度（关于科技伦理治理以及科技伦理（审查）委员会的设立，请见本书《人工智能（AI）：科技伦理治理走起》《科技伦理（审查）委员会：如何设立？》等文章）、AI 内容安全基础审查制度（包括 AI 语料安全审查制度、AI 模型安全制度等）、数据安全制度（包括数据采集合规制度、数据处理合规制度、算法安全及伦理合规制度等）、内容生态治理制度、用户权益保护制度等。关于 AI 合规制度的设立，请见本书《AIGC：合规引领探索之路》一文。

（二）外部合作层面

AI 领域的企业在对外合作过程中还会与诸多合作方签署不同的商业合作协议。该等商业合作协议根据 AI 领域的企业自身的性质与合作事项的不同包括 AI 模型开发 / 许可协议、数据交易协议等诸多类型，AI 领域的企业需要根据不同的协议类型以及自身的商业安排与合作方就各自的权利义务达成一致，在此过程中应当考虑 AI 安全方面的要求。例如，如企业自身属于网络安全审查的对象，为加强对于上游网络产品或服务提供者的管理，企业应当在与该等上游网络产品或服务提供者签署的商业合作协议中对于上游网络产品或服务提供者的配合网络安全审查义务、保密义务、安全保护义务等进行明确约定。

（三）互联网应用层面

倘若 AI 领域企业的业务涉及 AIGC 平台等互联网应用，还应当考虑到 AIGC 平台对于信息安全、数据安全造成的风险，结合 AI 业务的具体情况，针对用户协议、隐私政策、平台治理规则等文件进行修订。例如，为治理违法和不良信息，AIGC 平台的运营方可在用户协议中明确约定，用户需要为其所输入的内容承担相应的责任，如保证不侵犯第三方权利，保证不得含有违反法律法规或公序良俗的信息，并且不得通过大量的负面数据输入企图扭曲模型的输出等，并制定相应的处置措施，例如警告、终止服务等等。

结语

AI 技术在金融、医疗、教育等日常生活领域以及生物、医学、材料科学等学科的学术研究领域的广泛应用，为 AI 领域的企业创造了巨大机遇，但同时也蕴含着不容忽视的风险，如《布莱切利宣言》中所述，人工智能可能表现出的造成安全问题的能力可与当今最先进的模型相提并论甚至可能超越它们。因此，无论企业属于 AI 产品的研究开发者、设计制造者还是部署应用者，必须高度关注 AI 安全，通过搭建 AI 安全与合规体系，履行保障 AI 安全的义务和责任，共同维护国家安全的新疆域。

感谢律师冯宝宝、单文钰、蒋孟菲、张一凡、朱佳蔚，律师助理米华林，实习生缪逸泓、张文溢、何一辰、黄若湉对本文作出的贡献。

境内平台使用 ChatGPT？至少注意这些

张逸瑞 冯宝宝 张一凡

引言

2023 年 9 月，实践中，针对部分调用 OpenAI 的 ChatGPT 产品接口的境内平台，部分地区的互联网信息办公室（即网信办）根据《网信部门行政执法程序规定》第三十八条[1]、《生成式人工智能服务管理暂行办法》（“《AIGC 暂行办法》”）第十七条[2]、第二十一条[3]，与该等境内平台的平台运营方进行了约谈。与此同时，如我们在本书《大模型合规之现实初探》一文中所述，自 2023 年 7 月底，苹果 APP Store 对中国内地中大量提供类 ChatGPT 服务的应用进行集中下架，其中不少亦是利用 OpenAI 提供的 API 服务所开发的应用，根据部分平台运营方收到的苹果官方下架通知，应用下架的主要原因在于应用与 ChatGPT 相关，而 ChatGPT 在中国没有获得运营所必需的许可。

该等实践意味着在《AIGC 暂行办法》正式生效后，境内平台利用境外技术支持方提供的服务向中国境内用户提供生成式人工智能服务的合规要求正在逐渐全面跟上。以下，我们针对该等具体合规要求展开讨论，供境内平台的平台运营方在与境外技术支持方进行合作的过程中进行参考。

总体而言，境内平台使用境外生成式人工智能服务主要通过直接调用境外生成式人工智能服务、接入境外生成人工智能服务的可编程接口（即 API）等方式实现。以 OpenAI 为例，根据 OpenAI 官网的说明，目前 OpenAI 提供的服务包括 ChatGPT、DALL-E 等非 API 服务（non-API consumer services）以及 API 服务这两种类型。不过，目前 OpenAI 的前述两种服务均未面向中国内地以及中国香港地区的用户开放；实践中，境内平台主要借助于虚拟专用网络（即 VPN）通过中国境外 IP 地址使用 OpenAI 提供的服务，该种使用方式本身存在一定的违规风险，例如违规建立或租用虚拟专用网络等国际联网相关风险。

---

[1] 《网信部门行政执法程序规定》第三十八条规定，网信部门对当事人作出行政处罚决定前，可以根据有关规定对其实施约谈，谈话结束后制作执法约谈笔录。

[2] 《生成式人工智能服务管理暂行办法》第十七条规定，提供具有舆论属性或者社会动员能力的生成式人工智能服务的，应当按照国家有关规定开展安全评估，并按照《互联网信息服务算法推荐管理规定》履行算法备案和变更、注销备案手续。

[3] 《生成式人工智能服务管理暂行办法》第二十一条规定，提供者违反本办法规定的，由有关主管部门依照《中华人民共和国网络安全法》、《中华人民共和国数据安全法》、《中华人民共和国个人信息保护法》、《中华人民共和国科学技术进步法》等法律、行政法规的规定予以处罚；法律、行政法规没有规定的，由有关主管部门依据职责予以警告、通报批评，责令限期改正；拒不改正或者情节严重的，责令暂停提供相关服务。构成违反治安管理行为的，依法给予治安管理处罚；构成犯罪的，依法追究刑事责任。

即使境外主体提供的生成式人工智能服务向中国境内用户开放，在境内平台使用该等服务向境内用户提供生成式人工智能服务的情况下，如本书《大模型合规之现实初探》一文所述，利用生成式人工智能技术向中华人民共和国境内公众提供生成文本、图片、音频、视频等内容服务的组织、个人均属于《AIGC 暂行办法》项下的生成式人工智能服务提供者，应当履行生成式人工智能服务提供者的责任与义务。因此，境外主体以及境内平台的平台运营方均属于生成式人工智能服务提供者，应当遵守《AIGC 暂行办法》《互联网信息服务深度合成管理规定》（“《深度合成管理规定》”）《互联网信息服务算法推荐管理规定》（“《算法推荐管理规定》”）等中国相关法律法规的规定，倘若存在违规情况，根据《AIGC 暂行办法》，国家网信部门有权通知有关机构采取技术措施和其他必要措施对境外的技术支持方予以处置，而该等处置措施极有可能会对使用境外生成式人工智能服务的境内平台造成较大影响。

当前，境内平台使用境外生成式人工智能服务的合规要点主要在资质要求、数据跨境合规、数据训练合规等方面，具体如下：

一、资质要求

（一）算法备案

为满足生成式人工智能服务的透明度要求，根据《AIGC 管理办法》《深度合成管理规定》《算法推荐管理规定》，平台运营方、技术支持方均应当履行算法备案的义务。因此，在境内平台使用境外生成式人工智能服务的场景下，境外的技术支持方以及境内的平台运营方均应当进行算法备案，具体而言[4]：

- 在算法类型方面，境外的技术支持方以及境内的平台运营方均应当选择“生成合成类（深度合成）算法”这一算法类型进行算法备案；
- 在备案角色方面，境外的技术支持方应当作为深度合成服务技术支持者进行算法备案，境内平台的平台运营方应当作为深度合成服务提供者进行算法备案；

《算法推荐管理规定》进一步明确，如技术支持方以及平台运营方未依法履行备案手续的，由网信部门和电信、公安、市场监管等有关部门依据职责给予警告、通报批评，责令限期改正；拒不改正或者情节严重的，责令暂停信息更新，并处一万元以上十万元以下罚款。

（二）安全评估

根据《AIGC 管理办法》《深度合成管理规定》《算法推荐管理规定》《具有舆论属

[4] 详见《互联网信息服务深度合成管理规定》备案填报指南，链接地址：https://beian.cac.gov.cn/api/file/fileDownLoad?noticeId=notice_4b62813c-b5cd-4bf0-b1ff-5c140decda7f，最后访问日期：2024 年 3 月 19 日。

性或社会动员能力的互联网信息服务安全评估规定》（“《安全评估规定》”），境外的技术支持方、境内平台的平台运营方还应当完成以下两种类型的安全评估：一是按照《安全评估规定》通过全国互联网安全管理服务平台完成具有舆论属性或社会动员能力的互联网信息服务安全评估；二是针对生成式人工智能服务进行新技术新应用安全评估（即双新评估），而关于双新评估的具体流程以及要求仍有待监管部门进一步公开。

根据《安全评估规定》，如技术支持方、平台运营方拒不依法开展安全评估的，网信部门和公安机关将通过全国互联网安全管理服务平台向公众提示其提供的服务存在安全风险。

二、数据跨境合规

在境内平台使用境外生成式人工智能服务的场景下，中国境内用户在境内平台的输入端口提出问题后，该问题会传输到位于境外的技术支持方，技术支持方模型给出相应回答后，该回答便会传输到境内平台的用户端口以实现对问题的反馈。按照该服务模式，境内平台的平台运营方向境外技术支持方传输用户输入数据的过程中，平台运营方有可能涉及将中国境内用户的个人信息传输至境外。

在此种情形下，对于平台运营方而言，平台运营方应当按照《中华人民共和国个人信息保护法》（“《个人信息保护法》”）《数据出境安全评估办法》《个人信息出境标准合同办法》等相关法律法规履行个人信息跨境传输相关的合规要求，包括数据出境安全评估 / 个人信息保护影响评估、个人信息出境标准合同签订和备案、用户告知等。

对于境外技术支持方而言，在技术支持方与平台运营方签订个人信息出境标准合同的情形下，技术支持方应当履行该等标准合同项下境外接受方的义务，例如确保个人信息的保存期限为实现处理目的所必要的最短时间，保存期限届满的，应当删除个人信息（包括所有备份）。同时，技术支持方还应当按照标准合同的约定，结合其所在国家或者地区的个人信息保护政策和法规，对于该等政策和法规对于技术支持方履行标准合同约定义务的影响进行评估。

根据《个人信息保护法》，如平台运营方、技术支持方违反上述要求，平台运营方、技术支持方将被中国境内的主管部门处以责令改正、给予警告、没收违法所得、罚款等行政处罚；违法处理个人信息的应用程序，将被责令暂停或者终止提供服务；直接负责的主管人员和其他直接责任人员将被处以一万元以上十万元以下罚款。此外，平台运营方、技术支持方还可能因违反技术支持方所在国家或者地区可适用的个人信息保护政策和法规面临相应的处罚风险。

三、数据训练合规

如我们在本书《大模型合规之现实初探》一文中所述，数据是大模型最底层的“原料”，而数据训练是对“原料的使用”，数据训练合规是满足服务生成内容合规的重要前提，技术支持方、平台运营方应当按照《AIGC 暂行办法》[5] 的要求，开展预训练、优化训练等训练数据处理活动。

在境内平台使用境外生成式人工智能服务的场景下，境内平台运营方应当特别关注数据和基础模型来源合法、知识产权合规以及个人信息保护等方面的要求。具体而言，

（一）数据和基础模型的来源合法

关于数据和基础模型，一般由境外技术支持方提供基础模型以及该等基础模型的训练数据。为满足相应的合规要求，平台运营方应对于境外技术支持方提供的基础模型和数据来源的合法性进行必要的审查，对技术支持方数据安全保护能力开展尽职调查。在平台运营方与技术支持方签署的相关技术服务合同中，平台运营方可以要求技术支持方对数据和基础模型来源的合法合规性进行陈述保证，明确双方的权利义务，避免因技术支持方所提供的基础模型和 / 或数据来源合法性问题影响平台运营方业务的持续开展。

（二）知识产权合规

如我们在本书《ChatGPT 许可应用，知识产权和数据怎么看？》一文中所述，在模型训练的过程中，在数据收集阶段、数据预处理阶段、结果生成阶段分别可能涉及对于数据的复制、翻译、改编、汇编、信息网络传播等受到著作权法等知识产权相关法律法规规制的行为。而在模型的训练数据库涉及未经授权使用他人享有知识产权的数据及内容的情形下，天然具有知识产权侵权风险。以 ChatGPT 为例，ChatGPT 的数据源包括用户输入内容和训练数据库。其中，用户输入内容包括用户使用 ChatGPT 等非 API 服务提供的数据；训练数据库则包括以下三种类型的数据：公有领域内容、通过签订合同获得合法授权的内容、未经授权的信息及内容。倘若技术支持方提供的模型的训练数据库涉及未经授权的信息及内容，在境内平台生成内容与该等信息及内容存在实质性相似的情形下，技术支持方、平台运营方往往并不属于合理使用，从而均有可能承担相应的侵权责任。为降低前述侵权风险，在要求技术支持方确保数据来源合法合规性的同时，我们也建议技术支持方、平台运营方对于生成内容进行一定程度的审核，确保生成内容在表达方面与原始的信息及内容在存在显著区分。

---

[5] 《AIGC 暂行办法》第七条规定，生成式人工智能服务提供者 (以下称提供者) 应当依法开展预训练、优化训练等训练数据处理活动，遵守以下规定：(一) 使用具有合法来源的数据和基础模型；(二) 涉及知识产权的，不得侵害他人依法享有的知识产权；(三) 涉及个人信息的，应当取得个人同意或者符合法律、行政法规规定的其他情形；(四) 采取有效措施提高训练数据质量，增强训练数据的真实性、准确性、客观性、多样性；(五)《中华人民共和国网络安全法》、《中华人民共和国数据安全法》、《中华人民共和国个人信息保护法》等法律、行政法规的其他有关规定和有关主管部门的相关监管要求。

（三）个人信息保护

如前文所述，在部分情形下，用户输入内容将成为技术支持方模型的训练数据。例如，根据 OpenAI 官网的说明，用户通过 ChatGPT、DALL-E 等非 API 服务提供的数据将成为 ChatGPT 的训练数据，除非用户选择关闭训练模式；而用户通过 API 提供的数据将不会作为 ChatGPT 的训练数据，除非用户另行进行授权[6]。而该等用户输入内容中可能包含用户的个人信息。参考全国信息安全标准化技术委员会发布的《信息安全技术 机器学习算法安全评估规范》的要求[7]，我们建议平台运营方针对将用户个人信息用于数据训练取得用户同意 (针对人脸信息等敏感个人信息还应取得用户的单独同意)，并向用户提供不使用个人信息用于数据训练的选项；此外，平台运营方、技术支持方还可以考虑对于收集的个人信息进行必要的匿名化处理，以降低数据训练活动对于用户个人权益的影响。

如技术支持方、平台运营方未按照《AIGC 暂行办法》的规定开展训练数据处理活动，除《个人信息保护法》《中华人民共和国著作权法》等法律法规明确规定的法律责任以外，技术支持方、平台运营方还可能被处以警告、通报批评、责令限期改正、责令暂停提供相关服务等行政处罚。

结语

相较于中国境内的技术支持方，境外的技术支持方在落实相关合规要求的过程中面临更高的不确定性。同时，在涉及境外主体的情况下，技术支持方、平台运营方还可能需要满足其他国家或地区更为严苛的合规要求。因此，我们建议，境内平台与境外的技术支持方开展生成式人工智能服务相关合作时，厘清各方的责任与义务，及时关注所涉国家和地区的监管动态，在合法合规的前提下开展跨境业务合作。

---

[6] How your data is used to improve model performance，链接地址：https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance，最后访问日期：2024 年 3 月 19 日。

[7] 《信息安全技术 机器学习算法安全评估规范》第 5.1.1 条第 d) 项规定，机器学习算法提供者 (即利用机器学习算法实现特定功能的组织，包括算法技术提供者和算法服务提供者，算法技术提供者是指算法技术的开发和提供方，算法服务提供者是指使用应用算法技术的服务提供方) 不应将个人信息用于算法生存周期各项活动 (即机器学习算法从设计到退役的演进过程)，以下情况除外：1) 已按法律法规要求取得个人信息主体同意；2) 法律法规规定无需取得个人信息主体同意。

AI 赋能零售——创新中的合规风险管理

赵新华 王哲峰 游婕

前言

近两年来，人工智能（AI）的快速发展正在众多行业领域掀起一场颠覆性的行业革新。在 2024 年举行的十四届全国人大二次会议上，李强总理在 2024 年政府工作报告中提出，深化大数据、人工智能等研发应用，开展“人工智能 +”行动。这是我国政府工作报告第一次提出“人工智能 +”。“人工智能 +”的概念，意在鼓励各行各业重视人工智能技术的应用和落地。企业需要整合多方资源，在自身产品研发过程中将人工智能技术更好应用，以增强其市场竞争力。

在零售领域，AI 技术已展现出巨大的商业应用价值。从行业实践来看，现阶段售前、售中到售后等环节均已出现了较为成熟的 AI 技术方案，在许多具体业务场景中已得到广泛应用。零售行业参与者们已意识到 AI 技术在降低成本、提升效率、加快创新等方面的重大作用，并纷纷将其作为寻求突破、改变竞争格局的新赛道。

本文拟从“AI 生成营销素材”、“AI 智能导购”和“AI 定价”等零售行业典型的 AI 应用场景出发，探讨伴随 AI 应用产生的相关法律风险。

一、AI 应用场景中的主要角色

目前我国与 AI 直接相关的法规规范主要包括《互联网信息服务算法推荐管理规定》（“《算法推荐规定》”）、《互联网信息服务深度合成管理规定》（“《深度合成规定》”）和《生成式人工智能服务管理暂行办法》（“《生成式 AI 办法》”）。三部法规规范的适用范围和主要义务主体略有差异，简要总结如下：

		《算法推荐规定》	《深度合成规定》	《生成式 AI 办法》
适用范围		在中华人民共和国境内应用算法推荐技术提供互联网信息服务。 应用算法推荐技术，是指利用生成合成类、个性化推送类、排序精选类、检索过滤类、调度决策类等算法技术向用户提供信息。	在中华人民共和国境内应用深度合成技术提供互联网信息服务。 深度合成技术，指利用深度学习、虚拟现实等生成合成类算法制作文本、图像、音频、视频、虚拟场景等网络信息的技术。	利用生成式人工智能技术向中华人民共和国境内公众提供生成文本、图片、音频、视频等内容的服务。 生成式人工智能技术，是指具有文本、图片、音频、视频等内容生成能力的模型及相关技术。
主要义务主体	服务提供者	√	√	√
	技术支持者	×	√	√ ** 本办法未专门规定“技术支持者”概念，但规定“服务提供者”包括通过提供可编程接口（API）等方式提供生成式AI服务的组织、个人，即通常所说的“技术支持者”。*
	服务使用者	×	√	√

在常见的零售行业参与者中，大型零售平台往往自行开发 AI 产品为消费者和平台内商家提供服务，其角色更偏向于 AI 服务提供者和 / 或 AI 技术支持者。相较而言，一般品牌方通常选择从其他 AI 技术开发者处直接采购 AI 工具搭载于自身产品和服务上再提供给消费者，或采购 To B 型 AI 工具用于自用目的，因此品牌方作为 AI 服务提供者和 / 或 AI 服务使用者的情况较为多见。

二、AI 生成营销素材

一个天才创意能够为营销活动积累巨大的口碑，极大地提升产品销量和品牌影响力，而生成式 AI 凭借其高效的内容产出能力和源源不断的创意能力，能够显著降低创意成本、

提高营销效率。

目前市场上已有较为成熟的生成式 AI 工具，可根据用户输入的提示文本撰写广告文案或生成广告图片，往往在几秒钟之内就可提供十几种方案供选，还可根据使用者的进一步要求进行调整优化。某些专注于营销领域的生成式 AI 工具还可根据不同行业领域和应用场景提供多种素材模板，甚至能够整合其他广告监测工具来追踪这些素材的广告投放效果。

根据 2024 年 4 月 2 日国家互联网信息办公室《关于发布生成式人工智能服务已备案信息的公告》以及此前陆续发布的深度合成服务算法备案清单，国内部分大型电商平台已进行了 AI 大模型的研发并完成了相关备案。这些 AI 大模型可提供商品营销文案创作、商品图片和卖点等营销素材生成等功能，帮助商家提高运营效率和营销内容质量。不少品牌方 / 零售商们已不同程度地使用过这些 AI 创作工具，有效降低了营销物料制作周期。

由于在 AI 应用中的角色不同，零售行业参与者所承担的合规义务也有所区别：

- **AI 服务提供者**

  AI 服务提供者除了自身不得实施非法使用个人信息等侵犯个人信息主体权益的行为外，也应提示用户注意输入内容中若包含未公开的个人信息，应取得相关个人信息主体的同意。在技术层面，服务提供者也应当采用技术或者人工方式对用户的输入数据和合成结果进行审核[1]，确认是否可能存在个人信息，并过滤或对相关个人信息进行脱敏处理。如接到第三方关于个人信息侵权的举报，应及时处理。

  AI 服务提供者如果自行开发 AI 工具的，还应当履行以下 AI 技术支持者的相关合规义务。

- **AI 技术支持者**

  AI 工具的训练和使用依赖于对大量数据的处理，因此对于 AI 技术支持者而言，确保数据来源合法是进行后续开发的首要合规前提。AI 模型训练的数据来源通常包括以下三种：

  （1）使用网络爬虫获取公开数据：需注意网络爬虫不应故意突破或绕开网站设置的访问限制措施和反爬虫技术措施；

  （2）直接收集个人信息：如从用户、消费者或测试志愿者等个人处收集个人信息用于模型训练，需确保已向相应个人信息主体告知其个人信息将被用于 AI 模型开发等目的并获取其授权同意，或在能够满足训练目的的前提下，对训练数据中的个人信息进行匿名化处理；

[1] 《互联网信息服务深度合成管理规定》第 10 条。

（3）从第三方采购数据集：需要求该等第三方就其数据来源合法性做出陈述与保证。

此外，与 AI 技术支持者的数据处理活动相关的主要义务包括：（1）依法开展训练数据处理活动，加强训练数据管理，采取必要措施保障训练数据安全[2]；（2）提供人脸、人声等生物识别信息编辑功能的，应当提示服务使用者依法告知被编辑的个人，并取得其单独同意[3]；（3）提供具有生成或者编辑人脸、人声等生物识别信息功能的模型、模板等工具的，应当依法自行或者委托专业机构开展安全评估[4]；（4）在生成式 AI 技术研发过程中进行数据标注的，应当制定清晰、具体、可操作的标注规则，开展数据标注质量评估，抽样核验标注内容的准确性，并对标注人员进行必要培训，监督指导标注人员规范开展标注工作[5]。

- **AI 服务使用者**

品牌方等 AI 服务使用者在使用营销素材生成、虚拟主播等生成式 AI 工具时可能也需要输入人脸、声音等个人信息，为此同样需要向相关个人履行告知—同意义务，特别是包括对处理其人脸、声音等敏感个人信息的单独同意。

三、AI 智能导购

AI 生成营销素材主要是服务于商家的 To B 模式，而 AI 智能导购则是直面消费者的 To C 模式。目前国内部分大型电商平台已推出了基于大语言模型的 AI 智能导购工具，以期解决传统人工导购 / 客服培训周期长、回复效率低、回复结果不理想等问题，帮助平台用户更快速准确地找到符合其偏好的商品。

AI 智能导购利用多轮对话不断挖掘用户需求，归纳总结用户的购物意图，圈定用户需求特征后检索相关商品，最后按照用户偏好排序并向用户进行推荐。以购买家用冰箱的场景为例，用户提问时一般仅会给出少量表层需求，例如意向品牌、意向价格区间等，智能导购会进一步对其他需求特征进行询问，例如开门方式、尺寸、容量、制冷模式、特殊功能要求等。当明确用户需求后，智能导购会在平台内检索相关商品，计算用户需求和商品特征的匹配度来筛选出符合条件的商品，并根据用户对话反馈不断调整推荐商品，最终实现交易转化。

技术层面上，AI 智能导购的本质是基于深度学习算法，根据用户输入的信息进行自动化决策并输出决策结果的一类生成式 AI 工具。同时，因 AI 智能导购可能涉及生成合成类、

[2] 《互联网信息服务深度合成管理规定》第 14 条。
[3] 《互联网信息服务深度合成管理规定》第 14 条。
[4] 《互联网信息服务深度合成管理规定》第 15 条。
[5] 《生成式人工智能服务管理暂行办法》第 8 条。

个性化推送类、排序精选类等算法技术，属于《算法推荐规定》定义的“算法推荐技术”，其服务提供者和技术支持者除了需要遵守生成式 AI 服务相关法规以外，还需要注意算法推荐管理方面的合规义务：

- **算法备案义务**

  《算法推荐规定》规定，具有舆论属性或者社会动员能力的算法推荐服务提供者应当在提供服务之日起十个工作日内通过互联网信息服务算法备案系统填报服务提供者的名称、服务形式、应用领域、算法类型、算法自评估报告、拟公示内容等信息，履行备案手续[6]。《算法推荐规定》并未进一步明确“具有舆论属性或者社会动员能力”的具体标准。2018 年国家互联网信息办公室和公安部联合发布的《具有舆论属性或社会动员能力的互联网信息服务安全评估规定》（“《2018 年安全评估规定》”）提供了相关解释：“本规定所称具有舆论属性或社会动员能力的互联网信息服务，包括下列情形：

  （一）开办论坛、博客、微博客、聊天室、通讯群组、公众账号、短视频、网络直播、信息分享、小程序等信息服务或者附设相应功能；

  （二）开办提供公众舆论表达渠道或者具有发动社会公众从事特定活动能力的其他互联网信息服务。”

  此外，从网信部门公布的算法备案信息中也可看出，已备案算法的应用产品类别非常广泛，除新闻资讯、搜索引擎、贴吧论坛、网络视频等应用以外，还包括求职招聘、外卖配送、网络购物类应用，几乎覆盖了绝大多数的互联网常见应用类型。因此实践中，算法推荐服务提供者应充分理解上述标准，对于包含评论互动和信息分享功能且面向社会公众提供的算法推荐服务，宜主动履行算法备案义务，避免受到监管处罚。

- **安全评估义务**

  具有舆论属性或者社会动员能力的算法推荐服务提供者应当按照国家有关规定开展安全评估，还应当依法留存网络日志，配合网信部门和电信、公安、市场监管等有关部门开展安全评估和监督检查工作，并提供必要的技术、数据等支持和协助[7]。《2018 年安全评估规定》规定了应当进行安全评估的五类情形，包括“具有舆论属性或社会动员能力的信息服务上线，或者信息服务增设相关功能的；用户规模显著增加，导致信息服务的舆论属性或者社会动员能力发生重大变化的”等。

---

[6] 《互联网信息服务算法推荐管理规定》第 24 条。
[7] 《互联网信息服务算法推荐管理规定》第 27、28 条。

- **其他主要合规义务**

  告知公示：AI 导购服务提供者应当在相关功能界面或相关网站 / 应用程序的其他显著位置向用户告知算法推荐服务的情况，并以适当方式公示算法推荐服务的基本原理、目的意图和主要运行机制等。

  不得操纵：AI 导购服务在向用户推荐商品或商品榜单时，应注意不得过度推荐或操纵榜单。因 AI 导购功能可直接促成交易，品牌方可能会向这一渠道进行更多广告投放，获取更多曝光量，但 AI 导购服务提供者不应向投放过广告的商品进行过分的资源倾斜，不得利用算法屏蔽其他商品信息或干预其他商品信息的呈现。

  算法机制和内容审核：AI 导购服务提供者应当定期审核、评估、验证算法机制机理、模型、数据和应用结果等，避免出现算法操纵、诱导过度消费和其他违反法律法规或者违背伦理道德的情形。此外，AI 导购服务提供者在不断收集更多数据、优化相关算法模型时，还应注意识别和过滤语料库中的违法内容和不良信息，避免生成违法和不良信息，例如在面向用户生成商品推荐文案时，应避免生成违反广告法的用语或内容。

  保障用户权利：AI 导购服务提供者应当始终保障用户权利，在向用户推荐商品时应当提供不针对其个人特征的选项，或者提供选择或删除用于 AI 导购服务的针对其个人特征的用户标签的功能。

四、AI 定价

商品定价策略对商品销量和企业营收至关重要。传统的定价策略通常是在商品成本上增加一定的利润率，同时结合竞品价格、市场需求和趋势等外部因素综合确定价格，而这往往依赖人工判断和经验总结。在 AI 赋能零售的技术浪潮下，商品定价也可以由 AI 大模型完成。

AI 大模型定价简单来说可分为三个环节：（1）收集和处理数据；（2）建立定价模型、确定定价策略的目标；（3）实施定价策略。基于底层数据的更新、不同期间定价目标的差异，AI 生成的定价策略也将随着市场变化和用户行为而动态变化。比如，在商品销量较低或用户反馈价格过高时，AI 可能会自动降价；在竞品价格普遍上涨时，AI 可能会自动跟随涨价。又比如，AI 也可能根据某一用户的行为进行个性化定价：当 AI 通过消费者的历史交易数据识别出消费者对某品牌有更高的忠诚度时，可能会向此类消费者提供更为优惠的价格。然而在我国目前的监管体系下，“千人千面”的个性化定价存在着较为明显的合规风险。

通过 AI 进行个性化定价的本质是通过自动化决策技术，就同一商品或服务对交易条件相同的个人提供不同价格，因而可能构成差别待遇。《个人信息保护法》明确规定，利用个人信息进行自动化决策，应当保证决策的透明度和结果公平、公正，不得对个人在交易价格等交易条件上实行不合理的差别待遇[8]。另外，利用消费者的个人信息进行自动化决策也应当根据《个人信息保护法》的要求向消费者告知个人信息将被用于进行自动化决策目的，并获取其同意，除非个人信息已经过匿名化处理不再构成个人信息。

个性化定价还可能涉及反垄断方面的合规风险。《反垄断法》明确禁止具有市场支配地位的经营者没有正当理由，对条件相同的交易相对人在交易价格等交易条件上实行差别待遇[9]。《国务院反垄断委员会关于平台经济领域的反垄断指南》对此作出了细化规定：（1）分析是否构成差别待遇，可以考虑以下因素：基于大数据和算法，根据交易相对人的支付能力、消费偏好、使用习惯等，实行差异性交易价格或者其他交易条件；（2）平台在交易中获取的交易相对人的隐私信息、交易历史、个体偏好、消费习惯等方面存在的差异不影响认定交易相对人条件相同；（3）针对新用户在合理期限内开展的优惠活动以及能够证明行为具有正当性的其他理由，可以被视为实施差别待遇行为的正当理由[10]。以上禁止差别待遇的规定针对的是具有市场支配地位的经营者，相较于一般品牌方，大型零售平台在此方面可能面临着更高的合规风险。

另外，消费者知情权也是一个重要的关注方面。2024 年 3 月 15 日发布的《中华人民共和国消费者权益保护法实施条例》明确规定，经营者不得在消费者不知情的情况下，对同一商品或者服务在同等交易条件下设置不同的价格或者收费标准[11]。因此，经营者如具有正当理由使用个性化定价方式，例如针对新客户提供优惠价格，则应当明确告知消费者，保障其知情权和选择权。

结语

“AI+ 零售”的行业前景广阔，除了本文举例的应用场景外，还有很多正在创新的应用场景，其发展空间和巨大潜力将进一步释放。品牌方和零售商们在积极使用 AI 赋能零售的同时，还应密切关注这一新兴技术所带来的法律风险，采取合理防范措施，在智能零售转型之路上行稳致远。

感谢实习生孟泽萱对本文作出的贡献。

[8] 《个人信息保护法》第 24 条。
[9] 《反垄断法》第 22 条第（六）项。
[10] 《国务院反垄断委员会关于平台经济领域的反垄断指南》第 17 条。
[11] 《中华人民共和国消费者权益保护法实施条例》第 9 条。

境外的 AI 发展

天下事预则立，不预则废——香港私隐公署开展人工智能合规检查，明确 AI 发展指引和提升产业信心

宁宣风 吴涵 方禹

引言

2024 年 2 月 21 日，香港私隐公署完成了对 28 家机构的人工智能（AI）合规检查。该检查自 2023 年 8 月启动，历时半年左右，主要对相关机构开发或使用人工智能系统时，收集、使用或者处理个人信息（香港称为“个人资料”，本文不做区分）的风险和影响进行检查和评估。这是人工智能合规实践的一次官方行动，并给出了整体结论和合规建议。

我们理解，行政行动是企业合规的外源性动力和重要合规价值之一。香港私隐公署此次对人工智能的合规检查行动，无疑对于人工智能健康发展，以及如何在人工智能环境下合理使用数据具有典型的指导借鉴意义。

一、检查对象

本次检查涉及多行业、多性质主体，包括电信、金融、保险、美容服务、零售、交通、教育等行业主体以及政府部门。从行业来看，本次检查以服务业为主，属于个人信息处理的密集型行业，大多是“互联网 +”较早应用且发展势头强劲的领域，未来也是生成式人工智能向产业落地的重点发力领域。从性质来看，这 28 家机构既包括私营部门，也包括政府部门，这也延续了香港《个人私隐条例》的监管思路：香港《个人私隐条例》对政府也具有约束力。然而，香港私隐公署并没有披露这 28 家机构的具体名单。

二、香港私隐公署的总体意见

香港私隐公署在本次合规检查中，并未发现违反《个人私隐条例》的行为。这表明大多数企业在部署 AI 系统以提高业务效率的同时，暂不存在显著侵害个人隐私的行为。香港私隐公署简要公布了相关检查意见，具体如下：

（1）该 28 家机构中有 21 家应用了 AI 系统，占比 75%。

（2）应用 AI 系统的 21 家机构中，有 19 家机构建立了内部 AI 治理框架，占比约

90%，而在所有被检查的 28 家机构中，占比约 68%。具体而言，已建立的内部 AI 治理框架包括成立 AI 治理委员会、指定专员监督 AI 产品、服务的开发和应用。

（3）应用 AI 系统的 21 家机构中，只有 10 家机构通过 AI 产品、服务收集个人信息，占比不到 50%。而此 10 家机构收集个人信息时，均向用户告知了收集情况，明确了使用目的及共享的第三方。

（4）通过 AI 产品、服务收集个人信息的 10 家机构中，有 8 家开展了影响评估，占比 80%。

（5）通过 AI 产品、服务收集个人信息的 10 家机构，全部采取了适当的安全措施，占比 100%。相关安全措施包括：1）仅允许获授权的人员查阅个人资料；2）对存储、传输的个人资料进行加密；3）进行定期的安全风险评估和渗透测试；4）为雇员提供书面指引和培训。通过这些保障措施，能够保障资料使用者所持有的个人资料在 AI 产品、服务的开发、应用中受到保护，减少在未获授权或意外情况下被查阅、处理、删除、丢失或者使用。

（6）通过 AI 产品、服务收集个人信息的 10 家机构中，有 9 家会存储其通过 AI 产品、服务收集的个人资料，其中有 8 家明确了存储期限，而剩下 1 家允许用户自行删除其个人资料。

从检查公布数据来看，相关机构的 AI 系统部署程度较高，体现了 AI 应用具有通用性、普及性的趋势。但是，通过 AI 系统收集个人信息的比例不高。而相关机构通过 AI 系统收集个人信息时，具有较强的数据保护意识，采取了影响评估、安全保障等相关措施，也对个人信息存储、删除等权益予以了保障。香港私隐公署也表示通过本次检查，发现 AI 部署情况良好，个人信息保护情况也令人满意。

三、检查程序

根据《香港个人私隐资料条例》，香港私隐公署可主动发起调查或根据投诉举报启动调查程序。本次调查属于私隐公署主动发起的行动，根据条例规定，香港私隐公署可以对资料使用者所使用的任何个人资料系统进行监督检查，并向有关资料使用者作出遵守该条例的建议。不过，香港私隐公署并没有公布检查的具体措施、对象、程序等细节。

四、香港私隐公署的结论及建议

香港个人资料私隐专员钟丽玲表示：“AI 在推动生产力和经济增长中有巨大潜力，但也会造成不同程度的个人资料私隐和道德风险。我很高兴地知道，在所调查的机构中，

大多数都建立了内部 AI 治理框架，监督 AI 产品、服务的开发和应用。资料使用者在开发、使用 AI 系统时，负有数据安全保障义务，他们应当及时审查和评估 AI 系统对个人资料私隐的风险。”

针对本次检查，香港私隐公署对开发、使用 AI 系统的机构发出如下合规建议：

（1）有关机构在开发、使用 AI 时收集或处理个人资料，应采取措施确保遵守《个人资料（私隐）条例》，并持续对 AI 系统进行监测。

（2）制定开发、使用 AI 的战略和内部 AI 治理结构，并为所有相关人员提供充分的培训。

（3）进行全面的风险评估（包括私隐影响评估），以便在 AI 开发、使用中系统地识别、分析及评估风险，并采取与风险相称的管理措施，如对风险较高的人工智能系统采取更高水平的人工监测措施。

（4）与利益相关方进行有效沟通，以提高人工智能使用的透明度，并根据利益相关方提出的关切，对人工智能系统进行调整。

此外，钟丽玲专员还提醒各机构在开发、应用 AI 产品、服务时，应当遵守香港私隐公署发布的《开发及使用人工智能道德标准指引》。

五、香港私隐公署《开发及使用人工智能道德标准指引》

不同地区的人工智能发展速度和普及程度各有不同，而人工智能技术对不同行业和界别的影响亦不尽相同。不同地区和机构都在适应人工智能的最新发展，研究不同措施，以应对人工智能所带来的影响和挑战，并在促进科技创新及保障规范之间作出平衡。基于此，香港私隐公署于 2021 年发布了《开发及使用人工智能道德标准指引》，以协助机构在开发、使用人工智能时，理解并遵从《个人资料（私隐）条例》有关保障个人资料私隐的相关规定。

《开发及使用人工智能道德标准指引》的内容包括数据管理价值及人工智能道德原则，并提供人工智能治理策略的实务指引，帮助机构制订合适的人工智能策略及管理模式，并作出风险评估及制定相关监督保障措施等。

《开发及使用人工智能道德标准指引》确定了“374”的 AI 道德框架，即 3 个数据管理价值观，7 个 AI 道德原则和 4 个主要业务流程。

（1）数据管理价值观。价值观决定了机构如何采取行动以实现保护个人资料私隐的目标。具体而言，指引确定了以下 3 个数据管理价值观：

- 尊重：尊重人的尊严、自主权、权利、利益和尊严，以及个人对其数据被处理的

合理期待，是至关重要的。因此，每个人都应该被道德地对待，而不是被当作一个物体或一段数据。

- 福利：应当向利益相关者提供福利，利益相关者包括受人工智能使用影响的个人，以及更广泛意义上的社会整体。同时，不应伤害任何利益相关者，或者应使伤害最小化。
- 公平：公平应当覆盖过程和结果。就程序而言，“公平”意味着做出合理的决定，不应有不公正的偏见或非法的歧视。应为个人建立无障碍和有效的途径，以获得受到不公平待遇的赔偿。就结果而言，做到“公平”意味着相似的人应该被同等对待。对不同的个人或不同的群体采取差别待遇，应当有合理的理由。

（2）AI 道德原则。根据数据管理价值观，机构可结合本身组织文化，确定相应的原则。指引鼓励机构遵守以下 7 项 AI 道德原则：

- 有责任：机构应对其行为负责，能够为其行为提供合理的理由。同时，应当在高级管理人员的参与下，通过跨部门协作，评估和解决人工智能的风险。
- 人工监督：AI 系统的使用者应该能够根据 AI 的建议或决策，作出有根据的、自主的选择。人工参与程度应该与使用人工智能系统的风险和影响相称。高风险 AI 系统中应当始终有人工干预。
- 透明且可解释：机构应明确、显著地告知其使用 AI 的情况，以及相关隐私保护实践，以提升 AI 自动决策和辅助决策的可解释性。透明且可解释，是承担 AI 责任的工具，也是保护个人权利、自由和利益的手段。
- 隐私保护：隐私是一项基本人权。开发和使用人工智能过程中，应该保护个人隐私，以实现有效的数据治理。相关个人资料处理活动应当遵守《个人资料私隐条例》。
- 公平：个体应被合理平等地对待，而不应有不公正的偏见或者非法的歧视。对不同的个人或不同的群体采取差别待遇，应当有合理的理由。
- AI 福利：AI 应对人类、商业和社会有好处。不产生伤害也是一种福利。使用 AI 时不应对利益相关者造成损害，或者应将损害最小化。
- 可靠、鲁棒和安全：机构应确保 AI 系统在预期寿命内可靠运行。AI 系统应在运行过程中应有容错能力，以防止有关伤害或将伤害最小化。同时，还应当防止对 AI 系统的攻击，如黑客攻击和数据中毒等。对此，应制定应急计划，以应对人工智能系统无法正常运行的情况。

（3）业务流程

- AI 治理策略：AI 系统需要大量使用个人数据，机构应当根据其 AI 部署程度和水平，调整其内部治理结构，并组建 AI 治理委员会或者类似部门。同时，应在 AI 应用的全生命周期确定或调整隐私保护政策及数据保护措施。
- 风险评估和人工干预：AI 的使用目的和方式决定了 AI 系统的风险高低，机构应当系统地识别、分析和评估 AI 风险。对于高风险 AI，应当建立全生命周期的风险应对机制。风险评估的目标是采取相应措施以降低风险，其中人工干预的措施十分有必要。人工干预是降低 AI 风险的关键措施之一。任何时候，人都应当最终对 AI 的决策负责。
- 开发 AI 模型和管理 AI 系统：AI 训练所使用的数据对于 AI 模型的精确性、可靠性具有十分重要的影响。开发 AI 模型时，应当经过 6 个步骤：a. 收集数据；b. 准备数据；c. 选择机器训练模型和算法；d. 通过数据分析训练 AI 模型；e. 测试、评估和调整 AI 模型；f. 将 AI 模型投入使用。而管理 AI 系统的关键是保证可靠性、鲁棒性和安全性，同时需要人工干预。具体而言，机构应当做好相关书面记录，及时针对新风险进行再评估，定期检查 AI 模型并使用新数据调整、再训练 AI 模型。
- 与利益相关者沟通协作：沟通协作的关键是保证透明性、可解释性。对于利益相关者而言，AI 系统应当具有透明性。机构应当向利益相关者清晰、显著地告知 AI 系统的使用情况，包括目的、好处、限制和影响等。对于可能对个人造成重大影响的 AI 系统，机构还应当建立相关机制，允许个人进行修正、提供反馈、寻求解释或者要求人工解释，以及选择退出 AI 系统。

值得注意的是，香港政府资讯科技总监办公室参考《开发及使用人工智能道德标准指引》后，也于 2023 年 8 月发布了《人工智能道德框架》，以对人工智能和数据所涉及的道德标准进行规范，促进科技发展的同时避免产生不良影响。

六、一些延展信息

2021 年，香港私隐公署发布《开发及使用人工智能道德标准指引》的同一天，还发布了一份消费者个人信息保护的调查报告。根据这份报告，私隐专员对公共设施服务企业的个人信息保护提出了一些建议。人工智能向通用性发展的过程中，也有向一般基础设施演变的可能性。因此，这些建议可能也具备人工智能合规的参考价值。具体如下：

（1）对未知的个人数据隐私风险有所准备。

（2）实施个人数据隐私保护治理方案。

（3）任命数据保护官（DPO）。

（4）保存个人数据清单。

（5）设计系统安全策略和流程。

（6）设置不同的内部数据访问权限。

（7）采取预防监控措施。

（8）同时保护电子数据和纸质数据。

（9）采取措施提升员工保护意识。

结语—— 对企业的合规启示

人工智能技术，特别是生成式人工智能技术的发展还在不断演进。相信越来越多的行业将开始应用人工智能技术，以提升生产经营效率，增加业务产出，改善现有的业务流程和商业模式。实现更大的利益增长空间。人工智能技术使用大量数据，与数据合规密不可分，同时也有很多人工智能合规的新问题，这些问题构成了人工智能应用的风险。实践中，风险评估 / 影响评估等是应对人工智能风险的主要和常见手段。

香港私隐公署此次的人工智能合规检查，从官方层面明确了人工智能合规的重点方向，同时也给出了相关合规建议。内地《个人信息保护法》《互联网信息服务算法推荐管理规定》《生成式人工智能服务管理暂行办法》等也构建了相关风险评估 / 影响评估机制。事实上，面对人工智能发展浪潮，如何确保安全与发展平衡的原则，是各方都在讨论的议题。在发展人工智能的过程中，需要有效应对人工智能风险，也基本形成了各方的共识。

企业在人工智能深入发展的趋势之中，既要把握技术产业发展先机，充分释放人工智能潜力和优势。同时，在治理策略、业务流程、风险识别和应对方面，也应提前布局，积极采取相关行动，构建相关机制，做好保障措施。

从美国商务部云计算管控新规看美国 AI 监管新趋势

楼仙英 戴梦皓 姚爽 郑子懿 张欣悦 曹逸晨 张嘉柔

2024 年 1 月 29 日，美国商务部产业安全局（“BIS”）发布了关于《“采取额外措施应对重大恶意网络活动方面的国家紧急情况”（Taking Additional Steps To Address the National Emergency With Respect to Significant Malicious Cyber-Enabled Activities）拟议规则的通知》（notice of proposed rulemaking）[1]，以加强相关美国云计算服务提供商对外国人客户的身份识别和报告义务。该拟议规则的公众征求意见期截至 2024 年 4 月 29 日。根据拟议规则，提供美国“基础设施即服务”（Infrastructure as a Service，“IaaS”）产品的美国 IaaS 提供商、美国经销商及相关外国经销商须核验外国人客户身份信息，并在特定情形下向美国商务部报告外国人客户的详细身份信息和人工智能（“AI”）大模型训练活动情况。

本次拟议规则的发布正值 BIS 主管信息和通信技术和服务交易（information and communications technology and services，“ICTS”）审查的信息和通信技术服务办公室（“OICTS”）首位执行主任 Elizabeth Cannon 入职之际，令人不禁联想 BIS 在 ICTS 审查领域的执法是否已经箭在弦上。而同时鉴于本次新规所针对的云服务和 AI 相关话题与先前 BIS 在出口管制领域关于先进计算芯片的相关管制措施紧密相连，同样也让人关注 BIS 在出口管制后续执法上是否会有进一步联动措施。在此，我们结合本次拟议规则的相关内容以及美国相关机构和国会的近期表态，为大家带来关于云服务和 AI 领域相关管控的趋势分享和提示，希望对大家有所帮助。

一、本次拟议规则的主要背景和法律依据

早在 2021 年 1 月 19 日，特朗普政府发布了第 13984 号总统行政令（“E.O.

[1] 参见：https://www.federalregister.gov/documents/2024/01/29/2024-01580/taking-additional-steps-to-address-the-national-emergency-with-respect-to-significant-malicious。

13984”）[2]，授权商务部制定法规，要求美国 IaaS 提供商验证其外国人客户的身份，以及采取特别措施阻止外国恶意网络行为者使用美国的 IaaS 产品。美国商务部随后曾于 2021 年 9 月 4 日发布了相关的预约立法通知（Advance notice of proposed rulemaking，“ANPRM”），向公众征求关于如何执行 E.O.13984 第 1 节、第 2 节规定及第 5 节关键词定义的相关意见[3]。就此 ANPRM，商务部收到了 21 条公众评论意见（“ANPRM 评论意见”），而本次拟议规则的拟议条款文本也融合了诸多前述评论意见。

在此之后，拜登政府于 2023 年 10 月 30 日发布第 14110 号总统行政令（“E.O. 14110”）[4]。该法规第 4.2 节 c 项规定，商务部应在该行政令发布后的 90 天内制定法规，要求 IaaS 产品提供商在外国人与其进行交易以训练具有潜在能力可用于恶意网络活动的 AI 大模型时，向商务部进行报告。

参考 E.O.13984 的相关 ANPRM 评论意见，BIS 于 2024 年 1 月 29 日发布了本次拟议规则，以执行 E.O. 13984 的第 1、2 和 5 节内容以及 E.O. 14110 中可适用的相关条款，并向公众征求意见。

二、本次拟议规则的主要内容

（一）关于相关重要概念的定义

结合 E.O. 13984 第 5 节及 E.O. 14110 第 3 节的定义，并经 BIS 进行进一步澄清和确认后，拟议规则中的重要概念整理如下，供各方参考：

主要概念	相关定义
IaaS 产品	指向消费者提供的处理、存储、网络或其他基本计算资源的任何产品或服务，包括消费者不管理或控制底层硬件，而是与第三方签订协议以访问硬件的所有服务产品例如内容交付网络、代理服务和域名解析等服务，但不包括域名注册服务。
用于恶意网络活动的潜在能力的 AI 大模型	指具有军民两用基础模型技术条件的任何 AI 模型，或老具有其他技术参数、可用于帮助或促使恶意网络活动自动化的 AI 模型。

[2] 参见：https://www.federalregister.gov/documents/2021/01/25/2021-01714/taking-additional-steps-to-address-the-national-emergency-with-respect-to-significant-malicious。
[3] 参见：https://www.federalregister.gov/documents/2021/09/24/2021-20430/taking-additional-steps-to-address-the-national-emergency-with-respect-to-significant-malicious。
[4] 参见：https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence。

主要概念	相关定义
美国人	指任何美国公民、合法永久居民、基于美国法律或美国境内法域所设立的实体，包括该等实体的境外分支机构以及任何位于美国境内的人士。
美国 IaaS 提供商	指提供任何 IaaS 产品的美国人 (包括个人和实体)，包括美国 IaaS 供应商和经销商。我们理解云服务安全拟议规则的规制对象不限于主流的云服务厂商，如若任何美国主体直接或间接对外提供 IaaS 产品，都会受到该规则约束 值得关注的是，作为对部分 ANPRM 评论意见的回应商务部在拟议规则中表明，其计划说明美国 IaaS 提供商的外国子公司不被视作“美国 IaaS 提供商”。
外国人	指非美国人士，即“美国人”之外的其他公民或实体。
受益所有人	商务部要求美国 IaaS 提供商收集并验证客户的“受益所有人”信息，判断标准包括 :(a) 对客户行使实质性控制或者 (b) 拥有或控制客户至少 25% 的所有权利益者。
外国人的身份信息	包括客户的姓名、地址、主要付款方式和来源、电子邮件地址和电话号码，以及用于访问或管理账户的互联网协议 (IP) 地址。
客户身份识别计划 (“CIP”)	指由美国 IaaS 提供商或外国经销商设立的计划，该计划规定美国 IaaS 提供商将如何收集其客户的身份识别信息如何验证其外国人客户的身份、如何存储和维护身份识别信息，以及如何通知其客户有关身份识别信息的披露。

（二）关于建立 CIP 的义务和相关具体要求

根据本次拟议规则，美国 IaaS 产品的：i）美国 IaaS 提供商（包括美国供应商及美国经销商），以及 ii）其外国经销商须建立 CIP，以执行有效的客户验证机制并维护其外国人客户的身份信息，并且这些 CIP 需要向美国商务部进行报告。

考虑到避免给 IaaS 提供商带来不必要的负担，商务部目前允许每个提供商在 CIP 符合某些最低要求的前提下创建与其服务产品和客户群体相匹配的 CIP。

1. 关于 CIP 的主要要求

具体而言，相关主要要求如下：

主要要求	具体内容
数据采集要求	根据拟议规则，其至少包括收集如下信息：**客户的名称、地址、每个客户账户的支付方式和来源、电子邮件地址和电话号码，以及用于访问或管理账户的 IP 地址**。laaS 提供商可更改其 CIP，要求潜在客户提供其他必要信息，以核实任何外国人的身份，但所有 CIP 必须至少收集前述信息。
客户身份的验证要求	美国商务部建议要求 **CIP 包括确保美国 laaS 提供商及其外国经销商核实所有外国账户所有者和外国受益所有者身份的程序**。根据拟议规则，提供商可自行制定程序和方法来验证其潜在外国人客户和受益所有人的身份，但 CIP 必须包括基于风险的程序，使提供商能够合理地相信每个客户和受益所有人的真实身份。这些程序必须基于提供商对相关风险的评估，包括提供商提供的各类服务所带来的风险、开立账户的方法、提供商可获得的不同类型的身份识别信息以及提供商的客户群。 根据拟议规则，CIP 必须包括提供商在无法核实任何客户身份时将采取的步骤，包括拒绝开立账户和 / 或在尝试核实前进行额外监控。此外，其必须进一步规定在提供商试图核实客户身份期间客户可以继续使用账户的条款，以及提供商在试图核实客户身份失败后何时关闭账户。进一步，CIP 还必须说明补救措施和问题处理措施，以解决合法客户可能无法通过身份验证的情况，或客户信息被泄露并建立欺诈性账户的情况。
记录保存要求	拟议规则目前要求美国 laaS 提供商和外国经销商建立维护、保护和获取在验证客户身份过程中获取的相关客户信息记录的程序。该记录至少必须包括客户首次尝试开立账户时提供的身份证明的描述、为验证客户身份而采取的任何措施的方法和结果的描述，以及在验证身份信息时发现的任何实质性差异的解决方法的描述。拟议规则允许 laaS 提供商自行设计记录保存程序，只要这些程序能够获得前述的基本信息。美国商务部要求美国 laaS 提供商及其外国经销商的 CIP 包括安全保存这些记录的要求，并说明为确保信息安全而采取的措施。**而对于保存期限，美国 laaS 提供商需要在账户最后一次被访问或关闭之日起两年内保留这些记求。**
确保外国经销商的验证要求	根据 E.O.14110，**美国商务部建议要求美国 laaS 提供商只能与维护和实施符合拟议规则中的 CIP 的 laaS 产品外国经销商建立或继续经销关系**。对此，美国商务部认为，美国 laaS 提供商需要时间来教育、协调和收集外国经销商有关 CIP 要求的信息，因此预计将允许美国 laaS 提供商在一年内实施最终规定并通知商务部。 根据拟议规则，美国 laaS 提供商必须在收到商务部的要求后十日内，向该部提供任何外国经销商的 CIP 副本此外，在收到表明外国经销商未能维护或实施 CIP 或表明存在恶意网络活动的证据时，美国 laaS 提供商必须报告恶意网络活动，关闭与该活动相关的账户，并必须在 30 日内终止经销商关系。

主要要求	具体内容
CIP 的更新和认证	本次拟议规则要求美国 IaaS 提供商**向美国商务部提交有关其 CIP 及其外国经销商 CIP 的相关信息**，包括验证客户身份和检测恶意网络活动的程序，以及有关其提供 IaaS 产品的信息和数据。 此外，**美国 IaaS 提供商及其外国经销商还被要求每年更新其 CIP**，以防范新的网络威胁和漏洞，提高效率和数据安全性，并向商务部证明已进行了此类年度更新。美国 IaaS 提供商也必须将其 CIP 或其外国经销商的任何 CIP 的任何更新通知商务部。 在这些年度认证中，美国 IaaS 提供商还应向该部证明自上次认证日期以来，其已审查 CIP 并更新其 CIP，以说明其服务产品的任何变化以及威胁环境的变化。认证将包括一份证明，证明当前的 CIP 符合拟议规则的规定。

2. 关于 CIP 义务的豁免

E.O. 13984 第 1 节 c 项允许商务部长在与国防部长、司法部长、国土安全部长和国家情报局局长协商后，根据商务部长认为适当的因素，豁免任何美国 IaaS 提供商或任何特定类型的账户或承租人遵守拟议规则下的规定要求。此外，E.O. 14110 第 4.2 节 d 项还规定，商务部长可以“免除美国 IaaS 提供商对其美国 IaaS 产品的任何特定外国经销商或任何特定类型的账户或承租人遵守根据本条款发布的任何规定的要求”。而根据拟议规则，美国商务部建议豁免美国 IaaS 提供商及其任何外国经销商在 CIP 下的标准和程序义务，并建议寻求豁免的提供商以电子方式提交书面申请。

（三）关于 AI 训练情况的报告要求

E.O. 14110 第 4.2 节 c 项指示商务部长提出法规，要求美国 IaaS 提供商在外国人与该美国 IaaS 提供商进行交易以训练具有可能用于恶意网络活动的潜在能力的大型人工智能模型时，向商务部长提交报告。此类报告应至少包括外国人的身份、符合 E.O. 14110 的标准 [5] 或商务部长确定的标准的人工智能模型的任何训练运行情况，以及商务部长确定的其他信息。此外，该条文亦规定，除非外国经销商向提供商提交此类报告且提供商将这些报告提供给商务部长，美国 IaaS 提供商禁止向其外国经销商提供美国 IaaS 产品。

对此，本次的拟议规则也将要求此类提供商在相关模型具有可用于恶意网络活动的潜在能力时向商务部报告有关外国人对大型人工智能模型进行训练的信息。应报告的信息包括有关训练运行的识别信息（即客户的姓名、地址、客户账户的付款方式和来源、电子邮件地址、电话号码和 IP 地址）以及训练运行本身的情况。

[5] 根据 E.O. 14110，对于具有“潜在用于恶意网络活动能力的 AI 大模型”的阈值设定，相关大模型需要超过 1026 次的计算能力，并且是在一群强大的计算机上训练，这群计算机放在同一个数据中心里的，相互之间可以非常快速地交换数据（速度超过每秒 100GB），且整个数据中心的计算速度可以达到每秒 1020 次。

（四）关于特别措施的规定

1. 特别措施要求

在本次拟议规则中，美国商务部建议制定法规，以落实 E.O. 13984 中赋予商务部长的权力，即如果商务部长确定存在合理理由，可以断定美国以外的司法管辖区或个人“有大量外国人提供美国 IaaS 产品用于恶意网络活动，或有大量外国人直接获取美国 IaaS 产品用于恶意网络活动”，则可以采取其中任何一项特别措施[6]。该拟议规则中建议允许商务部自行启动调查，或接受其他行政部门机构或提供商的转介，以评估有关特定外国司法管辖区或外国人不合规的证据，从而决定是否实施特别措施。此后，商务部将评估其掌握的信息以及从公共和其他来源获得的有关外国人或外国司法管辖区的信息，以确定实施特别措施是否适当。

2. 合理理由的确定

E.O. 13984 规定，在确定某一特定外国管辖区是否“有大量外国人提供美国 IaaS 产品用于恶意网络活动，或有大量外国人直接获取美国 IaaS 产品用于恶意网络活动”时，商务部长必须考虑其他相关信息：

- 外国恶意网络行为者在该外国司法管辖区获得美国 IaaS 产品的证据，包括此类行为者是否通过经销商账户获得此类美国 IaaS 产品；
- 该外国司法管辖区在多大程度上是恶意网络活动的源头；
- 美国是否与该外国司法管辖区签署了法律互助条约，以及美国执法官员在获取有关源于或途经该外国司法管辖区的美国 IaaS 产品的活动信息方面的经验。

对于外国人，商务部长必须评估：

- 外国人使用美国 IaaS 产品进行、促进或推动恶意网络活动的程度；
- 外国人提供的美国 IaaS 产品在多大程度上被用于促进或推动恶意网络活动；
- 外国人提供的美国 IaaS 产品在多大程度上用于该司法管辖区的合法商业目的；
- 对于涉及提供美国 IaaS 产品的外国人的交易，在多大程度上不采取特别措施就足以防范恶意网络活动。

3. 特别措施的选择

商务部建议要求商务部长的调查过程包括 E.O. 13984 中提及的机构，即国务卿、财政部长、国防部长、总检察长、国土安全部长、国家情报局局长以及商务部长认为适当的其他行政部门和机构的其他负责人进行磋商，以决定实施哪项特别措施。磋商将包括：

---

[6] 根据 E.O. 13984，特别措施包括对某些美国以外的辖区内账户施加禁令或条件，以及对某些外国人施加禁令或附加条件。

- 审查现有证据，以确定是否对外国司法管辖区或外国人实施特别措施；
- 考虑实施特别措施是否会给提供商造成重大竞争劣势，包括与合规相关的任何不当成本或负担；
- 确定实施特别措施或特别措施的时间安排在多大程度上会对涉及外国司法管辖区或外国人的合法商业活动造成重大不利影响；
- 最后，该决定将包括评估任何特别措施对美国供应链、公共卫生或安全、国家安全、执法调查或外交政策的影响。

三、拟议规则外的 AI 监管新趋势

如前提示的，拟议规则的发布可能只是美国在云服务和 AI 监管领域的第一步，伴随着拟议规则的提出，美国政府和国会近期的一系列动作预示着后续美国将有更多更全面的云服务和 AI 管控和执法在酝酿之中。其中，最值得关注的主要包括以下方面：

（一）ICTS 审查关于 AI 监管执法箭在弦上

2024 年 1 月 22 日，BIS 下属的 OICTS 迎来了首位执行主任 Elizabeth Cannon。作为主要负责 ICTS 审查的部门，OICTS 也是本次拟议规则的主管和执行部门。作为一名已在出口管制领域深耕十余年的专业人士，Cannon 在加入 OCITS 前，曾在微软担任全球贸易合规高级顾问，并负责风险情报小组的监督工作，专注于与贸易合规相关的调查和取证，并且还在美国司法部拥有超过十年的公共服务和国家安全领域经验，曾担任出口管制和制裁的副主管，监督全国范围内所有出口管制和制裁违规案件。其在司法部期间还负责起诉多起涉国家安全案件，包括间谍活动、经济间谍活动、不当处理机密信息以及制裁、网络安全和出口管制违法案件。根据 BIS 的说明，Cannon 本人“展现了其对出口管制、制裁以及其他国际贸易和安全政策问题的深刻理解”。由一名有着执法经验丰富的执行主任负责 OICTS，也意味着 OICTS 未来的重心很可能将不仅仅限于新规起草，执法落地也将是其中非常重要的一环。因此，不仅是云计算相关业务，ICTS 所涉及的其他行业，特别是 AI 广泛运用的 TMT 行业以及 AIGC 相关应用也需要关注未来 ICTS 审查的执法走向。

（二）云计算和 AI 训练的进一步出口管制限制

虽然 ICTS 审查同样由 BIS 负责，但是在 BIS 内部，其与出口管制执法分属不同条线管理，BIS 也已经明确表示关于针对中国等国家 AI 跨境训练的云计算出口管制限制措施正在逐步推进之中。例如在 2023 年 12 月 12 日，美国众议院外交事务委员会监督和问责小组委员会所举行题为“审查 BIS，第二部分：战略竞争时代的美国出口管制”的听证会

上，BIS 出口管制助理部长 Thea D. Rozman Kendler 表示，BIS 正在与利益相关方合作，准备根据《国防授权法案》的授权对云计算进行出口管制，限制从中国国内访问云计算以获得 AI 在线生成物，但这还需要国会给予其更多授权。而在 2024 年 1 月 26 日，BIS 国家安全与技术转移管制办公室总监 Eileen Albanese 在马萨诸塞出口中心的会议上表示，BIS 在今年将持续推进云计算出口管制相关工作。目前 BIS 正在去年 10 月发布出口管制新规所征询的公共意见基础上，拟就新规及云计算相关内容做相应调整和澄清，并明确表示“云计算的出口管制毋庸置疑是 BIS 重点关注领域”。

在 BIS 推动相关出口管制同时，美国国会也在努力推动相关出口管制措施的落地。在 2023 年 12 月 12 日，众议院美中战略竞争委员会发布了一份 53 页的两党报告，提出近 150 项政策建议，旨在“重置”美中两国日益紧张的经济和技术关系。在出口管制方面，报告建议国会应要求政府对 AI、量子技术、生物技术和其他新兴技术“迅速建立全面控制”，并要求 BIS 制定一项新的云计算最终用途规则，从而堵上中国公司用来远程获取出口控制技术的漏洞，同时还要求美国与盟国就扩大对 AI 等技术的多边管控进行谈判，以加强对该等新兴技术的多边管控。显然，美国国会议员们并不满足于现有框架下的行政部门出口管制思路，希望建立更为全面的管制措施，限制中国公司使用美国的云计算服务。

（三）国会关于 AI 立法持续推进

除持续敦促 BIS 等监管机构加强执法力度外，美国国会在本届任期内也大量推出与云计算和 AI 监管相关的法案，不少法案目前已经通过了委员会审议，正等待排期表决。其中，比较重要的法案包括：

时间	具体情况
2023 年 2 月 21 日	众议院引入两党法案《打击海外不受信任电信法案》（Countering Untrusted Telecommunications Abroad Act，H. R. 1149），提议修订《1934 年证券交易法》，要求相关发行人披露与“不受信任”的电信设备相关的某些活动，包括所涵盖的电信设备或服务是否用于云计算或数据存储（该条后经修订）。
2023 年 3 月 7 日	参议院引入《限制信息和通信技术安全威胁出现法案》（Restricting the Emergence of Security Threats that Risk Information and Communications Technology Act，S. 686），要求优先考虑对互联网托管服务、基于云或分布式计算及数据存储、内容交付服务等领域采取行动，以解决造成“对手国家”及相关实体所造成的不当或不可接收风险的信息和通信技术产品及服务。

时间	具体情况
2023 年 7 月 17 日	众议院引入两党法案《填补人工智能海外使用和发展漏洞的法案》（Closing Loopholes for the Overseas Use and Development of Artificial Intelligence Act，H.R. 4683），指出中国在芯片管制规则出台后仍然利用云服务来远程访问美国在线技术开发 AI 工具及模型，要求禁止中国（包括澳门特别行政区）实体远程使用或云使用出口管制编码为 3A090 和 4A090 的集成电路。

上述法案足以已经显示出美国国会对云计算和 AI 应用的高度关注，特别是对中国等“对手国家”以及“不受信任”的中国企业利用美国云计算服务的担忧。同时，美国国会多个委员会正筹划设立专门的工作组，并开展 AI 相关立法前置活动，以提高 AI 相关的立法效率。例如，众议院财政服务委员会在 2024 年 1 月 11 日设立了两党关于 AI 的立法工作组，相关工作组将由数字资产和财务技术小组委员会主席 French Hill（阿肯色州共和党众议员）和小组成员 Stephen Lynch （马萨诸塞州民主党众议员）共同领导，以监督 AI 相关的立法。而在 2024 年 1 月 10 日，美国参议院国土安全委员会和司法委员会分别召开了 AI 相关的全员听证会，讨论了规范 AI 后续运用的一系列问题。后续国会在 AI 的立法活动预计将持续加强。

四、中国企业需要未雨绸缪

基于美国近期针对云计算和 AI 相关领域的一系列立法和执法新趋势，中国的相关企业需格外重视，并在以下方面做好准备：

（一）提前评估潜在风险

虽然对于 AI 服务而言，本次拟议规则的相关限制仍显克制，其相关限制局限于提供硬件租用服务的 IaaS 提供商，而不包括“软件即服务”（SaaS）和平台即服务（PaaS），同时也不包括美国提供商的外国子公司，但需要注意的是，拟议规则的核心在于提出了对于云计算监管的全新思路，即采取了金融监管上常用的“了解你的客户”（Know Your Client，“KYC”）的方式，对 IaaS 提供商提出了全新的合规监管要求。由于云服务提供商可以了解其客户所使用的算力，并且可以随时限制或关闭云计算服务，因此，如同商业银行可以监管客户的资金用途一样，美国完全可以通过美国 IaaS 服务商监控甚至阻断中国 AI 公司利用云计算开展的业务活动。而如果利用美国境外的算力开展业务，考虑到目前 BIS 正在筹划的云计算出口管制新规，总部位于中国的相关公司在使用该等算力的过程中，可能因底层硬件原因而需要面临额外的出口管制限制。有鉴于此，虽然目前相关规

定均未正式落地，但中国企业仍需要提前妥善评估自身相关业务模式中的关键节点，如目前主要的算力来源和物理服务器所在地，是否主要客户位于中国，以了解其可能面临的监管挑战和算力供应风险，并基于此进行必要的业务调整，以降低潜在风险。

（二）持续关注政策变化

目前美国国会和行政部门关于 AI 的相关立法层出不穷，基于相关立法前准备（如立法听证会、专业会议讨论等）可以大致了解美国国会和行政部门对于 AI 和云计算领域的后续立法思路和态度。这有助于企业可以提前预警相关风险，并做出相应准备，而不至于在相关规则落地时猝不及防，应对失据。由于美国立法活动相对复杂，寻求专业机构的支持来进行政策监控能够大大提高企业的信息监控能力，避免关键信息遗漏。

（三）全方位考量业务布局风险

需要提示的是，AI 业务作为全球范围内的新兴业务，除美国方面的持续监管压力外，包括中国、欧盟在内的诸多法域目前也在推进自身的 AI 立法。因此，对于 AI 公司而言，在进行业务模式构建时，需要对业务可能涉及的国家、类型、数据传输方式等进行全面评估，除了出口管制等特殊限制外、关于数据安全、隐私保护等相关要素的评判也必不可少，加强合规体系和管理机制更是重中之重，切不可一味业务优先，而将合规风险放诸脑后。在目前的国际贸易和政治形势下，任何合规问题均可能被放大成一场风暴，对企业发展和生存带来无法承受的影响。

结语

AI 领域作为美国与中国进行战略竞争的核心领域之一，美国逐步完善和加强 AI 相关立法和监管，以谋求和中国的竞争优势的意图也早已昭然若揭。ICTS 审查新机制下的 AI 拟议规则从长远来看，只是美国此等努力中的拼图一块而已。如我们反复强调的，在如今的国际局势下，中国企业，特别是涉及 AI 等关键行业和领域的企业，更需要具有国际化视野的风险和危机意识，才能积跬步而致千里。

感谢实习生杨雅岚对本文作出的贡献。

全球人工智能治理大变局之欧盟人工智能治理监管框架评述及启示

宁宣凤 吴涵 陈琳珺 吴仁浩 董方倩

引言

继 ChatGPT 发布后，全球人工智能产业就生成式人工智能的研发和创新掀起一股浪潮。国内多家大型平台也相继推出自己的大模型产品。人们在惊叹于当前 AI 技术的高效便捷的同时，AI 技术所带来的潜在风险也日益凸显。2023 年 5 月 8 日，布鲁金斯学会（Brookings Institution）发表一篇名为《人工智能的政治：ChatGPT 和政治偏见（The politics of AI: ChatGPT and political bias）》的评论文章[1]。该文章指出，类似于 ChatGPT 这种基于大型语言模型（large language models, LLM）的聊天机器人甚至可能存在政治偏见。

2023 年 6 月，欧洲议会通过了《人工智能法案》的折衷修订草案（“2023 年《AI 法案》折衷草案”），并就法案的具体条款组织包含欧洲议会、欧盟理事会和欧盟成员国在内的“三方会谈”，于 2023 年 12 月 9 日达成有关《人工智能法案》的临时协议。2024 年 3 月 13 日，欧洲议会以 523 票赞成、46 票反对和 49 票弃权的表决结果通过了《人工智能法案》（Artificial Intelligence Act）[2]。由于本文成文时间较早，撰写依据为 2023 年 6 月及 6 月之前公开的《人工智能法案》各版本。彼时《人工智能法案》正式法案尚未出台，鉴于法律法规更新的时效性，如本文相关规定与最新规定不一致的，请以最新立法规定为准，下文不再赘述。

近年来，我国也不断建立自己的人工智能治理监管体系，国家网信办于 2023 年 4 月发布的《生成式人工智能服务管理办法（征求意见稿）》[3] 以及 2023 年 7 月 13 日最新发布的《生成式人工智能服务管理暂行办法》[4] 正是对目前如火如荼的人工智能产业监管作出的及时回应。本文将从欧盟《人工智能法案》（使用“《人工智能法案》”或“《AI 法案》”

---

[1] https://www.brookings.edu/blog/techtank/2023/05/08/the-politics-of-ai-chatgpt-and-political-bias/.

[2] 具体详见本书中《历时三年，欧盟 < 人工智能法案 > 通过欧洲议会表决》一文。

[3] 对该意见稿的解读具体详见《“不要温和地走进那良夜”——对 < 生成式人工智能服务管理办法 > 的思考》https://mp.weixin.qq.com/s?__biz=MzA4NDMzNjMyNQ==&mid=2653305510&idx=1&sn=bf690aaeef343cc1cf31cdce4529d2fd&chksm=843a510cb34dd81ad0bf6ff6b95766c5141161a3e9589dc4a4458c847c665b6c043b1eafc751&scene=21#wechat_redirect。

[4] 对该法规的解读详见本书中《卧看星河尽意明——全球首部生成式人工智能法规解读》一文。

概括指代目前为止所有《人工智能法案》的立法文件）的立法进程出发，以比较分析的方法梳理欧洲对人工智能监管的变革和重点制度，并总结其可鉴经验，以提出对中国人工智能治理的进一步展望，供读者参阅。

一、欧盟人工智能治理立法发展概述

（一）立法概况

欧盟自 2016 年起就不断探索推进对人工智能技术应用的监管体系建构。2018 年，欧盟建立人工智能高级专家小组（High-Level Expert Group on Artificial Intelligence，Al HLE），加快建立一个统一的人工智能法律监管框架的步伐。随着《人工智能法案》三次修改及谈判草案的发布，欧盟在世界范围内率先设计一系列措施以确立人工智能的治理规则体系，并尝试影响甚至塑造全球范围内的人工智能治理共识规则和标准。

表 1 欧盟的人工智能监管及算法治理规则一览表

文件名称	发布主体	发布 / 实施时间	内容介绍
European Civil Law Rules in Robotics（《欧盟机器人民事法律规则》）	欧盟法律事务委员会	2016 年	对基于人工智能控制的机器人，提出其使用的责任归属、伦理规则及对人类自身和财产的伤害赔偿等监管原则
Artificial Intelligence for Europe（《欧盟人工智能》）	欧盟委员会	2018 年	提出“以人为本”的人工智能发展路径
Artificial intelligence：Commission kicks off work on marrying cutting-edge technology and ethical standards（《欧盟委员会启动发展符合伦理标准的尖端技术的研究工作》）	欧盟委员会	2018 年	欧盟委员会在这条新闻中公布了即将启动人工智能专家小组的计划，并面向公众开放专家组的申请征集。
Commission outlines a European approach to boost investment and set ethical guidelines（《探索面向人工智能领域推动投资以及伦理标准建设的欧洲路径》）	欧盟委员会	2018 年	欧盟委员会为提高欧洲在人工智能技术领域的竞争力以及使这项技术更好地为欧洲人服务所做的一系列措施。包括加强财政支持、为人工智能带来的社会经济变化做好准备、确保适当的道德和法律框架。

文件名称	发布主体	发布 / 实施时间	内容介绍
Communication Artificial Intelligence for Europe（《关于发展欧洲人工智能的交流会》）	欧盟委员会	2018 年	这次交流探讨了欧盟在国际竞争格局中的地位，提出了《关于发展欧洲人工智能的倡议》。倡议内容包含：提高欧盟的技术和工业能力以及人工智能技术在经济中的渗透、为社会经济变化做好准备、确保适当的道德和法律框架、鼓励会员国制定人工智能战略、鼓励利益相关者共同组建欧洲人工智能联盟等内容。
Coordinated Plan on Artificial Intelligence（《欧盟委员会：关于人工智能的合作计划》）	欧盟委员会	2018 年	这份文件旨在基于伦理和社会价值观发展可信赖的人工智能，并为欧洲创造一个创新友好的人工智能生态系统。
Member States and Commission to work together to boost artificial intelligence “made in Europe”（《欧盟委员会（新闻）：发展欧洲制造的“人工智能”》）	欧盟委员会	2018 年	这条新闻中，欧盟委员会公布了一项协调计划，该计划旨在推动“欧洲制造”的人工智能。该计划提出了要在四个关键领域加强和推动更有效的合作，具体包含：增加投资、提供更多数据、培养人才和确保信任。
Coordinated Plan on Artificial Intelligence（《欧盟委员会：关于发展欧洲制造的“人工智能”的交流》）	欧盟委员会	2018 年	该计划基于三大支柱：增加对人工智能的公共和私人投资，为社会经济变革做好准备，并确保建立适当的道德和法律框架。

文件名称	发布主体	发布 / 实施时间	内容介绍
A Governance Framework for Algorithmic Accountability and Transparency（《算法的可问责和透明的治理框架》）	欧洲议会	2019 年	基于对现有算法系统治理提案的广泛审查和分析，提出了政策制定的四个选项，每个选项均旨在解决算法透明度和问责制的不同方面： 1. 提高认识：教育、监督者和举报人。 2. 公共部门使用算法决策的问责制。 3. 监管监督和法律责任。 4. 算法治理的全球协调。
Ethics Guidelines for Trustworthy AI（《可信人工智能伦理指南》）	人工智能高级别专家组（AI HLEG）	2019 年	提出了值得信赖的人工智能应当具有三个要求，且在人工智能系统的整个生命周期中均应得到满足： 1. 应是合法的，遵守所有适用的法律和法规； 2. 应是合乎道德的，确保遵守道德原则和价值观，以及 3. 无论从技术角度还是社会角度，人工智能系统应是稳健的。因为即使是出于善意，人工智能系统也可能造成无意的伤害。
Communication on Building Trust in Human-Centric Artificial Intelligence（《关于在以人为本的人工智能中构筑信任的讨论》）	欧盟委员会	2019 年	这次交流会议中，各方就 2018 年 4 月发布的欧盟发展人工智能的倡议（见上文 Communication Artificial Intelligence for Europe）以及 2018 年 12 月发布的《可信赖人工智能伦理准则》初稿的相关内容进行了讨论。

文件名称	发布主体	发布 / 实施时间	内容介绍
The first European AI Alliance Assembly（《首届欧洲人工智能联盟大会召开》）	欧盟委员会	2019 年	这次活动标志着欧洲人工智能联盟平台成立一周年。联盟汇集了包括普通公民在内的利益相关者和政策制定者，讨论人工智能政策的最新成就、欧洲人工智能战略的未来前景，以及其对经济和社会的影响。
Pilot the Assessment List of the Ethics Guidelines for Trustworthy AI（《试行“可信赖人工智能道德指引评估清单”》）	欧盟委员会	2019 年	“可信赖人工智能评估清单”是《可信赖人工智能伦理准则》的操作工具，旨在确保用户从人工智能 (AI) 中受益，而不会面临不必要的风险。
White Paper on Artificial Intelligence – A European Approach to Excellence and Trust（《人工智能白皮书——追求卓越和信任的欧洲方案》）	欧盟委员会	2020 年	法律草案一共 12 章 85 条，明确规定禁止使用的人工智能类型及要件、高风险人工智能的范围和种类、人工智能系统的透明度要求、系统入市后的监督和职责等。其由两项重要板块构筑：卓越生态系统（ecosystem of excellence）和信任生态系统（ecosystem of trust）。并且提出了在设计阶段和产品上市后阶段应落实的要求，同时革新安全和责任立法的展望。对于某些“高风险”行业，如医疗、交通、能源及特定公共领域，或影响劳动者（以及可能包括消费者）权利的情形，或涉及生物特征识别或监视的情形，可能会施以强制性要求。

文件名称	发布主体	发布 / 实施时间	内容介绍
A European Strategy for Data（《欧洲数据战略》）	欧盟委员会	2020 年	提出将就影响数据敏捷型经济体系中各主体关系议题探讨立法行动的必要性，解决包括企业间共生数据的共享（物联网数据）和建立数据池（用于数据分析和机器学习）的安全和信任问题。
Public consultation：Artificial intelligence – ethical and legal requirements（《对 < 人工智能伦理及法律要求 > 的公众问询》）	欧盟委员会	2020 年	欧盟委员会就如何确保人工智能安全、合法，并符合欧盟的基本权利向公众进行了问询。这一系列问询的总体目标是推动欧盟经济采用可信赖的人工智能。
Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment（《人工智能高级别专家组：可信赖人工智能最终评估清单 (ALTAI)》）	欧盟委员会	2020 年	评估清单经过多次修订和完善，转化为支持人工智能开发人员和部署人员开发可信人工智能的工具。该工具具有可操作性，符合 2019 年 4 月向欧盟委员会提交的人工智能高级专家组（AI HLEG）提出的《可信赖人工智能伦理准则》概述的关键要求。
Second European AI Alliance Assembly（《第二届欧洲人工智能联盟大会召开》）	欧盟委员会	2020 年	提出欧洲在人工智能领域建立卓越和信任生态系统的倡议。
Communication on Fostering a European approach to Artificial Intelligence（《关于推动人工智能发展的欧洲路径的交流》）	欧盟委员会	2021 年	目标是促进欧洲在人工智能方面的创新、竞争力和可持续发展。欧盟需要以面向未来的方式抓住人工智能的众多机遇并应对挑战。为促进人工智能的发展并平等地解决其对安全和基本权利构成的潜在高风险，制定该协调计划。

文件名称	发布主体	发布 / 实施时间	内容介绍
The Proposal for a Regulation laying down harmonized Rules on Artificial Intelligence（《人工智能法案》提案）	欧盟委员会	2021 年	在对人工智能系统进行分类的基础上，采取基于风险的方法（Risk-based Approach），主要对高风险的人工智能系统进行监管。 针对高风险的人工智能系统，法案提出的要求包括： 1. 建立和维持风险管理系统； 2. 数据和数据治理要求； 3. 透明度和向用户提供信息； 4. 人为监管等。
Coordinated Plan on Artificial Intelligence 2021 Review（《关于制定〈2021 年人工智能协调计划〉的审议》）	欧盟委员会	2021 年	欧盟委员会制定的《2021 年人工智能协调计划》是欧盟在可信赖人工智能领域建立全球领导地位的下一步行动。该行动旨在：加快对人工智能技术的投资，在采用新的数字解决方案的帮助下，推动有弹性的经济和社会复苏；就人工智能策略和计划采取行动，全面及时地实施这些策略和计划，以确保欧盟充分受益于先发采用者的优势；调整人工智能政策，增强国际协作，应对全球挑战。

文件名称	发布主体	发布 / 实施时间	内容介绍
European Commission：Proposal for a Regulation on Product Safety（《欧洲委员会关于一般产品安全的立法提案》）	欧盟委员会	2021 年	该立法提案为人工智能系统在欧盟的投放、投入使用和使用制定了统一的规则。这些规则需要确保高度保护公共利益，特别是在健康和安全方面，以及人民的基本权利和自由。它规定了高风险人工智能系统必须遵守的具体要求，并规定了此类系统的提供者和用户的义务。
European Economic and Social Committee, Opinion on the AI Act（《欧洲经济及社会委员会 (EESC) 关于〈人工智能法案〉的意见》）	欧洲经济及社会委员会 (EESC)	2021 年	EESC 认为该提案需要改进的地方包括： • 禁止人工智能操作的范围、定义及清晰度； • 与“风险金字塔”相关的分类原则的含义； • 对高风险人工智能的要求的风险缓解效果； • 欧洲《人工智能法案（AIA）》的可执行性；以及 • 与现行监管及其他近期监管建议的关系。

文件名称	发布主体	发布 / 实施时间	内容介绍
Committee of the Regions, Opinion on the AI Act（《欧洲区域委员会关于欧洲人工智能法案的修改意见》）	欧洲区域委员会	2021 年	欧洲区域委员会的修改意见主要包括以下几个方面： 强调定义人工智能系统是一个持续的过程，应考虑到人工智能运行的背景，跟上该领域的社会发展，不忽视卓越生态系统与信任生态系统之间的联系；呼吁委员会更深入地考虑拟供公共当局使用的人工智能系统的高风险分类； 要求符合性评估透明化，并向公众开放。此外，地方当局还应该能够参与人工智能系统的监测，报告其在当地的实施情况，并为欧盟委员会对该法规的应用进行的评估作出正式贡献。
European Central Bank, Opinion on the AI Act（《欧洲央行关于〈人工智能法案〉的意见》）	欧洲央行	2021 年	欧洲央行承认，人工智能支持的创新在银行部门越来越重要。考虑到银行活动中人工智能创新固有的跨境性质和机会，欧洲央行作为联盟层面的审慎监管机构，强烈支持在涉及审慎风险和要求时，确保信贷机构协调实施拟议监管的必要性。同样，鉴于人工智能的重要性日益增加，需要联盟立法者在未来考虑在联盟一级建立一个独立的人工智能管理局的可能性，负责就健康、安全和基本权利的具体问题在整个单一市场协调适用拟议的条例。

文件名称	发布主体	发布 / 实施时间	内容介绍
European Parliament, ENVI opinion（《环境、公共卫生和食品安全委员会关于〈人工智能法案〉的意见》）	环境、公共卫生和食品安全委员会	2022 年	《人工智能法案》应该维护欧洲的价值观，促进人工智能在整个社会的利益分配，保护个人、公司和环境免受风险的影响，同时促进创新和就业，并使欧洲成为该领域的领导者。委员会希望强调沙盒在某些领域（例如卫生）的重要性，委员会还强调了人工智能系统对心理健康的潜在影响。
European Parliament, ITRE opinion（《工业、研究和能源委员会关于《人工智能法案》的意见》）	工业、研究和能源委员会	2022 年	该委员会提议法案应在以下四个方面做出改进： • 加强支持创新的措施，如监管沙箱的预期措施，特别关注初创企业和中小企业； • 为人工智能系统提供简明和国际公认的定义，并为准确性、鲁棒性、网络安全和数据设定高要求同时符合实际的标准； • 通过强调社会信任和价值链责任，鼓励行业采用人工智能系统； • 通过更好地与绿色转型以及产业、技术和人工智能建立更多连接，以验证《人工智能法案》的作用。

文件名称	发布主体	发布 / 实施时间	内容介绍
European Parliament, TRAN opinion（《交通和旅游委员会关于〈人工智能法案〉的意见》）	交通和旅游委员会	2022 年	该委员会提议法案应在以下三个方面做出改进： • 确保《人工智能法案》不与部门立法重叠，对运输行为者施加双重 / 冲突的义务； • 促进和维护对运输行业特别重要的国际标准； • 促进研究和创新，以确保欧盟的运输部门在实施人工智能方面开发自己的专门知识，同时坚持最高的道德标准。
Proposal for an AI Liability Directive（《欧洲消费者组织：关于人工智能责任指令的提议》）	欧洲消费者组织	2022 年	欧盟的产品责任规则必须建立一个明确和可执行的法律框架，让消费者能够诉诸司法。在此背景下，欧洲消费者组织针对人工智能责任指令提出了建议。
Proposal for a Regulation laying down Harmonized Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts（《人工智能法案》妥协版本）	欧盟委员会	2022 年	妥协版本突出强调了欧盟与国际接轨的人工智能系统定义，并提出了对于小型企业的豁免要求。
Proposal for a regulation of the European Parliament and of the Council on harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts（《人工智能法案》折衷草案）	欧盟委员会	2023 年	折衷草案注重基于风险来制定监管制度，进一步细化和补充了风险管理制度，以平衡人工智能的创新发展与安全规范。此外，草案严格禁止对人类安全造成不可接受风险的人工智能系统，并扩大了人工智能高风险领域的分类，将对人们健康、安全、基本权利或环境的危害考虑在内。

文件名称	发布主体	发布 / 实施时间	内容介绍
Artificial Intelligence Act（《人工智能法案》）	欧盟委员会	2024 年	2024 年 2 月 13 日，欧盟议会内部市场和消费者保护委员会与公民自由、司法和内政事务委员会以 71 票赞成、8 票反对和 7 票弃权的投票结果通过了与各成员国就《人工智能法案》达成的谈判草案。2024 年 3 月 13 日，欧洲议会正式批准了《人工智能法案》，其立法宗旨之一在于维护公平的人工智能竞争环境并有效保护欧盟个人的自由等基本权利。

起初，欧盟对于人工智能的监管主要基于对技术发展的憧憬及对其产生的担忧。2016 年发布的《欧盟机器人民事法律规则》从民事法律的角度对基于人工智能控制的机器人制定了相应的权利及义务规则。尽管该法律文件因机器人的法律地位存在较大争议而颇受非议[5]，但该文件作为欧洲乃至全球第一份专门规制人工智能的政策文件，对于欧洲的人工智能发展及人工智能伦理治理的发展均产生较大影响[6]。

随后，2019 年发布的《可信人工智能伦理指南》正式确立了“以人为本”的人工智能发展及治理理念，其提出的可信人工智能三大要素也进而成为世界范围内较为通行的人工智能评估准则。与此同时，《算法的可问责和透明的治理框架》也于同年发布，确立了人工智能的算法问责机制及透明义务的规则。

2021 年 4 月，《人工智能法案（提案）》（“2021 年《AI 法案》提案”）正式发布，该法案引入风险分级监管、市场准入制度、监管沙盒等制度，其目的在于应对突出的算法黑箱问题，以确保投放到欧盟市场的人工智能系统及其使用的安全性。在随后的两年，该法案陆续经过两次谈判修订，并于近日进入最终谈判阶段，该法案如正式获得批准，将有望成为全世界首部有关整体人工智能规制的法律。

此外，近年来，欧盟针对数据治理也陆续发布《数字服务法案》（Digital Service Act）、《数字市场法案》（Digital Market Act）、《数据治理法案》（Digital Governance Act）。这些法案旨在规范欧洲的数据利用和流转制度，与即将正式出台的《人工智能法案》将共同构成欧盟数据战略框架下的重要监管规则，一方面从底层逻辑入手，加

[5] 刘洪华：《人工智能法律主体资格的否定及其法律规制构想》，载《北方法学》，2019 年第 4 期第 56-66 页。
[6] 郭佳楠：《欧盟人工智能的政策、伦理准则及规制路径研究》，载《互联网天地》，2023 年第 1 期第 26-32 页。

强数据的安全保护，促进欧洲数据流动，防范算法自动化决策的潜在风险；另一方面也建立相关伦理价值标准，保障个人权利，构建监管与创新发展的平衡机制。

（二）欧盟人工智能治理框架下的相关法律概念的界定

1. 人工智能

目前人们对人工智能的定义并不统一。欧盟广泛使用的人工智能定义来自《2018 年人工智能战略》，该战略指出：“人工智能（AI）是指通过分析环境并采取行动（具有一定程度的自主性）以实现特定目标来展示其智能行为的系统。基于人工智能的系统可以完全依赖于软件，在虚拟世界中运行（例如语音助手、图像分析软件、搜索引擎、语音和人脸识别系统）或者也可以嵌入硬件设备中（例如高级机器人、自动驾驶汽车、无人机或物联网应用程序）。”[7]

2021 年《AI 法案》提案第 3 条对人工智能的定义为：“AI 系统指采用附录 1 中所列的一种或多种技术和方法开发的软件，该软件能生成影响交互环境的输出（如内容、预测、建议或决策），以实现人为指定的特定目标。”其中，附录 1 列举的技术方法主要包括：机器学习方法（包括监督、无监督、强化和深度学习）；基于逻辑和知识的方法（包括知识表示、归纳编程、知识库、影响和演绎引擎、符号推理和专家系统）；统计方法，贝叶斯估计，以及搜索和优化方法。

事实上，“人工智能”的概念自 1956 年于美国的达特茅斯学会上被提出后，其所涵盖的理论范围及技术方法随着时代的发展也在不断扩展。如今，人工智能技术也发展出多个技术分支，应用于不同的领域中。相比于《2018 年人工智能战略》，2021 年《AI 法案》提案对于人工智能的定义采取更加宽泛的界定标准。值得注意的是，在 2022 年《AI 法案》妥协版本中，欧盟理事会及欧洲议会对于上述界定的观点有进一步意见，其认为“AI 系统”的定义范围应适当缩窄，并侧重强调机器学习的方法。

我们理解，上述两种定义方法可能产生不同的实施效果，各有利弊。一方面，若采取更加宽泛的定义，则法案的适用范围也将更广，考虑到法律规制较之于技术的发展相对滞后，更宽的适用范围针对不断更新的技术能够提供相应的指导意见，以防止各类新技术“野蛮生长”；另一方面，若采取较为限缩的定义，则可在一定程度上避免一些已经应用比较广泛的技术（例如视觉智能等）被进一步监管，从而限制其发展。从最新的修改方案来看，欧盟可能倾向于采用更加宽泛的定义以期制定更加广泛的适用范围，但具体的界定仍待《人工智能法案》的正式版本发布后确定。

---

[7] 郭佳楠：《欧盟人工智能的政策、伦理准则及规制路径研究》，载《互联网天地》，2023 年第 1 期第 26-32 页。

2. 通用型人工智能系统

通用型人工智能系统（General Purpose AI system）的概念于 2022 年《人工智能法案》妥协版本（“2022 年《AI 法案》妥协版本”）中提出，在该版本第 6 条中，通用型人工智能系统是指“一种人工智能系统，无论其如何投放市场或投入使用，包括作为开源软件，其目的是由供应商执行普遍适用的功能，如图像和语音识别、音频和视频生成、模式检测（pattern detection）、问答、翻译等; 通用型人工智能系统可在多个环境中使用，并可集成在多个其他 AI 系统中”。此外，该版草案进一步指出，通用型人工智能既可能单独作为高风险人工智能系统，也可能仅是某高风险人工智能系统的组件；如果某通用型人工智能系统提供者确信自己的系统不会用于《人工智能法案》规定的高风险场景，则其无需承担相应的高风险规制义务。

相较于上述版本，2023 年《AI 法案》折衷草案中将上述定义进行了简化和扩展，即“通用人工智能系统是指可用于和适应广泛应用的人工智能系统，而非经过有意专门设计的系统。”此处修改放弃了 2022 年《AI 法案》妥协版本中的“定义 + 列举”的方式，并强调了此类人工智能系统的广泛适用性。

3. 高风险 AI 系统

2021 年《AI 法案》提案首次提出了高风险 AI 系统（high-risk AI systems）的概念，但未给出明确定义，主要采用列举的方式列出了高风险 AI 系统的类型。提案中规定，高风险 AI 系统包含两种类型：一是附录 2 中根据某些欧盟法律作为安全组件或产品使用的 AI 系统；二是附录 3 中在特定领域使用的某些类型的 AI 系统。其中，附录 3 中列出了八个领域的高风险 AI 系统，具体为：

- 生物识别和以生物识别为基础的分类；
- 可能威胁人的生命和健康的关键基础设施的管理和运营；
- 可能决定人的受教育机会和职业培训就业；
- 就业、员工管理和获得自营职业；
- 获得和享用基本的私人服务和公共服务及福利；
- 可能影响人的基本权利的执法活动；
- 移民、庇护和边境管理；
- 司法和民主程序。

而在 2022 及 2023 年的草案中，附录 3 的分类形式都被作出了一系列调整。在 2022

年《AI 法案》妥协版本中，用于“执法部门的深度伪造检测、犯罪分析、旅行文件验证”的目的被排除在高风险 AI 系统列表之外，与此同时，增加了“在关键数字基础设施中作为安全组件使用的 AI 系统、用于评估生命和健康保险定价或资格的 AI 系统”的目的。

此外，该版本将“纯粹辅助”型的 AI 系统也排除在高风险范围之外，即如果一个 AI 系统在其他高风险环境中仅用于辅助人类决策，则该系统不会被认定为“高风险”。

2023 年《AI 法案》折衷草案对上述“高风险”类别进一步补充，包括：

- 情绪识别系统和其他利用生物识别数据对个体进行推断的系统（除生物识别验证系统外）；
- 铁路和空中交通的安全组件。

用于下列目的的 AI 系统：

- 评估职业教育培训资格
- 在教育中检测舞弊
- 健康医疗分级
- 边境检查
- 移民和庇护预测
- 用于替代性争议解决过程（除法律程序外）
- 影响选举和全民公决的结果
- 用于大型社交媒体在线平台推荐

此外，如果分销商、进口商或用户对一个未被指定为“高风险”的 AI 系统进行了重大修改，那么该 AI 系统将自动成为高风险 AI 系统[8]。

由上述分类可知，欧盟对于“高风险”的划分原则遵循《人工智能法案》中所提到的人工智能治理的基本原则，也就是“促进以人为本和可信任人工智能应用，并保证对于健康、安全、基本权利、民主以及法治的高度保护”。可以看到，欧盟目前对于法案中所提出的风险分类的具体方式仍处于斟酌阶段，但从各草案的修改大抵可以看出，欧盟对于人工智能可能带来的算法歧视、算法黑箱等问题，仍持较为谨慎的态度。

（三）欧盟人工智能治理的重点原则与制度

1. 算法可解释性和透明度原则

随着自动化决策技术的普遍应用，大数据杀熟、算法黑箱等问题屡有发生。此前，欧

[8] 洪延青，欧盟《AI 条例》的立法进展和现有三个版本重点内容比较（截止 2023 年 7 月）。

盟的《通用数据保护条例》（“GDPR”）便针对算法自动化决策的问题提出了相应的规制要求，将透明度及算法问责原则列入该法案的核心原则之一。该法案第 22 条[9]也规定，数据主体有权拒绝仅基于算法自动化处理的决策。

事实上，早在 1978 年法国就曾在第 78-17 号文件《法律 - 信息技术、档案和自由法》中提出“数据主体有权知道自动化处理的逻辑信息并有权拒绝自动化处理作出的决定”[10]。2017 年法国《公共行政关系法》进一步规定，运用算法作出的行政决定，行政机关应向公民提供算法决策所涉模型、大致参数、一般权重、数据来源等[11]。

在美国，其《公平信用报告法》提出算法决策不利结果告知规则，以及 2017 年《关于算法透明和算法问责声明》中规定的算法解释原则，也将算法的可解释性规则确认为算法治理的基本原则之一。美国 2022 年的《算法问责法案》形成了“评估报告—评估简报—公开信息”三层信息披露机制。这种层级披露机制一方面保留了监管所需的必要信息并保障了消费者基本的知情权；另一方面也限制了核心信息的流通范围，避免算法控制者商业秘密泄露，从而平衡了算法透明和商业秘密保护之间的冲突。

但目前，产业界仍对此规则有所争议，有不少学者认为，从技术的本质出发，算法的运行无法达到完全可解释的效果，且即便向用户披露算法的运行机制也无意义，因为用户不具备相应的专业知识[12]；同时，履行该义务也有可能侵犯算法服务提供者的商业秘密[13]。

2. 风险分类分级监管与算法安全评估

《人工智能法案》的三次修改草案中，重点引入了以风险为导向，对于 AI 系统的分类分级监管制度。该法案提出了四种风险类型的 AI 系统：不可接受的风险、高风险、有限风险和极低风险。该法案对于不可接受的风险以及高风险等级的 AI 系统提出了严格的规制措施，同时为人工智能设计了全生命周期的规制措施，要求人工智能产品入市前评估和入市后监测，以便从事前、事中和事后共同治理。

自 GDPR 发布以来，有学者提出将数据保护影响评估制度（DPIA）与 GDPR 中的公私渠道“协同治理”制度有效联结，构建个人数据权利与算法治理相结合的影响评估机

---

[9] GDPRArt. 22 Automated individual decision-making, including profiling. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.

[10] 魏远山：《算法解释请求权及其权利范畴研究》，载《甘肃政法学院学报》，2020 年第 1 期第 143-156 页。

[11] EDWARDS L,VEALE M. Slave to the algorithm? why a right to explanationn is probably not the remedy you are looking for[J]. SSRN Electronic Journal, 2017, 16(1)：1884.

[12] Frank Pasquale. The Black Box Society: The Secret Algorithms that Control Money and Information. Cambridge, MA: Harvard University Press,2015, pp.102-115.

[13] 张凌寒：《商业自动化决策的算法解释权研究》，载《法律科学》2018 年第 3 期。

制[14]。其后欧盟的《算法责任与透明治理框架》也提出针对公共机构强制要求实施算法影响评估的要求。

这种以识别潜在影响并提出对应解决措施的“算法影响评估制度”同样被美国的立法机关所采用，在美国，2019 年《算法问责法案》明确了算法影响评估的主要内容，包括：对算法的详细描述；实现数据最小化；保障消费者对决策结果的获取权和修改权；评估算法对个人信息隐私和安全的影响，以及可能产生的歧视性风险；算法主体采取的降低风险的补救措施。

在我国，2021 年 9 月国家网信办发布的《关于加强互联网信息服务算法综合治理的指导意见》中明确提出了风险防控和算法分级分类安全管理的要求，强调了对高风险类算法的有效识别。此外，国家网信办等多部门于 2023 年 7 月 13 日最新发布的《生成式人工智能服务管理暂行办法》第 3 条及第 16 条也提出了对生成式人工智能服务的分类分级监管要求[15]。

欧美的风险分级监管路径以及算法安全评估方法虽然在具体的分类分级方式以及评估内容上存有一定争议，但总体上对我国仍具借鉴意义。

3. 人工智能的外部问责机制

2019 年 4 月，欧盟《算法问责及透明度监管框架》就算法自动化决策明确了具体的规制路径和政策建议。《人工智能法案》除了上述的评估监测机制外，更是从平台主体的义务、监管机构职责、处罚措施等各方面构建了对人工智能算法的外部问责机制。

针对算法的外部问责机制，美国率先通过 2017 年的《关于算法透明性和可问责性的声明》确定了算法的可问责性原则，并提出对于算法问责的各项程序。纽约市 2018 年设置了“算法问责特别工作组”（Algorithmic Accountability Task Force），负责调查市政府使用算法的情况，并就如何加强纽约市算法应用的公共问责提出建议[16]。2019 年，美国《算法问责法》进一步明确了大型算法平台所应当承担的算法可问责性义务，以规制可能的算法歧视问题。

此外《2020 年新西兰算法章程》也通过外部问责方式提升算法透明度，要求签署该章程的政府机构必须保证算法驱动决策的过程公开、透明、道德，包括提供算法过程和数据存储相关信息、进行风险评级以评估偏见出现的可能性及影响程度等，以构建公众对于

[14] See Margot Kaminski et al., Algorithmic Impact Assessments under the GDPR: Producing Multi-layered Explanations, International Data Privacy Law, Vol.11, Issue 2,2021.
[15] 参见本书《卧看星河尽意明——全球首部生成式人工智能法规解读》一文。
[16] 陆凯：《美国算法治理政策与实施进路》，载《环球法律评论》2020 年第 3 期，第 26 页。

政府机构自动化决策行为的信任[17]。

（四）欧盟对人工智能的宏观治理特点

纵观前文对欧盟人工智能治理框架的梳理，欧洲当前对于人工智能的监管和治理重点主张以个人自治权与人格尊严的保护作为核心理念。其试图通过建立人工智能监管体系，保障数据主体的基本权利，并尝试构建统一的监管规则，防范人工智能发展可能带来的风险，同时探索创新发展与安全规制的平衡。具体而言，欧洲的人工智能监管框架主要有以下几个特点：

- 在治理理念上，主张“以人为本”，采取发展与规制并行的理念，在防范风险的同时鼓励创新；
- 在治理主体方面，欧盟试图建立统一的监管机构，以实现和 GDPR 的衔接下，数据治理和算法规制主体的统一；
- 在治理的内容和手段上，欧盟强调以风险预防为导向，通过事前评估，事中监测，事后救济的多元路径，分级监管人工智能系统并规避其可能带来的风险，同时强调个人赋权和外部问责的协同治理。

二、2023 年《AI 法案》折衷草案重点变化分析

（一）重点制度变化与立法趋势分析

与 2022 年《AI 法案》妥协版本相比，2023 年《AI 法案》折衷草案对一些制度作出了修订，包括不可接受的人工智能和高风险人工智能清单、通用型人工智能系统的义务、监管沙盒制度的创新等。此外，2023 年《AI 法案》折衷草案整体上呈现加强个人数据保护的立法趋势。

作为欧盟针对人工智能的首次综合性立法尝试，欧盟人工智能法案从安全、隐私等方面制定了详细规则。2023 年《AI 法案》折衷草案突出了对个人数据的保护，比如使用情感识别系统或生物特征识别系统的用户应当向可能受系统影响的自然人告知该系统的运行方式，并在此基础上增加适用欧盟规定（EU）2016/679、（EU）2016/1725 和指令（EU）2016/280 处理他们的生物识别特征和其他个人数据之前获得自然人的同意的要求[18]，以及一般性地禁止在公众可进入的场所使用实时远程生物识别系统[19]。

---

[17] 侯海军 刘晓：《域外两种算法治理机制的分立与兼容》，载《人民法院报》，https://www.chinacourt.org/article/detail/2023/06/id/7375208.shtml。
[18] Article 52 of the Artificial Intelligence Act.
[19] Article 5（1）（d） of the Artificial Intelligence Act.

1. 扩大了被禁止的人工智能应用名单

2022 年《AI 法案》妥协版本	2023 年《AI 法案》折衷草案	变化趋势
第 5 条（1）（c）禁止公共机关或其代表使用 AI 系统进行信任度评估或分类。	第 5 条（1）（c）加入了对基于已知、推断或预测的个人或个性特征的社会行为评估，以及对特定自然人或群体进行有害处理的概念。	欧盟的 AI 法案采用了基于风险的监管方法，根据 AI 可能产生的风险水平，为提供者和部署者规定了不同的义务。对社会评分系统等对人类安全构成不可接受风险的 AI 系统实施了严格的禁止。折中草案更加注重隐私权保护。相较于前两版，折中草案对公共空间实时远程生物识别系统采取了更为严格的立场，全面禁止其使用。折中草案也禁止了因准确性问题可能被滥用的情绪识别系统。此外，折中草案还禁止了有可能侵犯人权和法律原则的“预测性警务”。在欧盟以外的地区，被视为不可接受的 AI 系统也是被禁止的，折中草案禁止欧盟供应商向第三国出口不可接受的 AI 系统。
第 5 条（1）（d）实时远程生物识别系统在公共可访问空间的使用仅限于法律执行的严格必要目标。	第 5 条（1）（d）、第 5 条（1）（db）取消了关于在公共可访问空间中实时远程生物识别系统的使用仅限于法律执行的严格必要目标的规定。同时，增加了新的限制，如不得创建或扩展通过从互联网或闭路电视画面中无目标地刮取面部图像的面部识别数据库。	
	第 5 条（1）（dc）明确禁止在执法、边界管理、职场和教育机构中使用 AI 系统推断自然人的情绪。	
	第 5 条（1）（d a）明确禁止使用 AI 系统对自然人或群体进行风险评估，以评估自然人犯罪或重新犯罪的风险，或预测实际或潜在的刑事或行政犯罪的发生或重发。	

欧盟人工智能法案遵循基于风险的监管方法，根据人工智能可能产生的风险水平，为提供者和部署商规定了不同义务。因此，对人们的安全具有不可接受的风险水平的人工智能系统将被禁止，例如用于社会评分（根据社会行为或个人特征对人进行分类）的系统。2023 年《AI 法案》折衷草案严格禁止对人类安全造成不可接受风险的人工智能系统，包括部署潜意识或有目的操纵技术、利用人们弱点或用于社会评分的系统。从折衷草案对不可接受的人工智能风险的修改来看，折衷草案更加注重对隐私权的保护。

在 2021 年《AI 法案》提案和 2022 年《AI 法案》妥协版本下，“生物识别系统”——根据关于他们的生物信息对人进行分组的系统——被视为有限风险的系统，只需遵守透明

度义务。折衷草案进一步禁止基于敏感或受保护特征对人进行识别的生物识别系统[20]。相比之前两个版本的人工智能法案，2023 年《AI 法案》折衷草案对于如何处理在公共可访问空间中的实时远程生物识别系统（如实时[21]人脸识别摄像头）采取了更为严格的立场，禁止在公共可访问空间中使用任何实时远程生物识别系统。使用人工智能系统在公共场所对自然人进行实时远程生物识别，对有关人员的权利和自由具有侵扰性，会影响到大部分人的私人生活，且对自然人进行远程生物识别的人工智能系统在技术上的不准确性可能导致歧视性结果。而在 2022 年《AI 法案》妥协版本中，在搜索犯罪受害者、防止对人们生命或安全的特定、重大和即将来临的威胁、防止恐怖袭击、搜索特定严重犯罪的嫌疑人、防止对关键基础设施的攻击和对健康的威胁等情况下，可以使用实时生物识别系统。

由于情绪识别系统在检测情绪、身体或生理特征时使用的人工智能技术可能缺乏可靠性、特异性和普适性，折衷草案也禁止情绪识别系统的应用。在执法、边境管理、工作场所和教育机构等现实生活中部署该系统时，也可能出现可靠性问题，会有滥用的重大风险，因此，应禁止将这些用于检测个人情绪状态的人工智能系统投放市场、投入服务或使用。

2023 年《AI 法案》折衷草案引入了对“预测性警务”的禁止。所谓“预测性警务”是指根据对自然人的画像，或者基于个性特征的数据分析，包括个人的位置，或者自然人或群体过去的犯罪行为，进行预测、画像或风险评估以预测实际或潜在的刑事犯罪或其他违法行为再次发生，包括欺诈预测系统，该类人工智能系统具有歧视某些人或群体的风险，可能侵犯人类尊严以及无罪推定的法律原则。在“预测性警务”人工智能应用方面，美国对这类预测性警务人工智能的应用更为普遍和广泛，其中在较有代表性的 COMPAS（Correctional Offender Management Profiling for Alternative Sanctions）系统的相关案例中曾引发关于人工智能再犯风险评估系统对于被告人质证权等正当权利损害的广泛讨论和争议，COMPAS 系统以对犯罪者的访谈以及司法部门提供的信息为依据，来评估再犯的风险系数，被告人并没有机会了解和评估法院所使用的 COMPAS 的算法和结果的准确性，也没有机会与该工具的开发者进行对质和辩论，因此 COMPAS 被认为可能侵犯被告人的正当诉讼权利。因此，欧盟采取对“预测性警务”的禁止行动，体现了欧盟在平衡人工智能创新发展与安全规范方面更倾向于保护欧盟境内人员的安全和维护欧盟价值观。

值得注意的是，2023 年《AI 法案》折衷草案明确提出，不可接受的人工智能系统在欧盟以外也是不可接受的。因此，在欧盟境内开发、部署或使用这类系统都是非法的，折衷草案也禁止居住在欧盟的供应者向第三国出口不可接受的人工智能系统。

---

[20] 洪延青，欧盟《AI 条例》的立法进展和现有三个版本重点内容比较（截止 2023 年 7 月）。

[21] 应该对“实时“和”事后”远程生物识别系统进行区分。在“实时”系统的情况下，生物识别数据的采集、比较和识别都是即时的、接近即时的或没有明显延迟的。在“事后”系统的情况下，生物识别数据被采集在先，比较和识别是在经过比较明显的延迟后才发生。

2. 扩大了高风险人工智能系统的名单

2022 年《AI 法案》妥协版本	2023 年《AI 法案》折衷草案	变化趋势
第 6 条（1）（b）其安全组件是 AI 系统的产品，或者 AI 系统本身就是产品，需要进行第三方合规性评估，以便根据列入附录 II 的欧盟法规将产品投放市场或投入使用。	第 6 条（1）（b）在折中草案中，第三方合规性评估现在与健康和安全风险相关，并明确了该产品的安全组件是根据第 (a) 点的 AI 系统。	折中草案扩大了高风险 AI 领域的分类，包括对健康、安全、基本权利或环境的影响。在高风险 AI 监管方面，2023 年的折中草案不仅强调了提供者的责任，还增加了部署者的义务，如配合进行符合性评估，并允许主管部门访问系统日志。这显示了对 AI 监管的严格趋势。
第 6 条（2）除了第 1 段中提到的高风险 AI 系统外，附录 III 中提到的 AI 系统也应被视为高风险。	第 6 条（2）附录 III 的 AI 系统仅在对自然人的健康、安全或基本权利构成重大风险时被视为高风险。 如果属于附录 III 第 2 点的 AI 系统对环境构成重大风险，也被视为高风险。 欧盟委员会应在法规生效前六个月内，提供关于 AI 系统可能对自然人的健康、安全或基本权利构成重大风险的明确指导。	
	第 6 条（2）（2a）如果附录 III 中的提供者认为其在重要领域和使用案例中的 AI 系统不构成重大风险，需向国家监管机构提出通知，并解释为何不受本条例第 III 章第 2 节的约束。如果系统在多个成员国使用，通知应寄给 AI 办公室。监管机构应在三个月内回复通知，如果认为系统被错误分类，可通过 AI 办公室处理。	
	第 6 条（2）（2b）错误分类的系统在监管机构提出异议之前投放市场，将面临第 71 条规定的罚款。	
	第 6 条（2）（2c）国家监管机构需向 AI 办公室提交年度报告，包括收到的通知数量、相关高风险领域和处理通知的决定。	

2022 年《AI 法案》妥协版本	2023 年《AI 法案》折衷草案	变化趋势
第 7 条（1）委员会有权在满足第 7 条（1）（a）、（b）的情况下，通过委托行政法规更新《附件 III》的高风险 AI 系统清单。	第 7 条（1）委员会有权在特定风险出现的情况下，通过授权法案修改《附件 III》的高风险 AI 系统或使用案例。	折中草案扩大了高风险 AI 领域的分类，包括对健康、安全、基本权利或环境的影响。在高风险 AI 监管方面，2023 年的折中草案不仅强调了提供者的责任，还增加了部署者的义务，如配合进行符合性评估，并允许主管部门访问系统日志。这显示了对 AI 监管的严格趋势。
	第 7 条（1a）委员会还有权在满足第 1 段所述条件的情况下，通过授权法案从《附件 III》的清单中删除高风险 AI 系统的使用案例。	
第 7 条（2）（c）考虑 AI 系统已经对健康和安全或基本权利造成的危害或影响程度，通过报告或提交给国家有权机关的记录的指控展示。	第 7 条（2）（c）考虑 AI 系统已经对健康和安全、基本权利、环境、民主和法治造成的危害或影响，或对此类危害或影响的可能性引起的重大关注，以报告或提交给国家监督机关、委员会、AI 办公室、EDPS 或欧洲联盟基本权利机构的记录的指控的方式来证明。	
	第 7 条（2a）当委员会评估 AI 系统以便满足第 1 或 1a 段的要求时，应咨询 AI 办公室和相关的代表团体，行业、独立专家、社会合作伙伴和民间社会组织。委员会还应在此方面组织公众咨询，并公开咨询结果和最后评估结果。	

如本文第一章第二节所述，2023 年《AI 法案》折衷草案扩大了人工智能高风险领域的分类，将对人们健康、安全、基本权利或环境的危害考虑在内。

2021 年《AI 法案》提案对高风险人工智能系统采取多项监管措施，包括上市前的严格管控、进行风险评估、确保活动可追溯、加贴 CE 标志、监管机构评估、市场监督、建立数据库、故障信息共享、严格执法和处罚等措施。2023 年《AI 法案》折衷草案对高风险人工智能系统的各项监管要求也有所修改，除了规定提供者的义务外，还规定了部署商的义务，包括在高风险人工智能系统必须根据第 43 条进行符合性评估的情况下，部署商应当与提供者合作调查原因。根据主管部门的合理请求，提供者和部署商还应允许国家主

管部门在其控制范围内访问高风险人工智能系统的自动生成日志，主管部门对于获取的信息负有保密义务。该条新增规定体现了折衷草案对于人工智能监管趋势趋严。

3. 新增关于通用型人工智能的条款

2023 年《AI 法案》折衷草案	变化趋势
第 3 条（1）（1d）定义了“通用 AI 系统”为可被用于和适应各种非专门和特定设计的应用的 AI 系统。	草案提出两个新概念：基础模型和通用 AI。基础模型，包括 GPT-3 和 DALL-E 等，是大规模训练的 AI 模型，可适应广泛的任务。生成式基础模型是一种专用于内容生成的 AI。通用 AI，也称强 AI，具有广泛智能，可适用于各种任务。草案规定，基础模型提供者需满足在市场投放前后和用于生成式 AI 的各项义务。义务包括风险评估、使用符合标准的数据集、设计和开发满足高性能、安全等要求的模型，最小化能源使用，编写技术文档和指南，将模型注册至 EU 数据库。通用 AI，如 ChatGPT，需接受全生命周期的外部审计，检测其性能等是否符合严格要求。
第 3 条（1）（1e）定义了“大规模训练”为生产强大 AI 模型的过程，该过程需要超过非常高阈值的计算资源。	
第 28 条 b（1）提供者在将基础模型投入市场或投入服务前，需要确保它符合本条规定的要求。	
第 28 条 b（2）描述了基础模型提供者的义务，包括通过合适的设计、测试和分析来识别、减少和降低风险，只处理和使用适当的数据治理措施的数据集，设计和开发基础模型以实现生命周期内的适当性能等。	
第 28 条 b（3）基础模型的提供者需要在基础模型投放市场或投入使用后的 10 年内，向国家主管当局提供技术文档。	
第 28 条 b（4）描述了专门用于生成具有不同自主性的内容（如复杂文本、图像、音频或视频）的基础模型的提供者以及那些将基础模型专门化为生成 AI 系统的提供者需要遵守的额外规定。	
第 52 条（1）AI 系统的提供者需要确保与自然人交互的 AI 系统的设计和开发方式，能够使自然人在与 AI 系统互动时，能及时、清晰、可理解地知道他们正在与一个 AI 系统互动，除非从环境和使用情境中明显可以知道这一点。	

2023 年《AI 法案》折衷草案提出了两个新概念：基础模型和通用型人工智能系统，基础模型被定义为一种依托大量数据被规模化训练的 AI 模型，为确保生成结果之通用性而设计，并能适应广泛的特定任务。基础模型可以是单模态和多模态的，包括大型语言模型如 GPT-3 和多模态模型如 DALL-E。其中生成式基础模型是基础模型的一种，特指被特别设定的、以不同水准的自主性进行内容生成的人工智能，如复杂文本、图像、音频或视

频等内容。“通用型人工智能系统”，这种系统可以用于它最初未被设计出来的广泛的用途。也被称为强人工智能，具有跨领域的广泛智能，能够适用于不同任务与领域。折衷草案将通用性人工智能定义为一种可被用于和可适应于广泛应用的、但未因此被有意和专门设计的人工智能。

2023 年《AI 法案》折衷草案对基础模型的提供者施加了新的规则，提供者需要满足一系列在市场投放前、市场投放后以及用于生成式人工智能的基础模型提供者的额外义务。相关义务可被分为：基础模型被投放市场或投入服务之前的一般义务、投放市场或投入服务后一定期限内的后续义务、用于生成式人工智能的基础模型提供者的额外义务。主要包括进行风险评估，仅使用符合适当数据治理标准的数据集、设计和开发模型要达到一系列要求：高性能、可预测性、安全性和其他属性，最小化能源使用和浪费，为下游提供者使用模型编写“广泛的技术文档”和“易懂的指南”，将模型注册到拟议中的欧盟高风险人工智能系统数据库。

随着近期 ChatGPT 等生成式人工智能技术的快速发展，ChatGPT 等通用目的人工智能系统的整个生命周期都必须接受外部审计，以测试其性能、可预测性、可解释性、可纠正性、安全性和网络安全性是否符合法案的最严格要求。

4. 算法透明度要求

2022 年《AI 法案》妥协版本	2023 年《AI 法案》折衷草案	变化趋势
第 13 条（1）透明度的目的是让用户能够解读人工智能系统的输出。	第 13 条（1）透明度确保提供者和用户能够理解系统的运作方式。提供者和用户会得到关于可解释性的额外规定，以及用户解释决策的能力。	折衷草案对人工智能系统的透明度要求更具体。要求提供者和用户都能理解系统的运行方式，包括数据处理和决策解释。相比之下，旧版法案主要要求披露提供者身份和系统性能特征。折衷草案还强调提供者和用户需具备足够的人工智能素养。另外，制造商需要披露训练模型所使用的版权数据信息。这些趋势表明对透明度和合规性的要求越来越具体，以保护自然人的权利。
第 13 条（2）使用说明必须包含简明、完整、正确和清晰的信息。	第 13 条（2）使用说明应当易于理解，包含操作和维护信息，并支持知情决策。使用说明可以以持久介质的形式提供。	

2022 年《AI 法案》妥协版本	2023 年《AI 法案》折衷草案	变化趋势
第 13 条（3）(b)(iii) 关于已知或可预见导致对健康、安全或基本权利构成风险的情况的信息。	第 13 条（3）(b)(iii) 清楚地指明已知或可预见的情况，这些情况可能导致对健康、安全、基本权利或环境构成风险。还添加了一些说明性示例以说明此类限制。	折衷草案对人工智能系统的透明度要求更具体。要求提供者和用户都能理解系统的运行方式，包括数据处理和决策解释。相比之下，旧版法案主要要求披露提供者身份和系统性能特征。折衷草案还强调提供者和用户需具备足够的人工智能素养。另外，制造商需要披露训练模型所使用的版权数据信息。这些趋势表明对透明度和合规性的要求越来越具体，以保护自然人的权利。
	第 13 条（3）(aa) 额外的要求包括关于执行符合性评估的实体的信息，第 13 条 (3) (b)(iiia) AI 系统解释其决策的程度，第 33 条（3）(ea) 用户收集、存储和解释日志的机制，第 13 条（3）3a 以及确保足够的 AI 素养。	
第 52 条（1）未要求告知 AI 启用功能、人工监督存在、决策责任人信息，未提现有权利和程序信息	第 52 条（1）需明确告知交互对象：启用了哪些 AI 功能，是否有人工监督，谁是决策责任人，以及现有的反对权利和救济程序	
第 52 条（2）需告知情绪识别或生物识别系统的使用，无需获取个人数据处理前的同意	第 52 条（2）使用者需要告知情绪识别或生物识别系统的使用，且需在处理个人数据前获得同意	
第 52 条（3）用户须公开内容已被 AI 生成或操纵，未提及需揭示不真实性和生成 / 操纵者名称	第 52 条（3）用户须明确公开内容已被 AI 生成或操纵，揭示不真实性及生成 / 操纵者名称	
	第 52 条（3a）对创新、讽刺、艺术或虚构作品，透明度义务限于披露生成 / 操纵内容的存在及版权	
	第 52 条（3b）应在首次交互 / 接触时或前提供信息，考虑残疾人或儿童的可获取性，包含暴露在系统下者的干预或标记程序	

“透明度”是指人工智能系统的开发和使用方式应做到具有一定的可追溯性和可解释性。关于透明度的要求，2023 年《AI 法案》折衷草案对透明度的定义更具体，要求提供者和用户都能理解人工智能系统的运行方式。

2023 年《AI 法案》折衷草案明确规定，用户应知道人工智能系统如何运行以及它处理了哪些数据，以便向受影响的人解释人工智能系统做出的决定。在 2022 年《AI 法案》妥协版本中，要求披露的信息主要涉及提供者的身份和联系方式，以及人工智能系统的性能特征、能力和限制。而折衷草案要求的信息更为详尽，包括进行合规性评估的实体的身份和联系方式，以及任何可能影响系统性能的用户行为等。此外，折衷草案中新增了一条规定，即为了遵守法案中规定的义务，提供者和用户应确保有足够的人工智能素养。

总的来说，2023 年《AI 法案》折衷草案对人工智能系统的交互、披露要求和豁免条件提供了更具体、详细的指导，旨在更好地保护自然人和他们的权利。

ChatGPT 这类人工智能系统的制造商披露训练大模型所使用的版权数据信息。[22] 基础模型的供应商将被要求声明是否使用受版权保护的材料来训练人工智能。对于谷歌和微软等科技公司，若违反规定，罚款可能高达数十亿美元。

5. 人工智能系统人为监督

2022 年《AI 法案》妥协版本	2023 年《AI 法案》折衷草案	变化趋势
第 14 条（1）人工智能系统应该在使用期间由自然人有效监督	第 14 条（1）人工智能系统应该由具有足够 AI 素养的自然人进行与系统风险相称的有效监督。此外，还需要进行事故后的深入调查	人为监督是保障人工智能发展中不会侵犯道德、伦理和法律规定，以及防止对个人和环境产生严重危害的重要手段。同时，人为监督还有助于提高人工智能系统的透明度，提升公众对高风险人工智能系统的信任和接受度，推动 AI 的健康发展。
第 14 条（2）人工监督的目的是预防或最小化高风险 AI 系统使用过程中出现的健康、安全或基本权利风险	第 14 条（2）人工监督的目的不仅包括预防或最小化健康、安全、基本权利风险，还增加了环境风险，并考虑了 AI 系统对个人或团体产生法律或其他重大影响的情况	
第 14 条（3）人工监督应通过一项或所有的措施来确保	第 14 条（3）人工监督应考虑特定的风险，AI 系统的自动化程度和背景，通过一项或所有类型的措施来确保	

[22] https://www.thepaper.cn/newsDetail_forward_23490174.

2022 年《AI 法案》妥协版本	2023 年《AI 法案》折衷草案	变化趋势
第 14 条（4a）负责人工监督的个体需要充分了解高风险 AI 系统的能力和限制，并适当地监控其运行	第 14 条（4a）负责人工监督的自然人需要意识到并充分了解高风险 AI 系统的相关能力和限制，并以适当和相称的方式监控其运行	人为监督是保障人工智能发展中不会侵犯道德、伦理和法律规定，以及防止对个人和环境产生严重危害的重要手段。同时，人为监督还有助于提高人工智能系统的透明度，提升公众对高风险人工智能系统的信任和接受度，推动 AI 的健康发展。
第 14 条（4e）负责人工监督的个体需要能够干预高风险 AI 系统的运行或通过“停止”按钮或类似程序中断系统	第 14 条（4e）负责人工监督的自然人需要能够干预高风险 AI 系统的运行或通过“停止”按钮或类似程序使系统安全停止，除非人为干预会增加风险或对性能产生负面影响	
第 14 条（5）根据附录 III 的 1(a) 点，需要确保基于系统得出的识别结果的用户的任何行动或决定都已经由至少两个自然人核实和确认	第 14 条（5）根据附录 III 的 1(a) 点，需要确保基于系统得出的识别结果的用户的任何行动或决定都已经由至少两个具有必要能力、培训和权威的自然人核实和确认	

2023 年《AI 法案》折衷草案要求，对于高风险人工智能系统，在入市前的设计应当保证能够实施人为干预。

增加对人工智能人为监督的主要原因在于以下几个方面。首先，人工智能（AI）可能会做出涉及道德和伦理问题的决策，比如对个人隐私的侵犯、不公平的偏见和歧视等。人类不能放心地把这些决策交给人工智能，但如果有了人类的监督，则可以在充分利用人工智能的基础上预防相关的道德风险。其次，人为监督是降低人工智能风险的主要措施。如果 AI 产生错误决策甚至有违法行为，由于 AI 无法承担法律责任，责任的承担者只能是人类自己，所以人类需要监督 AI 以保证它们的行为符合法律规定。最后，尽管 AI 在很多领域都有显著的性能，但是它们并不能理解伴随着历史逐步形成的人类社会。AI 必须在人类的监督下，以保证它们能够正确地处理这些复杂的涉及人类社会的问题。

还有一点需要引起人们注意的是，风险较高的人工智能系统，往往需要更高程度的人为监督。高风险的 AI 系统可能在未经恰当处理的情况下，造成重大的个人、环境乃至社会的损害。人类监督可以确保在经验层面使 AI 系统的运行符合人类社会的价值观和道德观，最大降低高风险 AI 对人类社会的伤害。AI 系统的决策过程往往是不透明的，这使得

其决策的有效性和合理性难以判断。人类监督可以要求 AI 系统提供更多的透明度和可解释性，以便理解和质疑其决策。且人为监督也可以促进 AI 的普及与发展。高风险的 AI 系统本身可能会引发公众的恐慌和反感，人类监督能够帮助建立公众的信任，提高 AI 系统的社会接受度。

6. 监管沙盒制度

2023 年《AI 法案》折衷草案	变化趋势
第 53 条（1）-（1b）成员国须设立至少一个 AI 监管沙盒，且可在各级别或与他国合作设立更多沙盒。	《人工智能法案》要求欧盟成员国建立并使用监管沙盒，验证人工智能系统是否符合法规。2023 年的折中草案修改了建立沙盒的规定，并强制成员国按照标准建立。沙盒可跨国合作，保护个人数据仅为公共利益开发目的使用。提供者需承担实验中对第三方造成损害的责任。设立机构需每年向人工智能办公室提交报告，直到沙盒终止。
第 53 条（1c）-（1e）设立机构须确保资源充足，为 AI 系统开发提供创新环境，并设定目标和指导原则。	
第 53 条（1f）-（4）沙盒运作需保证风险识别与缓解，尤其针对基本权利、民主法治、健康安全和环境等领域的风险。此外，对于高风险 AI 系统的开发者，设立机构需要提供如何满足法规要求的指导和监督。最后，如果在沙盒内的实验对第三方造成任何损害，预期提供者将根据适用的联盟和成员国法律负责。	
第 53 条（5）-（5b）设立机构应在 AI 办公室的框架内进行合作，并公开沙盒信息以鼓励互动与跨国合作；设立机构每年向人工智能办公室和委员会提交报告，包括沙盒进展、成果、最佳实践、建议和修订，并公开报告或摘要。	
第 53 条（6）委员会开发 AI 沙盒信息接口，提供非约束性指导和协调服务。	
第 53 条（6a）-（a）委员会帮助成员国建立和运营 AI 沙盒，制定了详细规则和程序。	
第 54 条 提议规定在沙盒中处理数据的条件，及推动 AI 研究以支持社会和环境效益。	

2021 年《AI 法案》提案引入监管沙盒（AI Regulatory Sandbox）制度，要求欧盟成员国在法案生效前至少建立并投入使用一个人工智能监管沙盒，以在相应的人工智能系统投放市场前对其遵守《人工智能法案》的情况进行验证。

对比 2022 年《AI 法案》妥协版本，2023 年《AI 法案》折衷草案将“各成员国自行决定”修改为按照既定的标准强制建立。这种沙盒在覆盖国家层面的基础上，可以与其他一个或几个成员国联合建立，也可以跨成员国建立，以促进跨境合作和协同效应。

此外，在人工智能系统中个人数据是否可以在沙盒的保护下进行处理，折衷草案给出了答案，增加了个人数据用于开发人工智能监管沙盒需为公共利益开发之目的的要求。并且增加了在监管沙盒中，人工智能系统的潜在提供者需按照适用的欧盟和成员国责任立法，对于因在沙盒中进行的实验而对第三方造成的损害承担责任。折衷草案还规定了沙盒的设立机构应当每年向人工智能办公室（AI office）提交年度报告，从沙盒设立一年后开始，每年提交一次，直到沙盒终止。

7. 行政罚款

2022 年《AI 法案》妥协版本	2023 年《AI 法案》折衷草案	变化趋势
第 71 条（3）最高罚款金额为 3000 万欧元，或者是前一财年全球年营业额的 6%	第 71 条（3）最高罚款金额为 4000 万欧元，或者是前一财年全球年营业额的 7%	从 2022 年《AI 法案》妥协版本到 2023 年《AI 法案》折衷草案的变化趋势主要表现在两个方面：一方面，草案明显提高了行政罚款的上限，这显示了立法者对于人工智能实践中的违规行为的严肃态度和强烈惩治意图；另一方面，草案更详细地考虑了侵权行为的性质、严重性和持续时间，同时强调了与欧洲数据保护监察员的合作程度以及如何修正侵权行为，体现出了立法者对于公正、公平处罚及引导合规行为的诉求。
第 71 条（4）最高罚款金额为 2000 万欧元，或者是前一财年全球年营业额的 4%	第 71 条（4）最高罚款金额为 1000 万欧元，或者是前一财年全球年营业额的 2%	
	第 71 条（7）各成员国需要就公共机构和机构的罚款规则进行规定	
	第 71 条（8a）罚款及相关诉讼费用和赔偿要求不能成为提供商、分销商、进口商、部署商或任何其他第三方之间的负担分享协议的主题	
	第 71 条（8b）国家监管机构应每年向 AI 办公室报告其依据本条规定在该年度内已发出的罚款	

2022 年《AI 法案》妥协版本	2023 年《AI 法案》折衷草案	变化趋势
第 72 条（1）（a）违规的性质、严重程度和持续时间以及其后果	第 72 条（1）（a）违规的性质、严重程度和持续时间以及其后果，并考虑到相关 AI 系统的目的、受影响人数、他们所受的损害程度，以及任何相关的先前违规行为	从 2022 年《AI 法案》妥协版本到 2023 年《AI 法案》折衷草案的变化趋势主要表现在两个方面：一方面，草案明显提高了行政罚款的上限，这显示了立法者对于人工智能实践中的违规行为的严肃态度和强烈惩治意图；另一方面，草案更详细地考虑了侵权行为的性质、严重性和持续时间，同时强调了与欧洲数据保护监察员的合作程度以及如何修正侵权行为，体现出了立法者对于公正、公平处罚及引导合规行为的诉求。
第 72 条（2）以下的违规行为将被处以最高 500 000 欧元的行政罚款	第 72 条（2）不遵守第 5 条所提到的人工智能实践禁令将被处以最高 1 500 000 欧元的行政罚款	
	第 72 条（2a）AI 系统不符合第 10 条所规定的要求，将被处以最高 1 000 000 欧元的行政罚款	
第 72 条（3）AI 系统不符合本条例规定的任何要求或义务，除了第 5 条和第 10 条规定的，将被处以最高 250 000 欧元的行政罚款	第 72 条（3）AI 系统不符合本条例规定的任何要求或义务，除了第 5 条和第 10 条规定的，将被处以最高 750 000 欧元的行政罚款	
第 72 条（6）本条款中处以罚款所收集的资金应为欧盟总预算的收入	第 72 条（6）本条款中处以罚款所收集的资金应为欧盟总预算的贡献，罚款不应影响被罚款的欧盟机构、机构或机关的有效运作	
	第 72 条（6a）欧洲数据保护监察员应每年通知 AI 办公室其根据本条款所处的罚款	

对于不同风险等级，2022 年《AI 法案》妥协版本采取了不同程度的监管措施。对于被禁止的人工智能系统，法案严厉禁止，并对违反者处于上一财年全球营业额最高 6% 的处罚。2023 年《AI 法案》折衷草案将最高罚款提高到 4000 万欧元或前一财年全球年度营业额的 7%。

（二）欧盟人工智能法案的立法缺陷分析

1. 责任机制存在缺陷

2021 年《AI 法案》提案对人工智能产品设定市场准入门槛和建立全生命周期的监测

体系，是一种有效的人工智能监管路径，欧盟人工智能法案规定了不同责任主体，根据组织在人工智能供应链中的位置划分了提供者、部署商、进口商、分销商或其他第三方的责任。然而，欧盟人工智能法案的责任机制仍然存在缺陷。

首先，在规制对象上，责任界定不清晰。例如，法案可能没有充分考虑到人工智能系统的自适应性，这可能导致在某些应用场景下的主体责任归属不清晰。此外，法案未明确多方参与场景下多个责任主体间的责任分配，这可能导致责任确定困难[23]。《人工智能法案》使用的术语与 GDPR 中通常理解的数据控制者 / 处理者不同，这可能会导致合规性问题。根据《人工智能法案》，参与者包括提供者、部署商、进口商和分销商，因此开发或销售人工智能系统组织内的负责机构可能并未能明确对应 GDPR 下数据控制者 / 处理者角色，有可能会导致在两部法案同时适用的情况下合规责任不清晰。

在责任承担方式上，《人工智能法案》在责任承担方式上较为单一，处罚方式主要是罚款，责任体系仍需进一步完善。

2. 阻碍产业发展

欧盟在数字立法领域一直走在世界前列，此次欧盟人工智能法案引起不少质疑和担忧，有观点认为欧盟《人工智能法案》将使人工智能企业在欧洲承担过高的成本，而且大部分合规要求在技术上无法实现。有反对者担心监管过度可能给欧洲人工智能公司带来过高的成本和过重的负担。基于此，OpenAI 公司首席执行官萨姆·阿尔特曼表示，如果《人工智能法案》对人工智能进行过度监管，他将带领团队撤出欧洲市场[24]。欧洲议会议员 Brando Benifei 表示欧洲的人工智能立法已超前地对人工智能的风险给予具体回应。

尽管欧盟《人工智能法案》的初衷是保护公共利益并确保人工智能的安全和透明，但其过于严格的监管措施可能会阻碍产业发展。例如，过高的合规成本和严格的数据保护规定可能会影响创新和竞争力。在 2020 年的一项针对使用人工智能技术的欧洲企业的调查显示，对有计划但尚未采用人工智能技术的欧洲公司来说，43% 的欧洲企业将法律责任问题视为其使用人工智能技术的最大外部障碍之一[25]。欧盟的超前和过度监管可能会进一步强化欧洲人工智能产业落后的局面，诸多事前监管环节都将影响人工智能的开发进程。尽管欧盟委员会预测行业里大部分人工智能应用将落入低风险类别中，因而认为法案对人工智能产业发展的负面效果有限，但无论如何，诸多事前监管环节都将影响人工智能的开发进程。

---

[23] 曾雄、梁正、张辉：《欧盟人工智能的规制路径及其对我国的启示——以 < 人工智能法案 > 为分析对象》，载《电子政务》2022 年第 9 期，第 67 页。
[24] 王卫：《抢占全球人工智能监管先机，欧盟 < 人工智能法案 > 进入最终谈判阶段》。
[25] 王雪稚：欧盟《人工智能法案》立法及人工智能监管进展综述。https://www.ctils.com/articles/9349。

另一方面，2023 年《AI 法案》折衷草案旨在在欧盟层面为人工智能系统的开发、投放市场、投入使用制定统一的法律框架，并确保基于人工智能的商品和服务的跨境自由流动，除非《人工智能法案》明确授权，成员国不得对人工智能系统的开发、销售和使用施加限制。

总体而言，《人工智能法案》属于偏规制型的立法，其对产业促进方面的规定较少。当前人工智能技术发展仍处在探索期，如何更好地平衡与衔接技术发展和监管创新，还需在实践中不断探索和完善。

三、欧盟人工智能监管体系对中国的启示

（一）中国人工智能监管框架和发展进程

我国一直关注人工智能技术与产业的发展，早在 2017 年 7 月，国务院发布了《关于印发新一代人工智能发展规划的通知》（“《通知》”）。《通知》高度总结了人工智能的发展现状，以及中国在应对人工智能的发展应采取怎样的战略态势，既强调了人工智能对我国提高国际竞争力、促进社会经济发展的意义，但同时也提及了其对政府管理、经济安全和社会稳定乃至全球治理带来的不确定性因素。为应对这一情形，《通知》在顶层设计上采取“三步走”战略，其中第二步战略（2020 年 -2025 年）要求“初步建立人工智能法律法规、伦理规范和政策体系，形成人工智能安全评估和管控能力”。人工智能的法律法规和伦理规范作为人工智能发展的保障措施在《通知》中得到确认。

目前，我国在人工智能领域的法律规制工作正在稳步推进中。其中，作为整体规制的基础性法律，人工智能法草案已被列入国务院 2023 年立法工作计划，提请全国人大常委会审议。针对特定运用领域，比如推荐算法、深度合成等技术 / 产业，我国在立法上已经做出了初步的尝试。除此之外，数据作为人工智能发展的三驾马车之一，我国近年来的与其相关的《数据安全法》《个人信息保护法》等数据法律也在数据处理方面适用于人工智能。另外，对于科技伦理的有关规定也同样适用于人工智能。

表 2 中国的人工智能监管及算法治理规则一览表

文件名称	时间	制定部门	文件性质	主要内容
《数据安全法》	2021.09.01	全国人大常委会	法律	规范数据处理活动，国家要统筹数据安全与发展，建立数据保护制度以及开放政务数据，对数据处理者提出一系列数据处理安全保护义务。

文件名称	时间	制定部门	文件性质	主要内容
《个人信息保护法》	2021.11.01	全国人大常委会	法律	规范个人信息处理活动，提出了处理个人信息的基本原则和基本规则，确认了个人信息权利和个人信息处理者义务。
《工业和信息化领域数据安全管理办法（试行）》	2023.01.01	工业和信息化部	部门规章	规范工业和信息化领域的数据处理活动，对工业和信息化领域数据处理者在确立数据分类分级的基础上，提出了数据全生命周期的数据安全管理要求。
《互联网信息服务算法推荐管理规定》	2022.03.01	国家互联网信息办公室等四部门	部门规章	规范互联网信息服务算法推荐活动，对服务提供者提出了安全管理、内容审核等一系列义务要求，落实算法备案，同时根据群体特殊性规定了一系列特殊义务。
《互联网信息服务深度合成管理规定》	2023.01.10	国家互联网信息办公室等三部门	部门规章	规范互联网信息服务深度合成活动，对深度合成服务提供者提出了一系列义务要求，加强了数据和技术管理规范，落实提供者安全主体责任。
《生成式人工智能服务管理暂行办法》	2023.08.15	国家互联网信息办公室等七部门	部门规章	规范提供生成式人工智能产品或服务行为，对提供者算法训练义务、信息内容管理义务、用户管理等相关义务进行了规定。
《新一代人工智能伦理规范》	2021.09.25	国家新一代人工智能治理专业委员会	部门规范性文件	对人工智能管理、研发、供应、使用全生命周期提出具体的伦理道德要求。
《科技伦理审查办法（试行）》	2023.04.04	科学技术部	部门规章（征求意见）	对科技活动提出事前、事中和事后的伦理审查要求，覆盖涉及人的科技活动以及可能带来伦理风险挑战的科技活动，对伦理审查委员会的组织以及审查程序作出了一系列规定。

1. 依据主体的治理范式与依据风险的治理范式

中国与欧盟关于人工智能监管的体系框架表明，两者对人工智能进行规制的出发点有所不同。中国针对不同的涉及算法的互联网信息服务，以落实主体责任作为基本落脚点。实际上，我国针对特定人工智能产品或服务的规定，基本上是将“服务提供者”作为相关义务的履行主体。《互联网信息服务算法推荐管理规定》的义务主体是“算法推荐服务提供者”，《生成式人工智能服务管理办法》的义务主体是“生成式人工智能服务提供者”，《互联网信息服务深度合成管理规定》则明确形成了多义务主体规制模式，包括深度合成服务提供者、深度合成服务技术支持者以及深度合成服务使用者，但从具体条文来看，深度合成服务提供者、深度合成服务技术支持者是该《规定》最主要的规制对象。

欧盟《人工智能法案》则是首先确立以风险为基准的人工智能治理框架。通过对人工智能系统进行评估，人工智能系统将被划分为不可接受风险、高风险、有限风险和最小风险四个层级，并匹配了不同的责任措施和差异化的监管[26]。在风险分级之下，2023 年《AI 法案》折衷草案进一步界定了提供者、授权代表、分销商、进口商、部署商等主体（根据 Art3.1(8)，统称为“经营者”），明确在不同风险的人工智能系统统一的责任措施和差异化监管之下，各类经营者在其中具体应该承担何种义务。

2. 在人工智能系统特殊领域的特殊监管达成共识

算法推荐，深度合成，生成式人工智能是我国规制人工智能的具体领域。虽然 2023 年《AI 法案》折衷草案适用于所有的人工智能系统，但就上述某些产品或服务而言进行了特殊的回应。这些回应一定程度上印证了我国对这些领域进行特别监管的必要性。

面对深度合成，2023 年《AI 法案》折衷草案在第 52 条第 3 款第 1 项确定了深度合成的标记义务，使用深度合成时，不仅要告知接收者内容为使用深度合成系统下的输出物，同时要标记具体使用深度合成技术的主体的信息。这与《互联网信息服务深度合成管理规定》第 16 至 18 条所规定的标记义务保持一致，但相比之下 2023 年《AI 法案》折衷草案更加强化了系统使用主体信息透明度的要求，一定程度上将法律责任分配给了使用深度合成技术的主体。

对于生成式人工智能，2023 年《AI 法案》折衷草案将其视为“基础模型”的一种类型，在第 28b 条对基础模型所施加的一系列一般义务之外，还通过该条第 4 款规定了针对生成式人工智能的额外义务。具体而言包括第 52 条第 1 款所规定的透明性义务、在保障言论自由等基本权利的同时防止违法内容生成，以及尊重知识产权的要求。

---

[26] 张欣：《生成式人工智能的算法治理挑战与治理型监管》，载《现代法学》，2023 年第 3 期，第 108-123 页。

欧盟对于深度合成和生成式人工智能的特殊规定仅占据整部法案的极小篇幅，从规定内容上来看虽然抓住了规制重点但并不详尽。相较而言，我国的两个相关规定在义务上更为全面。不过由于缺少统一的横向立法，导致我国的相关规定存在大量重合，可能会导致规定交叉重复适用的问题。

（二）人工智能技术发展实践所需监管补位的难点和痛点

1. 供应链中不同经济运营商之间责任分配的不确定性

人工智能系统从研发到投放市场涉及多个主体，特别是当委托代理或授权关系进行介入的情况下主体之间的关系将更为复杂。就我国的相关具体人工智能规范而言，服务提供者往往是主要的责任主体。在《人工智能法案》中，人工智能系统供应链的参与主体更为细化，具体包括提供者、部署商、授权代表、进口商和分发商，它们被统称为“运营者”。

2023 年《AI 法案》折衷草案在法律义务分配设计上，特别是对于高风险人工智能系统，提供者，其次是部署商，将承担主要的义务（Art16）。其中，提供者将承担最广泛的合规义务，包括建立风险管理制度和质量管理制度等，涵盖人工智能系统生命周期的事前和事后环节。而部署商的义务则主要集中于确保对高风险人工智能系统的人工监督和日常检测义务，主要覆盖人工智能生命周期的事中环节（Art29）。

2023 年《AI 法案》折衷草案要求进口商采取一定措施，确保所进口的人工智能系统投入市场之前已经获取了折衷草案对提供者所要求履行的一切程序及其文件，包括评估程序、技术文件以及 CE 标志（Art26）。

另外，相关责任主体的界定是以“特定”的人工智能系统参照进行界定的。在人工智能系统供应链上，当非提供者的其他责任主体对提供者的人工智能系统进行实质性修改而使其成为高风险系统，或将基本模型嵌入到高风险系统中，或以自己的名字或商标放在已经投放市场或投入使用的高风险人工智能系统的情形下，其他责任主体将会被认定为新的提供者（Art28）。

《生成式人工智能服务管理办法（征求意见稿）》出台之后，对生成式人工智能服务提供者的义务体系设计引起了一定的讨论。如果对责任主体的界定过于简化，人工智能供应链上的相关主体都可能被要求承担同样的义务，进而也有可能引起责任主体内部难以分配责任的问题。针对此，国内有的观点建议区分基础模型开发者和利用者，原则上由利用者承担内容生产者的责任和算法备案义务[27]。另外，也有观点认为生成式人工智能产品应该秉承“各负其责”的原则，分配权利义务——服务提供者承担服务提供者义务，生成内

[27] https://mp.weixin.qq.com/s/a6txd2WvCJvNWUC12plBMA.

容使用者（用户）承担使用者义务[28]。我们理解，我国在尝试进行进一步的人工智能立法上，应当充分对人工智能开发及使用的各相关主体纳入考虑范围，合理分配各自应承担的法律责任，避免单一主体承担过重乃至于全部的相关责任，从而导致抑制创新或对产业链形成不良后果。实际上，正式生效的《生成式人工智能服务管理办法》也一定程度上回应了这个问题，通过第二条第二款和第三款的规定，一些主体被排除在了《办法》的适用范围之外[29]。

2. 通用型人工智能的监管问题

在以风险为导向的人工智能治理框架中，最容易受到的挑战是无法应对一些跨应用场景和不具有特定使用目的的人工智能系统。最初的 2021 年《AI 法案》提案并未对这类人工智能技术有所回应，这类人工智能系统往往也难以进行所谓的“风险评估”。因为其风险取决于部署商如何使用该系统。这一缺陷在 ChatGPT 等生成式人工智能涌现之后被发现并引起关注。

2023 年《AI 法案》折衷草案则对这一问题进行了回应。这类系统在 2023 年《AI 法案》折衷草案中被称为“基本模型”，是指在广泛的数据上进行规模化训练的人工智能模型，其设计是为了实现输出的通用性，并能适应各种不同的任务。基本模型可以独立使用（例如生成式人工智能），也可以成为其他人工智能系统的组件。针对“基本模型”的泛用性特征，2023 年《AI 法案》折衷草案对这类模型的相关义务单独设计条款专门规定，与以风险作为分类标准的人工智能系统的相关义务相互区别。这有效解决了这类技术无法纳入 AIA 最初风险治理框架的问题。

（三）对中国人工智能监管治理框架的建议

1. 纳入道德伦理和人权考量的以风险为基准的统一人工智能治理框架

欧盟以风险为基准的人工智能治理框架在一定程度上值得我国人工智能领域一般性立法进行参考和借鉴。

在一般性立法方面，人工智能的治理既有共性，但在具体的特殊领域的运用又具有特性，这是我们需要兼具横向立法和纵向立法的原因。缺乏统一的人工智能立法，将会导致不同特殊领域的立法在内容上高度重合，容易引起法律之间适用的困惑，也导致法律规定的重复和冗杂。提炼出人工智能治理的共性，将其作为一般性人工智能立法的规定内容，将有助于解决这一问题。

此外，伦理道德和人权向来都是人工智能技术发展绕不开的终极话题。其高度概括性、

[28] https://mp.weixin.qq.com/s/NCT-9LGEuJcJiBcXXYbVNA.
[29] https://www.kwm.com/cn/zh/insights/latest-thinking/china-first-interim-regulatory-measure-on-aigc.html.

抽象性和不确定性，导致如何将这部分内容融入人工智能治理考验着立法者的立法技术。而以风险为基准的人工智能治理框架提供了其中一种解决方案。《人工智能法案》将人工智能系统对伦理道德和基本人权的影响有机纳入规制框架和评估框架中，对相关责任主体的义务配置和履行起着决定性的作用。实际上，我国《科技伦理审查办法 (试行)》也是将伦理纳入包括人工智能开发在内的科技活动的积极探索。伦理审查委员会的审查结果也能够阻止违反道德伦理科技活动的开展。不过由于该《办法》适用于所有的科技活动，人工智能的特殊性可能无从得到体现。另外，应如何界定和解释“不可接受的风险”和“高风险”的人工智能系统也面临相当大的不确定性和模糊性 [30]。人工智能系统会带来的问题，具体取决于使用它们的人、地点和目的。在一定程度上可能难以统一进行风险分类 [31]。《人工智能法案》虽然通过举例的方式帮助解释和澄清，但仍可能无法应对快速发展变化的人工智能系统。因而以风险为基准的人工智能管理框架究竟成效如何仍有待进一步的观望和研讨。

2. 人工智能法律与现有数据保护法律，特别是个人信息保护法制度的衔接性

人工智能的研发和部署使用离不开对数据，尤其是个人信息的处理。但我国《个人信息保护法》的规定可能会对上述活动形成一定的合规障碍。在《个信法》第 13 条所规定的各类个人信息处理的合法性中，仅有个人同意和基于合同履行所必须可以作为使用人工智能进行个人信息处理的合法性基础。诚然，就人工智能的部署和使用而言，上述合法性基础的获取并不存在特殊的障碍。但如果在研究开发环节也要求获取上述合法性，则会极大增加相关责任主体的合规成本。

虽然欧盟《人工智能法案》明确不排除一般情况下对 GDPR 的适用（Art2.5a），但在合法性基础上，GDPR 有更多的合法性基础供个人数据处理者主张。例如“控制者和其他第三方的正当利益”以及“为了实现公共利益、科学或历史研究或统计目的处理中的处理”。

相较之下，《人工智能法案》在明确不影响 GDPR 的实施之下，在具体规定中对涉及个人数据的处理进行了解释和衔接。我国若计划进行统一的人工智能立法，那么个人数据处理的合法性问题将无法回避。比较好的方法当然是在人工智能立法中就合法性进行特别规定，以适用《个信法》第 13 条第 1 款第 7 项“法律、行政法规规定的其他情形”，为人工智能系统对个人信息处理提供额外的合法性基础。

3. 监管沙盒充当人工智能系统投入市场的“守门人”

监管沙盒是《人工智能法案》中所使用的，确保人工智能系统合规的重要措施。“监

[30] https://arxiv.org/ftp/arxiv/papers/2107/2107.03721.pdf.
[31] https://link.springer.com/article/10.1007/s12027-022-00725-6.

管沙盒”是指由公共机构建立的，在创新人工智能系统投放市场或根据监管部门监督的具体计划投入使用之前，在有限的时间内为其安全开发、测试和验证提供便利的受控环境。它通过监管创建一个受控制的环境对人工智能系统进行测试，并给予合规指导。这一监管措施将为人工智能的合规监管带来两层意义：其一是能够帮助监管者实现了解人工智能系统的全貌并收集相关信息，消除信息壁垒，防止人工智能的复杂性而引起的滞后性，促进敏捷治理；其二是能够帮助提供者评估人工智能系统的现实运作情况，从而进行针对性的完善并获取专业的监管指导，有效降低合规成本[32]。同时，2023 年《AI 法案》折衷草案中的监管沙盒制度，还为个人数据的处理提供了额外的合法性基础，其第 54 条规定，在符合特定条件的情况下，在人工智能监管沙盒中，为其他目的合法收集的个人数据可仅为在沙盒中开发和测试某些人工智能系统的目的而处理。

我们理解，监管沙盒作为开发环境和现实环境之间的缓冲带，能够非常好地帮助提供者评估人工智能系统的合规水平，并决定是否投入市场。监管沙盒一定程度上也能够作为信用背书，帮助提供者论证其合规能力。

4. 对中小企业的兼顾激励与监管的制度体系

在统一式的监管措施面前，虽然合规要求并无差距，但中小企业往往面临着难以承担的巨大合规成本。如果不采取适当的措施对中小企业进行保障，大型企业往往可以凭借自己的资源优势在实现技术创新的同时完成合规要求，而中小企业则只能在创新与合规中艰难抉择，这无疑可能会加剧大型企业的垄断局面。欧盟的数据立法向来意识到这一点，从 GDPR 到《数字服务法》，都采取了相应的措施适度降低中小企业的合规成本。《人工智能法案》也不例外。2023 年《AI 法案》折衷草案第一条进一步明确了要采取监管沙盒等措施降低中小企业的合规成本，促进科技创新。具体而言，通过第 28a 条制约单方面强加给中小企业和初创企业的不公平合同条款，通过采取规制格式合同的方式，一些显著不公平的条款将被视为无效约定，防止大型企业利用自身优势转嫁法律规定下本应自己承担的法律风险。在监管沙盒方面，2023 年《AI 法案》折衷草案第 53a 条特别提出应促进监管沙盒广泛而平等的参与，并减免参加费用和提供部署前服务和其他增值服务。最后，2023 年《AI 法案》折衷草案还通过适当降低中小企业的评估费用或其他合规要求，在处罚规定中要求将纳入中小企业的利益和经济活力，以实现降低合规成本的最终目标。2023 年《AI 法案》折衷草案还进一步吸纳了来自欧洲数字中小企业联盟的建议，要求相关准则的制定需要充分考虑中小企业的需求，以确保对其的合规负担不会因为中小企业参

[32] 毕文轩：《生成式人工智能的风险规制困境及其化解：以 ChatGPT 的规制为视角》，载《比较法研究》，2023 年第 3 期第 155-172 页。

与缺位而导致过重 [33]。

相较而言，我国的法律制度更多是在一般性规定的基础之上，强化对大型企业的监管。例如，就《个人信息保护法》而言，关键信息基础设施运营者和大型个人信息处理者将面临更多的合规义务，执行个人信息出境时也面临更严格的要求；而对于小型个人信息处理者的保障仅出现在第 62 条，属于个人信息保护工作推进内容之一，但目前尚未有具体规定。而专门针对人工智能领域制定的，针对算法推荐、深度合成、以及生成式人工智能的办法，均没有对中小型的服务提供者制定专门的规定，以控制其合规成本。

从防止垄断，促进人工智能技术创新的角度而言，欧盟的《人工智能法案》顾及到了中小企业在当中的弱势地位。我们认为，适当地将对中小企业的合规义务豁免规定以及合规支持规定纳入到未来的人工智能立法中，将有利于形成人工智能领域健康有序的公平竞争秩序，有效激发中小企业的科技创新活力，同时也能够在制度上有效防止过度监管，避免“放过老虎抓苍蝇”的行为。

以上，我们对欧盟整体的人工智能治理框架及 2023 年《AI 法案》折衷草案的重点变化进行了梳理。可以看到，欧盟对于人工智能的治理整体上由分散不断趋于统一。同时，也逐渐重视欧盟内各项法律法规的衔接问题，以期构建欧洲数据治理的整体格局。接下来，我们将继续分析《人工智能法案》的重点制度，并通过对比分析，为我国的人工智能治理的具体路径提供可行思路。

感谢实习生苏琦对本文作出的贡献。

[33] https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/F266557.

路未央，花已遍芳——欧盟《人工智能法案》主要监管及激励措施评述

宁宣凤 吴涵 吴舸 崔月柳 刘畅

如本书中《全球人工智能治理大变局之欧盟人工智能治理监管框架评述及启示》一文所述，欧盟《人工智能法案》（以下简称“《AI 法案》”）以风险为进路，将人工智能（以下简称“AI”）系统分为不可接受的风险、高风险、有限风险以及极低风险四类。针对每一类 AI 系统，《AI 法案》采取了不同的监管策略，并配以具体的制度使得监管可执行落地。具体而言，针对不可接受风险的 AI 系统，欧盟禁止任何企业或个人部署。针对高风险 AI 系统，欧盟允许相关主体在履行事前评估等义务后投放市场或投入使用，并进行事中、事后的持续监测。针对有限风险的 AI 系统，欧盟虽然不要求相应主体履行事前评估等义务，但其仍需遵循相应的透明度义务。针对极低风险 AI 系统，相应主体则可依据自由意志部署和使用。

总体而言，AI 系统的监管不能以牺牲或放弃 AI 技术发展为代价。实践中，为了向市场提供符合规定的高风险 AI 系统，企业一般需要承担极高的费用或负担，而中小企业基于其财力、市场份额、能力等的多重限制，可能无法实现相应的要求。因此《AI 法案》还提出了 AI 监管沙盒等策略，鼓励各企业尤其是中小企业的发展和创新。

需要提示的是，如本书中《全球人工智能治理大变局之欧盟人工智能治理监管框架评述及启示》一文所述，欧盟成员国、欧盟理事会及议会已于 2023 年 12 月 9 日达成有关欧盟《人工智能法案》的临时协议，且欧洲议会已于 2024 年 3 月 13 日以压倒性结果通过了《人工智能法案》[1]。本文成文时间较早，主要以 2023 年 6 月及 6 月前已公布的《人工智能法案》各版本为撰写基础，而最新公布的《人工智能法案》在条文序号、内容上均可能发生调整或更新，因此，如本文所涉条款与欧盟发布的最新《人工智能法案》文本存在区别，请以最新规定为准，本文不再一一赘述。我们理解，尽管最新的法案文本可能在

[1] 具体详见本书中《历时三年，欧盟 < 人工智能法案 > 通过欧洲议会表决》”一文。

规制内容上进行了调整，并修订了条款序号，但是，整体而言，欧盟仍然秉持“以风险为进路”的监管思路。本文承接《全球人工智能治理大变局之欧盟人工智能治理监管框架评述及启示》一文，以期为读者勾画出有关欧盟《AI 法案》整体监管机制，探寻我国的人工智能监管立法方向。

一、纵观：欧盟《人工智能法案》的监管思路

自欧盟 2021 年《AI 法案》提案起，其解释备忘录即对法案的监管思路进行了整体性描述，其中提到，“本提案对 AI 提出了平衡、相称的横向监管方法”，仅限于“与 AI 相关的风险和问题的最低必要要求”。并且，“建立了一个以明确界定的基于风险的监管方法为基础的相称的监管制度……同时，法律框架包括灵活的机制，使其能够随着技术的发展和有关新情况的出现而动态调整。”可以看出，欧盟《AI 法案》的监管思路是以横向监管为基础，风险规制为主要方式，同时兼顾监管与发展的动态平衡，具体而言：

（一）横向监管为基础

目前，在 AI 监管的相关法律文件中，主要存在“横向”监管与“纵向”监管两种主要方式。在横向监管方式中，监管机构将创建一个全面的法规，以尽可能涵盖 AI 可能产生的各方面影响；在纵向监管方式中，政策制定者采取“定制”的方法，针对不同应用或类型的 AI 制定不同的法规[2]。欧盟《AI 法案》则采用横向监管模式，具体而言，该法案以风险分级的方式将所有 AI 系统纳入监管范围（特殊 AI 系统除外），并允许监管机构随着 AI 的发展不断将新的应用纳入现有的风险类别，而没有针对特定 AI 应用领域制定具体的法律规范。风险分级方式使法案整体处于相对灵活的状态，既能够保持横向监管方式具有的统一性和协调性，同时，相对灵活的分类标准也弥补了传统横向监管方式下对具体的 AI 应用场景针对性不高的问题，兼具法律的确定性与灵活性，使得相关监管措施更易落地。

（二）风险规制为主要方式

欧盟《AI 法案》采用分类分级的风险规制路径，在规定统一监管框架的基础上，识别和评估 AI 系统可能引发的风险。法案将 AI 系统分成四个风险级别，分别是不可接受的风险、高风险、有限风险和极低风险，每个风险类别都有相应的应用场景和监管措施，具体分类情形如下表所示：

---

[2] 参见《全球两类人工智能治理实践的教训》，http://chinawto.mofcom.gov.cn/article/br/bs/202303/20230303399532.shtml，最后访问时间：2023 年 7 月 17 日。

风险级别	应用场景	监管措施
不可接受的风险	• 采用潜意识技术或有目的的操纵或欺骗技术，明显损害或实质性扭曲人的行为和能力的系统 • 利用个人或社会群体的弱点或已知的人格特征或社会经济状况用于损害或实质性扭曲该人或该群体的系统 • 利用人的社会行为或人格特征进行社会评分并对个人造成有害后果或歧视性待遇的系统 • 在公众场所的“实时”远程生物识别系统	• 完全禁止
高风险	1) 在欧盟统一立法规制范围内作为产品的安全组件或本身属于产品且其需要第三方符合性评估机构开展评估的 AI 系统 2) 符合分级标准且用于下述领域的独立 AI 系统： • 生物识别和分类 • 可能威胁人的生命和健康的关键基础设施 • 可能决定人的受教育机会教育或职业培训 • 就业、员工管理和获得自营职业 • 获得和享受基本的私人和公共服务 • 可能影响人的基本权利的执法活动 • 移民、庇护和边境控制管理 • 司法和民主程序	• 投放市场前，受到严格管控 • 履行符合性评估的要求 • 使用高质量训练、验证和测试的数据集 • 对活动记录确保可追溯 ……
有限风险	• 与人类互动的系统 • 情绪识别系统 • 生物特征分类系统 • 生成或操纵内容系统	• 透明义务要求
极低风险	• 允许自由使用 AI 的电子游戏、垃圾邮件过滤器等系统	• 不作干预

以上，不同风险级别的 AI 系统对应不同的监管措施，风险等级越高意味着监管措施越严。灵活的分类方式基本涵盖了目前绝大多数 AI 系统，并且该框架允许随着 AI 技术的发展不断更新和补充。该分类方法结合了不同应用场景的具体情况并对应设置针对性的监管措施，一方面避免笼统规范化的合规义务带来的规制失焦问题，另一方面，也为存在多种风险的复杂场景提供了较为清晰的解决思路，例如可以通过风险分级方式对应复杂场景中的不同类型的风险。但是，过于严格的风险分级方式也有可能对 AI 的创新发展造成一定限制，例如存在一些新兴的 AI 系统可能会被归类于高风险从而影响人们对于新兴 AI 技术探索的情形。

然而，我们理解，实践中各行业企业在某一具体场景中可能不仅适用一个 AI 系统，而是采取多 AI 系统耦合的方式，共同为客户提供服务。共同参与某个具体场景的多 AI 系统可能分属于不同的风险级别，此时，有必要为有机结合共同对外提供服务的 AI 系统重新划定风险级别。具体而言，当较高风险与较低风险的 AI 系统同时存在时，可以采用“就高不就低”的原则，对风险进行较严的监管规制。

（三）兼顾监管与发展的动态平衡

法案同时引入了多种灵活措施和例外情形用于促进 AI 技术的开发创新，在保障监管要求的同时，兼顾 AI 技术的发展。具体而言，《AI 法案》三个版本均提到 AI 监管沙盒，该措施可以保障企业在一个“安全空间”内测试创新性的 AI 系统，从而实现 AI 系统的开发、测试和验证。另外，法案也注重对知识产权的保护，例如 2023 年《AI 法案》折衷草案序言第（79）条规定：“国家监督机构应将获得的任何信息，包括源代码、软件和数据 (如适用)，作为机密信息处理，并尊重欧盟关于保护知识产权和商业秘密的相关法律。”2023 年《AI 法案》折衷草案序言第（83）条规定：“参与本条例应用的所有各方应以透明和公开为目标，同时尊重在执行任务时获得的信息和数据的保密性，制定技术和组织措施，以保护其活动中获得的信息的安全和保密性，包括知识产权和公共及国家安全利益。”保护知识产权即为保护创新，保护创新是推动发展进步的重要动力，由此可以看出，《AI 法案》力图在完善监管措施的同时，基于对个人基本权利的保障，推动实现在 AI 技术可信赖和 AI 技术发展之间的动态平衡。

二、评述：各风险级别人工智能系统的主要监管制度

（一）高风险人工智能系统的主要监管制度

欧盟《AI 法案》采取统一监管策略，主要体现在以下三个方面：责任主体统一监管、各行业统一监管以及 AI 系统全生命周期统一监管。

各责任主体统一监管	各行业统一监管	系统全生命周期统一监管
《AI 法案》对高风险 AI 系统价值链中各主要参与者（具体可参见下表）均设计了相应的义务，义务主体主要包括以下四类：AI 系统提供者，AI 系统部署者，AI 系统进口商以及分销商。 其中，某类参与者是否承担以及应承担何种义务，因其所涉及的 AI 系统风险等级和对 AI 系统的参与深度而异。基于此，**高风险 AI 系统提供者将承担最为全面和严格的义务。**	《AI 法案》适用于使用高风险 AI 系统的各行业主体。《AI 法案》作为一项广泛意义上预防 AI 产品系统性风险的立法，并不仅仅针对某个或某些特定领域，而是 AI 监管与治理的全领域通用性立法。 但是，为科学研究目的开发的 AI 系统将被排除在规制范围外。	高风险 AI 系统在包含投放市场或投入使用之前在内的整个生命周期中均需接受审查，其监管措施包括强制性的风险管理体系、严格的数据和数据治理要求、技术文档和记录保存要求，以及上市后监控和事件报告要求等。

首先，在各责任主体统一监管方面，《AI 法案》将在 AI 供应链的各个层面为相关主体施加一系列义务，包括 AI 系统的提供者、部署者、进口商和分销商。

AI 供应链责任主体	含义	《AI 法案》对主体的责任分配情况
AI 系统提供者	**开发或拥有已经开发的 AI 系统，以自己的名义或商标将其投放市场或在欧盟投入服务的自然人或法人（无论是付费还是免费）；** **值得注意的是，在特定情形下，部署者、进口商和分销商将被视为高风险 AI 系统的提供者，如以自身名义或商标将高风险 AI 系统投放市场或投入使用、投放市场后修改高风险 AI 系统的预期用途、对高风险 AI 系统进行重大修改等。**	在提供**高风险 AI 系统**及**有限风险 AI 系统**时均承担《AI 法案》项下的责任
AI 系统部署者	**在欧盟境内在其权限 / 被授权范围内使用 AI 系统的自然人或法人（不包括在个人非专业活动过程中），可理解为 AI 系统的正式用户。**	在使用**高风险 AI 系统**及**有限风险 AI 系统**时均承担《AI 法案》项下的责任
AI 系统进口商	**在欧盟设立 / 在欧盟境内，并将带有欧盟境外自然人或法人名称或商标的 AI 系统投放到欧盟市场或投入服务的自然人或法人。**	仅在进口欧盟境外的**高风险 AI 系统**时承担《AI 法案》项下的责任
AI 系统分销商	**供应链中提供者和进口商之外的在欧盟市场中提供 AI 系统且不改变其系统属性的自然人或法人。**	仅在分销高风险 AI 系统时承担《AI 法案》项下的责任

下文将主要讨论《AI 法案》为高风险 AI 系统提供者在 AI 系统全生命周期各阶段设置的义务。依据《AI 法案》对提供者的义务履行时间限制，我们将提供者在《AI 法案》框架下应履行的合规义务细分为提供者将高风险 AI 系统投放市场前、投放市场时以及投放市场后三个阶段。

义务主体	系统全生命周期各阶段	具体义务与责任
高风险 AI 系统提供者	投放市场前	• 确保高风险AI系统符合合规要求的义务（包含构建风险管理体系，保障训练数据质量要求，为 AI 系统制定技术文档及记录的要求等） • 构建质量管理体系的义务 ……
	投放市场时	• 履行符合性评估的要求 • 完成 AI 系统注册的要求 • 获取 CE 标识的要求 ……
	投放市场后	• 部署风险监测系统的义务 • 采取纠正措施的义务（产品召回） • 配合监管部门开展工作的义务 ……

由于篇幅所限，在下文中我们将重点介绍高风险 AI 系统提供者在全生命周期各阶段的重难点合规要求。

1. 投放市场前

基于《AI 法案》的整体性要求，一方面，投放市场或投入使用的高风险 AI 系统需遵循一定的技术处理要求，例如在开发设计时需配置系统运行日志的功能。另一方面，高风险 AI 系统提供者自身在将相关系统投放市场前，也需履行一定的合规义务，从而为判断 AI 系统是否符合相关规定的要求，以及提供者证明自身是否合规提供可行的路径。

（1）确保高风险 AI 系统符合相关技术要求的义务

《AI 法案》第二章规定了投放市场或投入使用的高风险 AI 系统应具备的要件，可进一步细化为基于风险动态变化形成的风险管理体系需求，基于核实验证目的形成的技术文

档需求，以及基于减少数据训练引发的算法偏见与歧视等问题形成的数据质量需求。而《AI 法案》第三章则规定，在相关系统投入使用前，提供者需确保其提供的高风险 AI 系统已具有上述能力或履行相关要求。

1）风险管理体系需求

《AI 法案》三次的修订和调整中，具备全生命周期的风险管理体系并进行定期审查与更新始终是《AI 法案》要求高风险 AI 系统具备的能力。

风险管理体系—风险识别步骤	
第一步	识别和分析高风险 AI 系统在预定目的和可合理预见范围内的目的中已知和可预见的风险，考量风险的要点在于 AI 系统对自然人的健康、基本权利等的影响。
第二步	结合为高风险 AI 系统配置的上市后风险监测系统收集的数据，对前述出现的重大风险进行评估。
第三步	采取适当和有针对性的风险管理措施解决上述风险。

高风险 AI 系统具备风险管理体系的能力是欧盟《AI 法案》采取以风险为进路的治理思路的题中应有之义。一方面，AI 系统和传统产品不同，基于自我学习、算法更新等原因，其始终处于动态发展、变化的过程。另一方面，在技术上可能无法完全消除基于研发、部署、使用 AI 系统引发的算法歧视、偏见等问题，需要通过管理体系来进一步降低风险。此外，“算法黑箱”也在一定程度上使得 AI 系统对人类的影响存在相当的不确定性和不稳定性。基于风险平衡的视角，并非全部的风险提供者均需采取措施予以解决，其仅需对存在的重大风险进行评估并处理，最终将重大风险控制在可接受的范围内。

严格意义上，风险管理体系是贯穿高风险 AI 系统全生命周期的技术要求，高风险 AI 系统提供者不仅需在设计、研发阶段即采取措施确保系统可识别未来的风险，还需在 AI 系统上市后持续进行风险识别与判断，确保已投入市场的高风险 AI 系统对个人的基本权利所可能引发的风险始终处于可控制的范围之内。但是，依据 2023 年《AI 法案》折衷草案第 16 条第（a）款，提供者应在高风险 AI 系统投放市场前，即已确保 AI 系统具有《AI 法案》第二章为 AI 系统设置的要求。

目前，我国算法与 AI 领域的治理规范已有类似于风险管理体系的要求。具体而言，《互联网信息服务算法推荐管理规定》（以下简称“《算法推荐管理规定》）第八条规定，算法推荐服务提供者应当定期审核、评估、验证算法机制机理、模型、数据和应用结果

等。但是，《互联网信息服务深度合成管理规定》和《生成式人工智能服务管理暂行办法》尚未规定前述的定期审核、评估义务。我们理解，定期审核等系列义务虽然并不完全等同于欧盟《AI 法案》所述的风险管理体系，但这两者本质上均涉及 AI 系统提供者定期对 AI 系统可能引发的风险进行识别与评估，从而采取针对性措施予以解决。有鉴于此，在未来立法中，我国可进一步要求设置 AI 系统及算法的风险管理体系，确保动态发展的 AI 系统与算法对我国网络安全、数据安全以及个人信息保护的影响始终处于可被接受的范围内。

2）训练数据质量要求

训练数据的重要性不言而喻，诸如提高训练数据质量真实性、准确性、多样性的要求不仅在《AI 法案》的三个版本中均进行了规定，在我国已经发布的《生成式人工智能服务管理暂行办法》中也有体现。具体而言，AI 系统依托数据的喂养而训练和优化，训练数据的偏差将镜像地体现在 AI 系统的运作中。为了避免 AI 系统的算法歧视和偏差，提高训练数据的质量具有极其重要的意义。例如，预期在律师行业协助律师进行文件翻译的生成式 AI 服务，如在训练 AI 时仅是为其提供文学行业的语料，由于法律英语具备相当的专有词汇和用法，因此可能导致生成内容出现一定的偏差。

我国《生成式人工智能服务管理暂行办法》第七条以行为为导向，要求生成式 AI 服务提供者“提高训练数据治理，增强训练数据的真实性、准确性、客观性、多样性”。《AI 法案》对训练数据的质量要求则存在不同的维度，除了应采取适当措施，发现、防止可能存在的偏见问题，2023 年《AI 法案》折衷草案还进一步要求数据具有相关性、充分的代表性，在考虑到预期目的的情况下尽可能地完整，并在高风险 AI 系统的预期目的或可预见的使用范围内考虑到其特定的环境或场景因素等。

和 2021 年《AI 法案》提案相比，2023 年《AI 法案》折衷草案的数据治理要求的亮点如下：

- **新增技术可行性**

  2021 年《AI 法案》提案一经发布即引起社会公众热议，对于研发 AI 系统相关的企业而言，数据治理要求将使得企业承担极高的负担。最终，在 2023 年《AI 法案》折衷草案中，第十条新增了技术可行性的说明，即利用数据训练高风险 AI 系统时，只要依据相应的市场或应用范围在技术上可行，则应采取相应的质量要求。

- **新增偏见考虑**

  《AI 法案》对 AI 系统进行风险分类的核心目的在于，维护个人的健康、安全和基本权利等。基于此，2023 年《AI 法案》折衷草案第 10 条第 2 款第（f）项，新增偏见考量要求，即研发者在选择训练数据时，应当考虑到对个人健康、安全和基本权利产生的负面影响或导致欧盟法律所禁止的歧视或偏见。《生成式人工智能服务管理暂行办法》虽然并未在数据质量要求一条中规定类似的偏见考量义务，但其在第四条明确规定，“在算法设计、训练数据选择、模型生成和优化、提供服务等过程中，采取有效措施防止民族、信仰、国别、地域、性别、年龄、职业、健康等歧视”。

  通过中欧相关条款对比，我们理解，虽然我国和欧盟均意识到训练数据的偏见将对 AI 系统产生实质影响，但基于歧视偏见的范围等，我们不排除可能存在具体的差异。此种差异本质上是基于各国对 AI 系统监管的合法利益考量差异而形成，并反映在具体的 AI 系统监管措施之中。

- **明确充分性要求**

  2021 年《AI 法案》提案中，其第十条仅规定训练数据需具有相关性、代表性、没有错误以及完整的。虽然这一条款旨在提高训练数据质量从而提高高风险 AI 系统的服务质量，但是实践企业和欧盟成员国均有代表反应，这一要求可能过于严苛，且缺乏一定的实践意义。例如，丹麦等国代表均认为，没有错误的完美数据是不存在的。

  有鉴于此，为提高数据质量条款的适用性，2023 年《AI 法案》折衷草案进一步调整表述，补充了“充分的代表性”，“适当的错误审查”以及“在考虑到预期目的的情况下尽可能地完整”，在一定程度上解决了企业在实践中适用数据质量保障条款的现实障碍。尤其是针对数据的完整性要求，2023 年《AI 法案》折衷草案明确规定完整性需考虑训练数据的预期目的也即训练高风险 AI 系统的预期目的，一方面有助于提高 AI 系统在预期使用领域的效能，另一方面可以减少提供者为了实现训练数据的完整性，从而过多地使用不必要的数据。

  相较而言，我国《生成式人工智能服务管理暂行办法》虽然相对其征求意见稿而言，已从以“结果为导向”调整至“以行为为导向”[3]，但是，我们理解，网信等监管部门仍可基于实践情况，进一步明确训练数据“真实性、准确性、客观性、多样性”的具体含义，避免歧义，并指导、帮助相关企业开展训练数据

[3] 参见本书中《卧看星河尽意明——全球首部生成式人工智能法规解读》一文。

处理活动。我们在一定程度上同意数据不可能在各个角度上都是完美的，但考虑到训练数据可能会影响人工智能系统的判断从而放大数据的不完美性，企业不仅应当在选取预训练数据集时考虑更多的因素，而且还应当在后续训练中持续不断的对训练数据的真实性、准确性等进行核查，尽可能的降低其对人工智能系统判断的影响。

- **强化个人信息保护要求**

  2023 年《AI 法案》折衷草案第 10 条第 5 款在 2021 年《AI 法案》提案的基础上，进一步强化了个人信息（欧盟《通用数据保护条例》称之为“个人数据”）的保护要求，并提出采取适当的技术和组织措施从而保障训练数据涉及的个人信息的安全性等要求。

  无论是《AI 法案》还是《生成式人工智能服务管理暂行办法》，其均规定个人信息的处理需遵循相应的法律要求，具体如《通用数据保护条例》和《个人信息保护法》。但依据实践经验，企业在利用涉及个人信息的训练数据时，可能存在较多障碍，具体而言，相关主体主张无需取得个人信息主体同意而处理个人信息的“例外”合法性基础的空间有限，但获取海量个人信息主体的同意似乎又缺乏可行路径。如企业认为确有利用个人信息进行 AI 训练、优化的必要，可考虑选择经匿名化处理的个人信息作为替代方案。除匿名化及知情同意以外，各国监管部门应当就人工智能训练是否可以合理使用已公开个人信息给予更明确的指导。我们理解，对于已公开的个人信息无限制地用于人工智能训练可能会引发比如个人信息主体权利侵害，公开互联网平台不正当竞争等社会问题，但如果对训练数据的使用从技术和自我管理约束中，能够做到不标识个人信息主体或者对公开数据源予以成本补偿等，可能也是促成人工智能训练数据发展与安全平衡的方式之一。

3）技术文档与运行日志要求

2023 年《AI 法案》折衷草案第 11 条和第 12 条规定，高风险 AI 系统在投入市场前即应制定技术文档，并设计具有自动记录运行日志的能力。一方面，制定技术文档并保持更新的目的在于证明高风险 AI 系统符合《AI 法案》第二章的要求；另一方面，日志记录能力能确保高风险 AI 系统在生命周期具有一定程度的额可追溯性，并能识别特定的活动。

技术文档要求（2023 年《AI 法案》折衷草案附件 4）	
对 AI 系统的一般描述	• 高风险 AI 系统的预期目的、提供者名称和系统版本，反应现版本与旧版本的关系等； • 可能涉及的数据类型，尤其是可能涉及的个人信息主体类型； • 对系统的主要优化目标的描述； ……
对 AI 系统的要素其开发过程的详细描述	• 为开发 AI 系统所采取的方法和步骤； • 描述训练方法和技术以及所使用的训练数据集，包含训练数据集的来源、范围和主要特征等； • 所使用的验证和测试程序等； ……
关于 AI 系统监测、运作和控制的详细材料	• AI 系统性能的能力和局限性； • 有关 AI 系统在开发阶段的能源消耗以及在使用过程中的预期能源消耗等； ……
其他	• 对风险管理体系的详细描述； • 说明提供者在该系统生命周期内所做出的任何改变； • 欧盟合规声明复印件； ……

为每个高风险 AI 系统配置技术文档与日志自动记录能力，无论是对 AI 系统提供者，还是高风险 AI 系统的主管监管部门而言，均具有积极意义。一方面，系统提供者可通过技术文档和自动记录的运行日志，自证其在高风险 AI 系统投入市场前，已履行完毕相关合规要求例如构建风险管理系统的要求。另一方面，技术文档也可为监管部门提供所有必要的信息，从而协助其判断特定的高风险 AI 系统是否符合评估要求。

我们理解，制定技术文档有助于 AI 相关企业构建完善的内部合规体系，明确识别已采取的合规措施与《AI 法案》规定之间存在的差异。我国《生成式人工智能服务管理暂行办法》第十九条虽然规定提供者应配合主管部门的监督检查，并“按要求对训练数据来源、规模、类型、标注规则、算法机制机理等予以说明”，和技术文档涉及的为监管部门提供必要信息的功能有相似之处。但是，其本质上是对监管部门偶发的监管活动进行响应的要求。如 AI 企业期待寻求长期稳定、可持续的发展，在设计、完善内部合规制度体系时，可考虑将制定描述 AI 系统相关情况（包含 AI 系统的整体情况，研发、开发环节具体情况，

以及投入市场后的更新情况等内容）的文件纳入企业内部合规制度体系当中，不仅有助于企业快速、高效应对监管部门的检查，还有利于企业自证合规。

（2）构建质量管理体系的义务

根据 2023 年《AI 法案》折衷草案，高风险 AI 系统提供者应建立、实施质量管理体系。具体而言，如下表所示，该版法案第 17 条第（1）款规定提供者应通过书面文件的形式（如政策、程序或指令等）系统有序地记录其检测和验证程序、适用的技术规范、数据管理体系、风险管理体系、系统投放市场后监测机制、事故和故障报告、监管配合、问责框架等。

构建质量管理体系（2023 年《AI 法案》折衷草案）	
1	用于高风险 AI 系统的**设计、设计控制和设计验证**的技术、程序和系统操作
2	用于高风险 AI 系统的**开发、质量控制和质量保证**的技术、程序和系统操作
3	在开发高风险 AI 系统**之前、期间和之后**要进行的检查、测试和验证程序及频率
4	拟采用的**技术规格**，包括标准，以及在未完全采用相关的协调标准或未涵盖所有相关要求的情况下，为确保高风险 AI 系统符合第三编第二章对高风险 AI 系统规定的要求而采取的措施
5	**数据管理的系统和程序**，包括数据获取、数据收集、数据分析、数据标记、数据存储、数据过滤、数据挖掘、数据汇总、数据保留以及在高风险 AI 系统投放市场或投入使用之前和为之进行的有关数据的任何其他操作
6	**风险管理制度**（根据第 9 条规定）
7	建立、实施和维持**系统投放市场后的监测系统**（根据第 61 条规定）
8	与**报告严重事故和故障**有关的程序（根据第 62 条规定）
9	处理**与相关主管部门，包括行业主管部门的沟通**
10	**记录所有相关文件和信息**的系统和程序
11	**资源管理**，包括与供应安全有关的措施
12	**问责框架**，规定管理层和其他工作人员在本段所列各方面的责任

整体上，在三版《AI 法案》中，就质量管理体系承担的角色而言，其监管维度与风险管理体系同理，几乎覆盖了高风险 AI 系统的整个生命周期。从系统设计、开发阶段的质量控制、标准控制，到日常的数据管理和风险管理，再到系统投放市场后的监测制度、事故报告制度、沟通制度等等，质量管理体系将《AI 法案》对高风险 AI 系统在各阶段的各类监管要求进行了有机的嵌套和整合，为高风险 AI 系统提供者设置了较为清晰、全面的、与预期组织规模相称的较高注意义务。此种全生命周期监管的方式要求提供者在任何情况下均保持合规的严格程度和保护水平，将有助于提供者及时发现系统存在的问题和风险，并采取合理方式避免风险的发生或损害的扩大，以始终将高风险 AI 系统可能造成的重大风险控制在可接受范围内。

就质量管理体系的要求内容而言，鉴于质量管理体系是企业在实践中广泛采用的标准化实践[4]，三版《AI 法案》规定的质量管理体系与立法部门的现有质量管理体系（例如ISO 9000系列标准[5]或特定行业质量管理体系等）存在交叉准用以及动态交互。一方面，对于上述 2023 年《AI 法案》折衷草案第 17 条第（1）款的内容，相关部门可将其纳入部门立法规定的现有质量管理体系中。另一方面，对于特殊行业既有的质量管理体系构建义务，《AI 法案》也提供了部分准用性条款，提供者可直接适用相关特殊行业立法，例如 2023 年《AI 法案》折衷草案第 17 条第（3）款规定，对于受第 2013/36/EU 号指令监管的信贷机构，提供者若根据该指令第 74 条规定落实了内部治理安排、流程和机制的规则，则被视为已经履行了《AI 法案》项下建立质量管理体系的义务。

从监管类型来看，尽管我国非常重视标准体系建设，例如《国家新一代人工智能标准体系建设指南》从顶层设计的角度指出加强人工智能领域标准化发展，并出台一系列人工智能伦理、训练、人工标注等方面的安全规范，但我国尚无针对 AI 系统整体生命周期的质量体系层面的系统监管体系。如果将质量风险比作自始至终悬在 AI 系统合规管理之上的一把“达摩克利斯之剑”，那么质量管理体系就是紧紧绑住这把剑的绳索。质量管理体系越完备、系统、综合，高风险 AI 系统提供者对质量风险的把控就越符合预期、合乎规范。因此，基于对 AI 系统质量规范问题的重点关注，监管部门可通过将普遍实践的质量管理规则纳入现阶段对 AI 系统的规制范围内，并结合 AI 系统本身的特殊性予以补充，形成较为体系性的、可供提供者直接参考的质量标准。

---

[4] See Michael Veale & Frederik Zuiderveen Borgesius: Demystifying the Draft EU Artificial Intelligence Act - Analysing the good, the bad, and the unclear elements of the proposed approach.

[5] https://www.iso.org/the-iso-survey.html，最后访问时间：2023 年 7 月 17 日。

2. 投放市场时

（1）履行符合性评估义务

从监管方式的角度看，《AI 法案》对高风险系统采取了事前评估的监管方式，这也是目前国际社会实践中各国对算法进行评估的探索方向之一。《AI 法案》规定，在将高风险 AI 系统投放市场或投入使用前，提供者或第三方评估机构应当进行针对高风险 AI 系统的符合性评估（Conformity Assessment）。

具体而言，2023 年《AI 法案》折衷草案第 43 条详细规定了主体开展符合性评估的规则，其本质是为验证该版法案第三编第二章及相关条款、附录中对高风险 AI 系统的有关要求是否得到满足的过程。如下表所示，2023 年《AI 法案》折衷草案中，根据 AI 系统类型的不同，相应开展评估的主体和具体程序也有所不同。值得注意的是，系统发生实质性修改后，《AI 法案》要求相关主体重新进行符合性评估。

高风险 AI 系统的类型	评估条件		评估方式
用于特定领域目的、符合《AI 法案》标准的高风险 AI 系统（详见 2023 年《AI 法案》折衷草案附录 3）	针对附录 3 第 1 条中的“生物识别和和基于生物识别的系统”	如提供者已应用欧盟协调标准 6 或通用规范 7	开展**自评估**（详见 2023 年《AI 法案》折衷草案附录 6）
		如（1）不存在欧盟协调标准或通用规范；（2）提供者未应用或未完全应用欧盟协调标准；（3）应用时收到相应限制等情形；**或**（4）提供者认为需要第三方核查	**接受第三方**进行符合性评估（可自主选择第三方，详见 2023 年《AI 法案》折衷草案附录 7）
	针对附录 3 中第 2 条至第 8 条的系统		开展**自评估**（详见 2023 年《AI 法案》折衷草案附录 6）
根据某些欧盟法律作为安全组件或产品使用的 AI 系统（详见 2023 年《AI 法案》折衷草案附录 2）	如果系统适用的特定行业立法规定了符合性评估		接受**第三方**符合性评估
	如果产品制造商适用特定行业立法，且符合欧盟协调标准或通用规范		可**退出**第三方符合性评估

符合性评估的监管模式主要呈现以下三方面的特点：

[6] 协调标准：是指欧盟第 1025/2012 号条例第 2(1)(c) 条中定义的欧洲标准，是由欧洲三大标准化组织，即欧洲标准化委员会（CEN）、欧洲电工标准化委员会（CENELEC）及欧洲电信标准委员会（ETSI）制定并经欧盟委员会官方公告批准实施的欧洲标准。

[7] 通用规范：是指欧盟委员会制定的除标准之外的文件，其中载有技术解决方案，提供了遵守《AI 法案》规定的某些要求和义务的手段。

1）因系统而异的分级分类制度

符合性评估作为对高风险 AI 系统的监管方式，是《AI 法案》整体采用风险进路作为监管思路的重要表现。当涉及到不同风险级别的 AI 系统时，《AI 法案》采取不同程度的分级规范模式。不同于针对有限风险 AI 系统监管措施，《AI 法案》对高风险 AI 系统采取的监管模式呈现出强监管、全生命周期监管和事前评估式监管的特点。

以对质量管理体系的符合性评估为例，在高风险的 AI 系统中，质量管理尤为重要。这些系统往往应用于决策关键领域，如医疗诊断、金融风险评估或自动驾驶等。对于这些系统，准确性和可靠性至关重要，一旦发生错误，将可能导致严重后果。由此，为了确保系统的输出符合高标准，对于高风险 AI 系统，严格的质量管理体系的符合性评估必不可少。

相比之下，在低风险的 AI 系统中，质量管理严格程度相对灵活。这些系统通常应用于日常生活中的辅助工具，如语音助手、社交媒体过滤器等。尽管这些系统的准确性和可靠性仍然很重要，但其可能出现的负面影响也相对较小。因此，在这种情况下，可以灵活调整质量管理措施，以平衡成本和效益。

另外，人工智能的应用场景非常广泛。随着技术的发展，越来越多的行业和领域开始采用人工智能技术，如医疗保健、农业、交通运输、零售等等。显然，如果对所有人工智能产品或服务都适用统一的监管规则，可能会对部分提供者施加过高的合规义务。这可能导致不公平市场竞争，并阻碍创新和发展。因此，有必要根据不同的应用领域和风险级别制定差异化的监管政策，以促进合理的市场竞争和技术创新。

我国早在《网络安全法》《数据安全法》《个人信息保护法》中就确立了分类分级保护制度，为不同风险层级的规制对象适配不同程度的保护或规范制度。而在人工智能领域，秉承《关于加强互联网信息服务算法综合治理的指导意见》的“分级分类安全管理”思路，《算法推荐管理规定》规定了对于算法推荐服务提供者的分级分类管理制度，《生成式人工智能服务管理暂行办法》也明确提出了整体性的分级分类监管理念。

值得参考的是，在制定规范模式和监管政策时，需要综合考虑 AI 系统的风险级别、应用领域、执行的功能、使用的具体目的和方式，以及涉及的利益相关方。这样可以更好地平衡技术创新与风险管理之间的关系，促进人工智能的可持续发展并最大程度地造福社会。

2）自评估和第三方评估相结合

符合性评估包括提供者出于内部控制的目的自行进行的自评估，以及由第三方机构进

行的评估。这两种评估方式在适用范围和效果上有所区别。这种区别或与各类高风险 AI 系统的特点和潜在风险相关。

对于 2023 年《AI 法案》折衷草案附录 3 列出的涉及生物识别系统以外的高风险 AI 系统，自评估允许 AI 系统提供者主动承担责任，监控和管理其系统的合规性。通过自评估，提供者可以更好地了解和掌握系统的特点，根据自身业务需求进行调整和改进。对于一般的高风险 AI 系统，这种方法具有合理性，因为提供者本身在系统开发过程中具有充分的专业知识和资源，实施自评估符合实际，且为提供者提供了灵活和主动管理的机会。实施自评估也有利于培养提供者的 AI 素养（AI Literacy）。

对于涉及生物识别的系统以及作为安全组件或产品使用的 AI 系统，因其可能对基本权利和环境等带来更高的风险或其风险更具有隐蔽性，需要采用更为严格的评估方式。第三方评估提供了一个独立、客观的审查机制，可以对系统进行全面的检查和验证。通过第三方评估，可以减少利益相关方的干预，一定程度上敦促提供者接受社会方面的审查，相对增强公众对 AI 系统合规性和安全性的信任感。当然，基于已有的欧盟协调标准与通用规范已经覆盖了法案对高风险 AI 系统的部分要求，出于降低评估成本和提高评估效率的考量，2023 年《AI 法案》折衷草案也为提供者第三方评估设置了退出机制。

目前，我国对 AI 系统的监管采取了“自评估 + 公权力机关评估”的综合治理思路。然而，不论是自评估还是监管部门评估，都存在缺乏具体规则指引的问题。为此，监管部门可参考《AI 法案》符合性评估的考量维度，如 AI 系统评估框架、相关法律衔接问题、是否引入第三方评估问题、评估程序和标准问题、评估可操作性问题，为 AI 系统提供者明确可遵循的标准、提供自我评估指引，也为监管部门提供科学、规范和有效的依据，提高监管效能。

3）嵌入现有欧盟标准的准用机制

《AI 法案》为高风险 AI 系统列出了一系列需要接受评估的事项，但也同时允许某些特定高风险 AI 系统提供者在就所有相关事项符合欧盟的协调标准或通用规范的情况下，退出更严格的第三方评估。

这意味着，如果 AI 系统提供者能够满足适用的欧盟标准或规范，他们可以部分地将这些标准视为等同于符合性评估的要求。这并非豁免 AI 系统的标准化要求，更不是豁免了提供者的符合性评估义务，而是通过准用已有欧盟标准等规则，提供了一种替代的途径来证明 AI 系统的合规性。

准用机制的优势在于，其一，准用机制利用已经存在的监管框架和标准，通过结合已有规则，允许 AI 系统提供者利用现有的标准和规范来证明其系统的符合性，不仅为提供者提供了便利而安全的技术标准，也省去了重复的评估流程，进而避免提供者成本的增加和评估结果的延迟。

其二，通过准用机制，欧盟将现有的标准和规范与 AI 系统的标准化工作相衔接，可以确保 AI 系统在技术和操作层面上符合行业最佳实践和欧盟现有的监管要求，实现对新兴技术监管制度的彼此衔接，有利于主管机关的灵活监管。

其三，符合性评估的准用机制也见证了欧盟针对 AI 系统监管标准化工作的进一步发展。在逻辑上，《AI 法案》的准用机制是一种充分利用现有行业立法和协调标准、通用规范，并进一步扩展 AI 系统标准化工作的方式。法案中专门针对高风险 AI 系统的要求构成了对已有规则的延伸，或可进一步拉动现有欧盟各项标准规范在 AI 系统方面的进一步发展。

但是，值得警惕的是，在法律适用层面，当欧盟标准化组织制定的与 AI 系统有关的标准可以成为准用性规则而具有符合性推定的“特权”时，这类标准化组织在解释规则时显然具有一定的话语权，然而少数的标准化组织是否经过充分的审查进而具备代表广泛主体利益的意见？这一过程是否可能构成欧盟委员会等立法机构对标准化组织实质性的委托或授权？这些问题均值得思考与质疑。

对于我国而言，当前，在我国人工智能相关产品和服务不断丰富的同时，也出现了标准化程度不足的问题。人工智能涉及众多领域，虽然某些领域已具备一定的标准化基础，但是这些分散的标准化工作并不足以完全支撑各领域人工智能系统全生命周期的监管要求。因此，布局、梳理并加快形成人工智能领域的完善的标准体系有利于我国抢占标准创新的高点。但是，在加快以标准化的手段促进我国人工智能技术、产业发展的同时，明确标准制定主体的地位，界定立法与标准之间的关系，厘清标准之间的依存与制约关系，也是在建立标准体系时不可忽略的重点。

（2）高风险人工智能系统备案义务

《AI 法案》提案第 51 条规定，高风险 AI 系统投放市场或投入使用前，应在欧盟委员会主导建立的欧盟公共高风险 AI 系统数据库中备案。《AI 法案》折衷草案相较而言，除了新增了特殊部署者的备案义务，还进一步补充了在 AI 系统发生实质变更（Substantial Modification）时的重新备案义务。实质变更具体指在高风险 AI 系统投入市场时，发生了提供者在最初的风险评估中没有预见到的变化，且此种变化可能致使高风

险 AI 系统不符合《AI 法案》规定的要求，或者这种变化将导致高风险 AI 系统的预期目的被改变。

我国自《算法推荐管理规定》以来，截至《生成式人工智能服务管理暂行办法》，始终坚持要求具有舆论属性或者社会动员能力的相关主体应履行算法备案义务。但是，我国已初具体系的算法备案和《AI 法案》的 AI 系统备案仍存在一定区别，即透明度义务程度可能存在区别。

具体而言，我国既有的算法备案是面向监管部门的信息公开，即使网信部门在实践中将定期公布算法备案的履行情况，但公众也仅能了解服务提供者名称、备案的算法服务类型以及备案编号等基础内容。相较而言，《AI 法案》的高风险 AI 系统备案义务除了面向监管部门的信息公开，其还暗含了面向社会公众的信息公开。具体而言，《AI 法案》不仅规定，高风险 AI 系统数据库应面向公众公开，还应有助于社会公众阅读、浏览。

高风险 AI 系统备案时应提交的部分信息（《AI 法案》附件 8）	
1	提供者的名称、地址和联系方式
2	AI 系统的商品名称和其他任何明确的参考资料，从而对 AI 系统进行识别和追踪
3	对 AI 系统简单易懂的描述，包括 AI 系统的预期目的，对 AI 系统逻辑的基本理解
4	可能或预期可能由 AI 系统处理的数据类型及性质
……	

向社会公开与 AI 系统相关的信息有助于社会公众理解 AI 系统的运行逻辑和预期用途，便于公众的实际使用。相较而言，虽然我国《生成式人工智能服务管理暂行办法》虽然在其第八条规定“提供者应当与注册其服务的生成式人工智能服务使用者签订服务协议，明确双方权利义务”，服务协议事实上也有助于用户了解生成式人工智能服务的运行逻辑、潜在的风险等，但是，这一规定并未具体规定服务提供者在服务协议里应向用户告知的内容颗粒度。有鉴于此，监管部门或可在实践中进一步制定规则或指引，引导、协助相关企业充分履行向用户的信息公开义务。

（3）CE 标识标志义务

针对前述符合性评估的认证，《AI 法案》自 2021 年提案起，即引进了欧盟合格认证

（Conformité Européenne，“CE”）机制。对存在高风险的 AI 产品和服务，经评估程序被认定合格后，可贴上 CE 标志投入使用。CE 标识的规则由来已久，其是欧盟对产品安全性的认证，而非质量的认证，由此可知，欧盟将 AI 系统的安全性置于极高的战略地位，这也符合欧盟计划建立“可信”AI 系统环境的美好愿景。

如前所述，我国算法治理领域采取了算法备案的监督方式，但其在行政法上仅属于行政事实行为，监管机关并非对备案算法的真实性、合法性、安全性等作出行政确认。有鉴于此，在未来，针对较高风险的 AI 系统，监管部门可考虑在备案的基础上增设类似的认证机制；这类机制将有助于保障公共利益，促进企业开发利用可信的 AI 及 AI 产业健康发展。

3. 投放市场后

（1）部署风险监测系统的义务

2023 年《AI 法案》折衷草案第 61 条规定，高风险 AI 系统提供者应采取与 AI 技术和高风险 AI 系统相称的方式，建立并执行上市后的监测系统。监测系统应当具有收集、记录和分析由部署者或者其他来源提供的有关高风险 AI 系统在整个生命周期内的性能的相关数据。提供者可根据上述数据和分析情况，评估高风险 AI 系统是否在投入市场后仍继续符合评估要求。

除此之外，AI 系统始终是动态活跃的系统，而非静止的物品，因此，监测其在投放市场后的性能、状态等各方面情况有助于提供者随时掌握系统的最新动态，在风险到来前及时响应，避免风险扩大，造成更多损失。

《生成式人工智能服务管理暂行办法》第十四条规定，“提供者发现违法内容的，应当及时采取停止生成、停止传输、消除等处置措施，采取模型优化训练等措施进行整改，并向有关主管部门报告”。这一规定明确了违法内容生成式提供者应尽的义务。但是，在此基础上，监管部门还可进一步明确服务提供者发现、识别违法内容的流程，并依此作为提供者是否履行相应义务的判断依据。延伸至生成式人工智能领域外的 AI 治理领域，和《AI 法案》相似，监测系统始终具有其积极意义。

（2）采取纠正措施等义务

根据 2023 年《AI 法案》折衷草案第 21 条和 22 条，高风险 AI 系统的提供者如果认为或有理由认为其投放市场或投入使用的高风险 AI 系统不符合《AI 法案》的要求，应立即采取必要的纠正措施（Corrective Actions），并尽到相应的通知义务及合作调查义务。

具体如下表所示：

场景	义务类型	高风险 AI 系统提供者应承担的义务内容	
如果**高风险 AI 系统**提供者**认为或有理由认为**其投放市场或投入使用的高风险 AI 系统不符合《AI 法案》的要求（具体可参见符合性评估要求）	**纠正措施**	**立即**采取必要的纠正措施，**使该系统恢复正常运行**，或**视情况撤回、禁用、召回该系统**	
	通知义务	通知对象：（1）**责任链上的相关主体**，分销商、进口商（如适用）、部署者（如适用）；及（2）提供 AI 系统或投入使用的成员国的**国家主管部门**	
		如根据第 25 条指定了**授权代表**	通知对象：**授权代表**
		如系统必须根据第 43 条进行**第三方符合性评估**	通知对象：**公告机构**
	合作调查	应与**部署者**合作**调查原因**（如适用）	

2023 年 6 月 12 日，欧洲议会通过的《通用产品安全法规》（General Product Safety Regulation，“GPSR”）正式生效，并将于 2024 年 12 月 13 日实施[8]。这一法规是对《通用产品安全指令》（General Product Safety Directive，“GPSD”）的修改，引入了处理诸如网络安全风险、新型科技产品相关风险的产品安全规定。与 GPSR 中的负责人制度、产品召回制度、事故报告制度相呼应，《AI 法案》中对高风险 AI 系统提供者也设置了记录技术文档、建立质量管理体系义务与上表中纠正措施等义务，契合了 GPSR 应对 AI 系统引起的风险、保护消费者等核心目标。

其中，产品安全信息强制报告制度值得重点关注。报告义务是整个产品安全监管中的关键环节之一，在推动相关主体落实产品安全主体责任方面起着关键性作用[9]。目前，实施产品安全信息强制报告制度是国际通行惯例，且各国报告制度呈现日趋精细化的趋势。

对我国而言，我国《生成式人工智能服务管理暂行办法》在第十四条为生成式人工智能服务提供者也设置了发现违法内容时的纠正、整改措施义务，且设置了相应的产品安全信息的强制报告义务，如“应当及时采取停止生成、停止传输、消除等处置措施，采取模型优化训练等措施进行整改，并向有关主管部门报告”，此外，该条还为提供者附加了对使用者相关违法行为的监督、处理、报告责任。此外，《消费者权益保护法》第十九条和《消费品召回管理暂行规定》也对产品召回及报告制度进行了规定。但应对目前人工智能

[8] https://commission.europa.eu/business-economy-euro/product-safety-and-requirements/product-safety/general-product-safety-regulation_en，最后访问时间：2023 年 7 月 18 日。

[9] 政府管理创新标准化研究所：《国外产品质量安全事故强制报告制度对我国的借鉴意义》，2021 年 12 月 27 日，参见 https://www.cnis.ac.cn/bydt/kydt/202112/t20211227_52572.html，最后访问时间：2023 年 7 月 18 日。

技术的发展与普及，各类新业态的消费模式迅速发展，目前我国的产品安全信息强制报告制度仍未形成完整的体系，相关法律文件的级别仅为部门规章，法律约束力不足，且规定内容在表述上仍然较为模糊，在实践中容易产生寻租空间。因此，我国可在继续完善产品安全法律法规体系的基础上，进一步明确针对人工智能领域的强制报告的内容、程序、时限以及报告对象等，并在法律层级的立法中纳入对新兴技术风险应对的考量。

（3）配合监管部门开展工作的义务

与欧盟《通用数据保护条例》类似，《AI 法案》作为一项欧盟范围内的立法，将对每个欧盟成员国产生法律约束力。在各阶段的草案中，《AI 法案》均提出各成员国须建立相应的国家监管机构，以监督《AI 法案》的规则在国内的应用和实施。相应的，《AI 法案》也为提供者设置了需要配合上述监管机构开展工作的义务。

综上所述，《AI 法案》对高风险 AI 系统提供者设置了种类繁多的监管要求。尽管目前对我国 AI 系统监管思路来说，欧盟拟议的《AI 法案》并不是我国搭建治理结构的唯一框架或参考，且自 2021 年《AI 法案》提案发布之日起，部分监管要求在实践中存在较多争议，如数据质量要求、准用欧盟协调标准等等，但是，《AI 法案》的提出和讨论，确实从整体上为高风险 AI 系统的治理提供了体系化的、基于风险的治理模式和精细化的监管要求，也为这类主体提供了在开发 AI 系统时可以实施的核心原则的示范，尤其是可以为人工智能监管的发展赋能的符合性评估等制度，对我国进一步细化对人工智能系统监管措施而言具有较高的参考价值。

（二）有限风险人工智能系统的主要监管制度

针对有限风险的 AI 系统（Certain AI systems）的监管，《AI 法案》对其提出了透明度义务要求，具体表现在以下几个方面：

2021 年《AI 法案》提案 52 条规定有限风险 AI 系统的特殊 / 额外透明度义务，2022 年《AI 法案》妥协版本对此进一步完善，2023 年《AI 法案》折衷草案对此无更新。根据 2021 年《AI 法案》提案序言 5.2.4 的解释，透明度义务旨在应对有限风险 AI 系统的具体操纵风险（Specific Risk of Manipulation），保障人们能够作出知情选择或退出特定场景。下表将从义务主体、义务内容和例外情形等方面拆分有限风险 AI 系统的透明度义务。如本书《全球人工智能治理大变局之欧盟人工智能治理监管框架评述及启示》一文所述，有限风险的 AI 系统包含生物特征分类系统、情绪识别系统等。

不同类型的有限风险 AI 系统的透明度义务的内容可以从义务主体、义务内容和例外

情形等方面进行区分，具体如下表所示：

系统类型	与人类互动 AI 系统	情绪识别 AI 系统、生物特征分类 AI 系统	生成或操纵内容 AI 系统
义务主体	系统提供者	系统使用者	系统使用者
义务内容	告知：正在与 AI 系统互动。	告知：上述系统运行情况的存在。	告知：内容是由人为生成或操纵的。
例外情形	从一个具有合理的充分信息、观察力和谨慎的自然人的角度来看，AI 互相行为是显而易见的；法律授权的在适当保护第三方权利和自由的情况下用于侦查、预防、调查和起诉刑事犯罪的使用的情形。	法律授权的在适当保护第三方权利和自由的情况下用于侦查、预防和调查刑事犯罪的使用的情形。	法律授权的为了侦查、预防、调查和起诉刑事犯罪的情形或者内容是明显具有创造性、讽刺性、艺术性或虚构性的作品或节目的一部分，并受第三方权利和自由的适当保障的约束。
履行要求	最迟应在首次互动或接触时、以清晰可辨的方式提供。		
与高风险 AI 系统耦合的处理原则	“就高不就低”，当出现不同风险 AI 系统耦合时，以较高的透明度义务要求为准，即第 52 条的规定不影响承担高风险 AI 系统的义务和其他透明度义务。		

除第 52 条规定的透明度义务之外，2023 年《AI 法案》折衷草案还要求有限风险 AI 系统遵守第 4a 条 1d 的一般原则的透明度义务，即 AI 系统的开发和使用方式应允许适当的可追溯性和可解释性（Traceability and Explainability），同时使人类意识到他们与 AI 系统的交流或互动，以及适当告知用户该 AI 系统的能力和限制，并告知受影响的人他们的权利。

三、人工智能创新激励制度— 人工智能监管沙盒（Regulatory Sandbox）

（一）监管沙盒制度总览

监管沙盒制度为英国金融行为监管局（Financial Conduct Authority）于 2015 年为监管 FinTech（金融科技）及相关创新产品应运而生的一项监管措施，监管沙盒旨在“建立一个安全空间，企业可以在其中测试创新性的产品、服务、商业模式和提供机制，而不会因从事所述活动而立即招致通常的监管后果”[10]。虽然目前很多国家例如美国、澳大利亚、新加坡等已采取了类似的监管沙盒制度，但该制度初面市场时仍为一项较为大

[10] https://www.fca.org.uk/publications/documents/regulatory-sandbox, p3.

胆地、相对创新性地监管措施，原因来自于监管沙盒制度的逻辑基础：一为破坏性创新（disruptive innovation），二为适应性监管（adaptive regulation）[11]。

金融科技的技术多样化和其自身快速发展与变革的特征决定了金融科技初创企业（尤其是中小企业）在产品设计或业务运营之初往往无法承担繁重的监管合规义务，这将会严重阻碍其技术创新；但另一方面，自由放任的监管态度又无法阻止新兴的金融科技的肆意发展、对公共社会利益及国家安全等方面的潜在系统性风险。而通过监管部门设置统一的准入限定性门槛，使得沙盒参与者在限定的运行期间、以竞争但资源共享互通的模式，利用模拟真实市场环境开展业务测试，最终向市场推广产品，对监管者和金融科技监管沙盒的参与者均有益处。对监管者而言，沙盒有助于其更好地理解金融科技的创新逻辑与知识，从“治理主体”转变为“协商主体”有助于其尽早发现对新兴技术存在的监管盲区，推动监管创新。而金融科技监管沙盒的参与者被允许在受控环境下测试其产品或服务，可以更好地了解与把握监管框架和趋势，有助于减少技术开发阶段因监管不确定性带来的可能损失与合规成本。

然而，因在监管沙盒制度下传统意义上利益对立的两面主体需要共同合作、互利互惠，随之也产生了一系列问题，例如[12]：

- 沙盒项目的准入标准 / 要求界定过严可能会直接排除抱有优秀创新计划的初创企业，标准界定过松也会招致不严肃、无关紧要的初创公司，从而浪费沙盒项目的时间与资源，破坏沙盒的目的；
- 沙盒项目的适用范围可能过于有限或存在滞后性，仍会阻碍全面地测试新技术的潜能；
- 监管沙盒旨在为初创企业提供一个安全的空间，以测试其产品和服务，初创企业无需面对全部监管合规负担。然而，这可能导致监管的不确定性，初创企业不知道离开沙盒后将面临哪些具体的监管要求。这种不确定性可能阻碍需要为未来做规划的初创企业，使其不愿投资于可能无法满足监管要求的产品或服务；
- 监管沙盒可能创造一种环境，使金融科技初创企业能够利用监管漏洞获得相对于例如银行和金融机构的成熟市场主体的竞争优势，可能产生对消费者、公共利益、国家安全的系统性风险。

由此可见，监管沙盒制度本身固有的模糊性与流动性决定了该制度在实践中必然会要求监管部门作出进一步解读或指示，同时，不同国家因政策导向与对某一领域的监管或鼓励趋势与态度均存在较大差异，利用监管沙盒制度管制何种行业、何种技术为佳同样需要

[11] 廖凡：《金融科技背景下监管沙盒的理论与实践评析》，载《厦门大学学报（哲学社会科学版）》2019 年第 2 期，第 12 页。
[12] The Advantages of Regulatory Sandboxes in FinTech., https://youtapinsights.com/the-advantages-and-disadvantages-of-the-regulatory-sandbox-in-fintech/，最后访问时间：2023 年 7 月 28 日。

监管部门进一步举措明确。

（二）设立人工智能监管沙盒的初衷与立法思路的转变

2023 年《AI 法案》折衷草案序言第（72）条列明了 AI 监管沙盒的（“AI 监管沙盒”或“沙盒”）立法初衷：即 AI 监管沙盒的目标应是通过在 AI 系统开发和上市前阶段建立一个受控的实验和测试环境来促进 AI 创新，以确保创新的 AI 系统符合《AI 法案》和其他相关的欧盟与成员国的法律规定。对于监管部门而言，监管沙盒可以增进其对技术发展的理解，改进监管措施，并向 AI 系统开发人员提供指导，监管部门也可在此受控环境中针对创新型 AI 系统进行更多的监管学习与理解，为未来法律框架的修订提供思路。针对潜在的 AI 系统提供者而言，受控的安全试验环境可以使其快速推进 AI 系统的测试与开发，也可以为其提供更多的法律确定性。尤其针对中小企业、初创企业而言，AI 监管沙盒可以消除资源与技术障碍，使得其能更加积极地参与创新型 AI 系统的测试与开发，贡献其经验与专业知识。

2023 年《AI 法案》折衷草案规定的 AI 监管沙盒制度具有如下突出特点：

强制性监管沙盒制度	2023 年《AI 法案》折衷草案第 53 条规定，欧盟成员国需至少在国家层面设立一个监管沙盒，该规定系《AI 法案》折衷草案新修订内容，2022 年《AI 法案》妥协版本与 2021 年《AI 法案》提案并没有针对沙盒制度的设立作出强制性规定，而是选择用“国家主管机关可以设立”的表述，由此也可以看出欧盟立法机构对于确立监管沙盒的态度从建议性转变为强制性，使得企业在泛欧盟区域至少可以选择参加任一国家级别的沙盒进行 AI 技术的测试与试验，从而获得相对于地区性或本地性监管沙盒适配水平更丰富的资源力量与指导性建议，而国家级监管沙盒制度背后的监管部门通常在其本国内的监管级别也会趋高，这对于一国监管部门统筹学习新兴技术，深化与改进合规要求，并对未来 AI 立法监管框架修订也起到促进作用。
新增对中小企业技术上市前的特别监管指导	2023 年《AI 法案》折衷草案第 53a（3）条新增规定，针对拟参与沙盒的提供者，特别是中小企业和初创企业，应当协助它们获得（1）技术部署前的服务，例如关于《AI 法案》实施的相关指导；（2）帮助其标准化文件、获得认证与咨询等增值服务；（3）接触其他数据单一市场（Digital Single Market）倡议的机会，例如 Testing & Experimentation Facilities、Digital Hubs 以及 Centres of Excellence 并符合欧盟基准能力。
为公共利益，允许在 AI 监管沙盒中基于处理数据原始目的以外的目的开发特定 AI 系统	本制度在 2022 年《AI 法案》妥协版本已有体现，2023 年《AI 法案》折衷草案对该项规定进行了细化要求，即第 54 条所列条件全部满足的前提下，才可以出于处理数据原始目的以外的目的在监管沙盒中开发特定 AI 系统。同时，第 54 条新增允许处理数据目的的种类，包括为公共安全与公共健康开展的疾病监测与诊断之目的、保护生物多样性、改进环境污染及全球变暖缓释措施之目的、公共交通、关键基础设施与网络的安全与适应力之目的等。此外，第 54 条删除了需要在参与沙盒终止后一年内仍需保留处理个人信息的日志记录要求。

从上述制度不难看出，2023 年《AI 法案》折衷草案对于设立 AI 监管沙盒制度、在监管部门与市场主体之间建立有效合规沟通机制等方面持积极态度，欧盟委员会将基于《AI 法案》的相关规定，从设立、开发、实施和监督 AI 监管沙盒等方面出台更为细致地规定，以规范沙盒准入标准、申请程序、选择与退出沙盒以及参与沙盒各方的权利与义务等事宜，以解决 AI 监管沙盒如何在欧盟层面得到一体协调化适用问题。

然而，部分学者对 AI 监管沙盒是否能理想化落地提出了一些疑问 [13]，例如：

1. 人工智能监管沙盒仅适用于人工智能系统的开发与上市前的有限时间段

就 AI 技术的开发与应用，换言之，AI 技术的创新与监管两个方面，AI 监管沙盒均应当是一项以保障技术安全为优先的监管措施与前提。由于 AI 系统具有快速更新与无法提前预期后果的自我学习能力的特性，当 AI 系统上市之后即脱离了 AI 监管沙盒的管控，即便当某项 AI 系统退出监管沙盒时，相关监管部门会出具一份合规退出报告书（exit report），以表明 AI 系统，特别是高风险 AI 系统已满足了相关合规要求，其他市场监管部门在对 AI 系统进行符合性评估时也应参考退出报告书的内容，但这并未从根本上排除或有效控制 AI 系统本身固有的无法预期性和可能衍生的新型风险。虽然 AI 系统离开监管沙盒后并非处于完全不受任何规范限制的状态，AI 系统的提供者仍需履行《AI 法案》规定的相关合规义务，但监管沙盒的试验性质及其“有限空间、有限技术及有限测试时间”的特征可能并不能覆盖 AI 系统上市后面临的数据处理场景的复杂与多样性，AI 系统在沙盒测试环境下也可能无法有机会与其他市场上已经较为成熟的 AI 系统与互联网环境进行互动与耦合，一旦 AI 系统离开 AI 监管沙盒的“温床”，对于国家安全、社会公共利益以及消费者个人权益的影响亦是一项未知谜题。

2. 加入沙盒不代表责任的完全豁免，企业自证善意合规实则增加负担

欧盟虽然通过设立监管沙盒机制为促进创新提供了立法背书，但却对沙盒参与者可能面临的法律责任配以有限的保护机制。2023 年《AI 法案》折衷草案第 53 条（4）明确规定，拟参与 AI 监管沙盒的提供者仍然应就沙盒中的试验对第三方造成的任何损害承担相关适用的欧盟和成员国层面的法律法规项下的责任。可以看出，即便 AI 监管沙盒的主旨是促进与激励 AI 的创新，但沙盒内并非法外之地，参与沙盒的 AI 系统提供者仍需注意规范测试、训练 AI 系统时的各项操作，并及时就可能存在的“灰色地带”向监管部门寻求意见，保持流畅的沟通机制，及时采取替代性方案或缓释措施，以避免产生法律责任。

与此同时，2023 年《AI 法案》折衷草案就上述法律责任的承担提议了一项折中操作：

---

[13] https://www.cambridge.org/core/journals/european-journal-of-risk-regulation/article/sandbox-approach-to-regulating-highrisk-artificial-intelligence-applications/C350EADFB379465E7F4A95B973A4977D#fn18。

即若 AI 系统的提供者系完全按照其与监管部门就加入沙盒项目达成的计划条款（Agreed plan）行事，并且善意（Good faith）遵从监管部门的相关指导，此种情形下若对第三方造成任何损害，监管部门不会就 AI 系统提供者违反《AI 法案》的行为进行罚款。本条新增内容看似是一种妥协后的法律责任“豁免”，但在实践中可能会面临诸多问题：例如，企业需要自证 AI 系统的开发、测试、训练等在各个方面均符合事先约定的计划条款，虽然约定的计划通常会就监管部门与提供者双方的权利与义务进行约定，然而参与沙盒项目的 AI 系统可能未必来自单一提供者，在训练 AI 系统的时候可能存在需要与其他提供者的 AI 系统进行耦合、互交互联的情形，对于 AI 系统自身学习而产生的后果甚至是侵权行为通常无法事先预期、区分与规范，而企业仅能尽最大限度确保 AI 系统的开发与训练的各项技术安全措施等合法合规，对此某一提供者需要自证行为合规性可能会花费大量的时间与经济成本，相关 AI 系统的训练与开发也需同步暂停，最终效果可能会使得更多持有新型想法的初创企业畏惧沙盒制度，阻却创新试验的积极性。

（三）人工智能监管沙盒制度对中国监管实践的参考与启示

2017 年国务院印发的《新一代人工智能发展规划》[14] 已经提到要统筹布局 AI 创新平台，包括 AI 开源软硬件基础平台、群体智能服务平台、混合增强智能支撑平台、自主无人系统支撑平台、AI 基础数据与安全检测平台，同时促进各类通用软件和技术平台的开源开放。此后，《关于加快场景创新以人工智能高水平应用促进经济高质量发展的指导意见》《网络安全标准实践指南——人工智能伦理安全风险防范指引》《生成式人工智能服务管理暂行办法》等法律法规均对上述开源开放合作的精神提出了具体回应，例如将大力支持专精特新“小巨人”、独角兽、AI 初创企业等积极开展场景创新，鼓励算力平台、共性技术平台、行业训练数据集、仿真训练平台等 AI 基础设施资源开放共享等等。从促进与激励技术创新角度，上述纲领与初衷与设置 AI 监管沙盒的意图具有相似之处，即打造一项技术与资源共享互利认知，以加速积累的技术能力与海量的数据资源、巨大的应用需求、开放的市场环境有机结合，形成我国 AI 发展的独特优势 [15]。

但是将类似 AI 监管沙盒制度嵌入我国现行 AI 规范法律框架中，对于监管部门与企业均提出了挑战：

对于监管部门而言，需要就沙盒的准入资质、申请与挑选“合格”AI 技术的程序、AI 技术 / 项目退出沙盒、结束试验的合规性评估要点等方面确立统一的规范制度，对于在沙盒中可能存在的技术耦合进行风险评估（例如是否允许交互耦合、耦合的程度等），沙盒试验环境的信息网络安全措施由哪一方保证落实（监管部门或另行指定具有技术能力的

[14] https://www.gov.cn/xinwen/2017-07/20/content_5212064.htm，国务院印发《新一代人工智能发展规划》。
[15] 同上。

第三方）也需要进一步研判，同时需要就发生数据安全事件、信息泄露、损害国家安全、公共利益、个人主体权益时的缓释措施与救济方案进行提前预设预期等等。

对于企业而言，监管沙盒制度带来的共享资源与技术支持虽然会推动创新型企业、中小企业的 AI 技术研发，但在缺少明确的立法框架（包括参与沙盒的各方权利与义务）与激励机制的前提下，市场参与度可能并不会达到立法预期效果，相比促进企业自身的技术沿革与发展而言，企业可能更加担心潜在地合规成本甚至是违法风险。此外，将新型技术投入至共享平台中也可能带来商业秘密、知识产权、算法算力专业知识等方面的泄露与抄袭，企业亦会担心自身的技术产品因此失去市场竞争力而不愿参与沙盒项目。此外，基于沙盒内其他 AI 技术而进一步得到开发、训练、测试的技术，如在未来投入市场后存在侵权行为，认定责任主体相对复杂困难也会是企业在加入沙盒之前考虑的问题之一。

此外，同上文所述，AI 监管沙盒制度如果仅适用于 AI 系统在上市前的开发与训练，基于 AI 系统自身学习与训练效果的不可预期性，事前监管仍存在滞后性，当 AI 系统上市之后即会面临更加复杂地真实世界场景与问题，不可预知地风险也会随着 AI 系统自身调节、训练与升级不断产生。对此监管部门可以考虑将 AI 监管沙盒的监管、指导期间延伸至相关 AI 系统上市后一定时间，这对于监管部门与企业双方而言均是一种交流与互惠互利的有效渠道，监管部门可以了解到 AI 系统上市后可能产生的问题与风险，从而加快修订有关规范制度，出台新行业标准以更加全面、迅速地对新风险做出立法应对，企业也可以就 AI 系统在真实世界应用过程中产生的合规问题与监管部门及时沟通，从源头做好合规整改；双方将就保护国家安全、造诣与提升公共利益方面共同挥发自身的角色与作用。

四、人工智能监管的未来方向

（一）设置安全标准或法律标准

我国目前阶段对于 AI 的治理规范主要集中于确保 AI 的安全性、使用的透明性、算法的可解释性以及符合伦理性等方面，已初步形成了包含法律、部门规章、地方性法规、国标、行业自律标准的多层次治理规范结构，在标准体系建设方面，例如国家标准化管理委员会等五部门印发《国家新一代人工智能标准体系建设指南》《人工智能伦理安全风险防范指引》《生成式人工智能预训练和优化训练数据安全规范》《信息安全技术生成式人工智能人工标注安全规范》等标准已经为 AI 的开发、测试、运行提供了指导性思路。而拓展中国 AI 发展的潜力与市场的前提是市场主体可以依赖相对细致与健全地技术标准与安全标准，按照既定的规则履行合规义务，尤其是若参考欧盟立法与监管的评估审查思路与逻辑，如何为一项发展迅速又具有固有不可预测性的技术从事前节点制定技术标准与安全标准将系监

管部门首要考虑的问题之一。其次，正因为 AI 技术涉及的数据种类与行业多样，各行业主管部门也应对本行业内特有数据处理场景及相关 AI 技术 / 系统出台标准文件为企业提供参考。第三，在选择制定强制性标准或推荐性标准文件方面，主管部门需要考虑企业在实践中落地标准要求的切实可能性、合理性与可行性，相对较严格的法律标准或强制性标准反而可能会阻却 AI 在我国开发与应用的进程。

（二）进一步落实动态平衡的要求

欧盟《AI 法案》旨在通过监管和促进发展的动态平衡，确保 AI 技术的合规性与可持续发展，在保障公共利益和个人基本权利的同时，推动 AI 技术的创新与运用。目前，我国也已经出台多部相关的法律和政策性文件，如《人工智能伦理安全风险防范指引》《新一代人工智能发展规划》《关于加快场景创新以人工智能高水平应用促进经济高质量发展的指导意见》等，在强化风险监管的同时，坚持推动 AI 技术的创新发展。虽然欧盟《AI 法案》通过制定具体的规定和机制用于满足这种平衡，但在实践中，动态平衡的实现仍具有一定的挑战性，具体而言，对于 AI 领域的监管，监管措施既不能过度严苛，也不能过于宽松，同时要具备一定的灵活性。过于严苛的监管措施会限制 AI 技术的创新，给相关义务主体造成较重的合规负担；过于宽松的监管措施会导致 AI 技术滥用，无法实现监管效果；固定且僵化的监管将无法满足 AI 技术的不断发展变化的现实要求。自《AI 法案》提案发布后，欧盟乃至世界范围对此具有义务过于严苛、不利于技术发展的负面评价。而欧盟在数次谈判和修订中也尽其努力，降低基于 AI 系统监管而对 AI 技术及产业发展带来的负面效应。

目前，我国仍处于对于 AI 领域的探索阶段， 许多新兴技术还无法做到准确的掌握和运用，未来国内 AI 的发展方向也无法准确预知，因此，对于 AI 领域的监管实际上无法且无需做到明确且清晰，需要保留一定的开放性。我们理解，我国最新出台的《生成式人工智能服务暂行管理办法》基于风险分级分类的思路和制度已经为此后基于生成式人工智能服务乃至 AI 技术的发展预留了动态平衡监管的空间。在风险分级分类的基础上，监管部门可依据实际情况适时调整风险分级分类清单，实现 AI 系统或服务在风险等级上的动态调整，在“保安全”的基本要求不变的情况下促使 AI 技术最大程度迸发其活力。

（三）监管措施或可更新：“以 AI 监管 AI”

无论是欧盟《AI 法案》，亦或是我国将于 2023 年 8 月施行的《生成式人工智能服务管理暂行办法》，在为服务或系统提供者施加合规义务时，仍是以传统的监管措施为主线。例如，安全评估、符合性评估均以评估为核心，同时具备国家评估和自评估两种要求；算

法备案实质上为一种行政事实行为，属于行政行为在算法治理领域的具体应用。但是，传统的监管手段往往以被监管者特定时间内较为固定的事实基础为依托，而基于 AI 技术和 AI 系统自我学习、不断变化的特性，在对其监管时若仍仅采取传统的监管手段，可能无法随时应对 AI 系统发生的新变化和新风险。同时，相对传统的活动而言，AI 系统的处理行为对人类而言更具不可知性。

为了解决 AI 系统在高速发展的过程中可能引发的等多样问题，尤其是大模型引发的 AI 伦理等问题，在未来，企业或监管部门或可采取以“AI 监管 AI”的方式，将人类的监管逻辑转换为机器可读的指令，构建一个“与人类水平相当的、负责模型对齐的‘AI 研究员’”。例如，在判断某 AI 系统是否存在显著风险时，监管部门可考虑设计特定的评估 AI 系统，用以判断被评估的 AI 系统是否具备相应的风险实时监测能力。

值得注意的是，OpenAI 于近日创建了一个名为 Superalignment 的研究团队，并计划打造一个如前所述的“AI 研究员”，并准备开展一系列诸如开发可扩展的训练方法等工作，从而打造一个解决超级 AI 对齐的“AI 研究员”。日后，OpenAI 对此的工作方法论和阶段成果均可为社会各界提供新鲜能量 [16]。

（四）人工智能伦理治理的发展

欧盟致力于制定人工智能治理规则的出发点之一，就是旨在通过制定明确的规则和标准来促进相关主体开发和使用负责任且合乎道德的人工智能。我们可以在欧盟《AI 法案》的各项规则中发现其实现伦理治理的思路，例如，《AI 法案》要求设计值得信赖的人工智能（Trustworthy AI），这体现了欧盟致力于为人工智能系统设立一个更为宏观的价值标准，希望通过将《欧盟基本权利宪章》（European Union Charter of Fundamental Rights）及以人为本的价值观、道德确立为更高级别的框架，以避免人工智能系统对公平、非歧视、保护隐私等社会权利等人类普世价值的褫夺；再如，《AI 法案》要求培养所有利益相关方的人工智能素养（AI Literacy），将其视为推动人工智能技术得以正确使用和发展的必然要求，也体现了欧盟对人工智能技术发展的价值引导思路。

简言之，《AI 法案》的伦理治理思路就是尽可能地实现人工智能与人类伦理在最大程度上的对齐（Alignment），即在人工智能系统的设计、开发和应用中，要确保人工智能系统与人类价值和期望一致，以促进人工智能技术的正确、负责和可持续发展，避免人类利用人工智能的初衷被与人类思维存在差距的算法规则吞噬。

---

[16] OpenAI 创新举措：用 AI 监督 AI，构建系统应对智能挑战，https://mp.weixin.qq.com/s/5EJq_cxJ2vjwbXfXPB6OVg，最后访问时间：2023 年 7 月 18 日。

对此，我国早在 2017 年就发布了《新一代人工智能发展规划》[17]，指明了人工智能可能带来的冲击法律与社会伦理等问题。目前，我国人工智能伦理治理呈现横向立法格局，《科技进步法》《个人信息保护法》《数据安全法》《生物安全法》等法律法规和《科学技术活动违规行为处理暂行规定》等相关制度均提出重视科技伦理治理的核心思路，科技伦理治理也随着国家科技伦理委员会的成立进入新的阶段。2021 年发布的《新一代人工智能伦理规范》体现了我国将伦理道德融入人工智能全生命周期，为从事人工智能相关活动的自然人、法人和其他相关机构等提供伦理指引的立法倾向。2022 年发布的《关于加强科技伦理治理的意见》我国首个国家层面的科技伦理治理指导性文件，为加强科技伦理治理划定了“红线”和“底线”[18]，将科技伦理审查和监管制度确立为贯穿人工智能生命周期治理的关键一环。2023 年我国还发布了《科技伦理审查办法（试行）（征求意见稿）》，对存在较高伦理风险的科技活动实施清单管理，并建立伦理审查结果专家复核机制。

目前，我国也在努力追求人工智能与人类伦理在最大程度上的对齐，“科技向善”的人工智能伦理治理思路日益清晰。从立法框架而言，我国目前人工智能伦理治理呈现横向发展格局，可考虑在立法过程中为行业立法留出足够空间，以构建更为纵深的各行业、各领域的伦理治理格局。从规范范围而言，我国可考虑进一步延伸对人工智能系统全生命周期的认定和规范，让所有相关方从编写第一行代码开始都有意识地设计负责、安全、值得信赖的人工智能系统和产品。从规范标准而言，我国在考量人类普适性价值同时，应立足我国科技发展的历史阶段和社会文化特点，限制不符合伦理的行为，促进人工智能系统高水平安全发展。从配套制度而言，我国应当加强对所有相关主体的人工智能伦理的教育和宣传，提高社会大众、从业人员和决策者的意识和理解，并加强对人工智能伦理原则和实践的培训，使相关人士具备认识和解决人工智能伦理问题的能力，推动人工智能技术的正确应用和发展。只有尽可能通过全方位发力，才可以建立起个体对人工智能的信心，并鼓励以安全且有益于社会的方式设计和使用人工智能技术。

感谢实习生刘阳、王艺捷对本文作出的贡献。

[17] 国务院关于印发《新一代人工智能发展规划》的通知，https://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm，最后访问时间：2023 年 7 月 18 日。

[18] 国家发布重磅科技伦理治理文件：基因编辑、人工智能等技术研发将得到规范，https://www.gov.cn/xinwen/2022-03/24/content_5681045.htm，最后访问时间：2023 年 7 月 18 日。

历时三年，欧盟《人工智能法案》通过欧洲议会表决

宁宣凤 吴涵 张浣然 刘畅 王储

引言

2024 年 3 月 13 日，欧洲议会以 523 票赞成、46 票反对和 49 票弃权的表决结果通过了《人工智能法案》（Artificial Intelligence Act）。

历时三年，经过成员国与欧盟理事会、欧盟议会等欧盟层面机构的多次磋商、讨论，面临欧盟、美国等国家 / 地区科技巨头企业的担忧与疑虑，欧盟人工智能监管立法终于迎来重大进展。

一、《人工智能法案》制定历程

欧盟始终期待打造可信任的人工智能生态环境，为实现这一愿景，欧盟陆续发布《欧盟机器人民事法律规则》《欧盟人工智能》《人工智能白皮书——追求卓越和信任的欧洲方案》等一系列人工智能政策，以期推动欧盟的人工智能产业发展，并形成健康、可持续的人工智能监管机制。

基于前述愿景，2021 年 4 月，欧盟首次发布《人工智能法案》提案，以“风险为进路”，基于人工智能系统的应用场景等因素，将人工智能系统划分为不可接受的风险、高风险、有限风险和极低风险四类，并对每类风险的人工智能系统提出差异化的监管要求。2022 年 12 月 6 日，欧盟理事会通过了关于《人工智能法案》的共同立场。2023 年 6 月 14 日，欧洲议会对《人工智能法案》草案进行表决，并以 499 票赞成、28 票反对和 93 票弃权的压倒性结果通过了其妥协立场。此后，欧盟成员国、欧盟理事会及议会就《人工智能法案》的具体条款组织“三方会谈”，进行多次谈判，并于 2023 年 12 月 9 日达成有关《人工智能法案》的临时协议。

达成临时协议并不代表欧盟本次人工智能立法程序的结束，而仍有待欧洲议会的最终表决。随着立法程序的进一步更新，2024 年 2 月 13 日，欧盟议会内部市场和消费者保护委员会与公民自由、司法和内政事务委员会以 71 票赞成、8 票反对和 7 票弃权的投票结果通过了与各成员国就《人工智能法案》达成的谈判草案。2024 年 3 月 13 日，欧洲议会正式批准了《人工智能法案》，扫清了欧盟人工智能监管立法的前置障碍。

未来，《人工智能法案》还将进行条款勘误，并预计在 2024 年 4 月交由欧盟理事会

批准以使得《人工智能法案》正式成为欧盟法律。

二、《人工智能法案》适用范围

《人工智能法案》的立法宗旨之一在于维护公平的人工智能竞争环境并有效保护欧盟个人的自由等基本权利。为了实现前述目标，《人工智能法案》确立了较为广泛的适用范围且具备域外效力：一方面，尽管某企业并非欧盟实体，只要其将人工智能系统在欧盟境内投入市场或投入使用，也需适用《人工智能法案》。另一方面，人工智能系统最终在市场发挥作用，除人工智能系统提供者外，离不开部署者、分销者等利益相关主体的参与。因此，人工智能系统价值责任链的相关主体均可能成为《人工智能法案》的规制对象。

图 1：《人工智能法案》适用范围

（一）适用于在欧盟境内将人工智能系统投入市场或投入使用的人工智能系统提供者

《人工智能法案》第 2 条规定，无论是欧盟境内还是欧盟境外的实体，只要其在欧盟境内将人工智能系统或通用人工智能模型投入市场或投入使用，均将受到《人工智能法案》的规制。场所位于欧盟境外的人工智能系统提供者，如其系统产生的产出将用于欧盟，相关主体也需适用《人工智能法案》。

此外，如提供者在欧盟境内尚未设立实体，根据《人工智能法案》的规定，其应指派一个位于或设立于欧盟的授权代表，代提供者履行相关合规义务。

（二） 适用于人工智能系统价值责任链上的多方主体

鉴于人工智能系统价值链的复杂性，人工智能系统的提供者、部署者、进口者、分销者或其他第三方均需履行其角色项下的合规义务。

具体而言，根据《人工智能法案》的相关规定，除了人工智能系统提供者，如下主要类型的主体也将受到欧盟人工智能立法的监管：

- **部署者**：部署者指经授权使用人工智能系统的公共机构、法人和自然人（自然人在非职业活动中使用人工智能系统除外），包含在欧盟内设立场所或位于欧盟的人工智能系统部署者，以及其人工智能系统产生的产出将用于欧盟的部署者；
- **进口者:** 将带有欧盟境外自然人或法人的名称、商标的人工智能系统投放欧盟市场、且位于或设立于欧盟的法人、自然人；
- **分销者:** 除人工智能系统提供者、进口者以外，在欧盟市场上提供人工智能系统的主体;
- **产品制造者**：以自己的名称或商标将人工智能系统与其产品一同投入欧盟市场或投入使用的产品制造者。

一方面，高风险人工智能系统的进口者、分销者等主体需履行《人工智能法案》第 23 条、第 24 条等相关规定对其施加的相关合规义务。另一方面，满足特定情形时，部署者、进口者等还将被视为高风险人工智能系统的提供者，履行提供者的相关义务。例如，相关主体对已投入市场或投入使用的高风险人工智能系统进行实质性调整，且调整后的人工智能系统也属于高风险人工智能系统，相关主体则需履行高风险人工智能系统提供者的义务。

三、《人工智能法案》主要内容

（一）以风险为进路，对不同风险等级的人工智能系统实施差异化监管

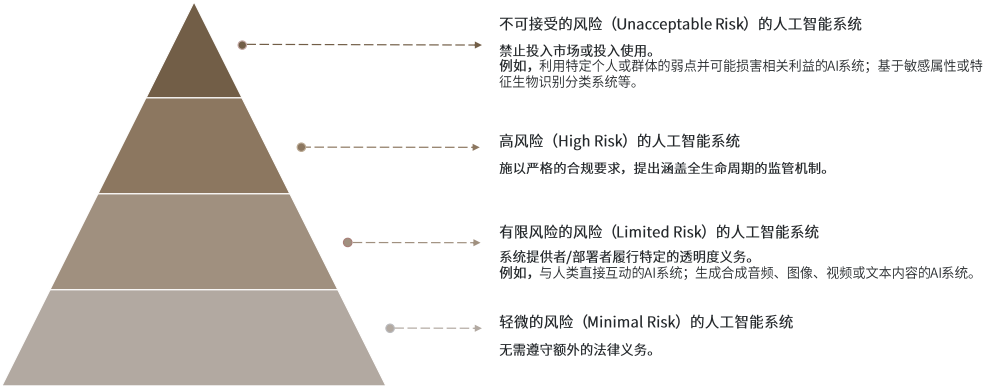


图 2：以风险为进路

最新通过的《人工智能法案》延续了自《人工智能法案》提案起即确立的人工智能系统四级监管机制，即将人工智能系统的风险等级划分为不可接受的风险、高风险、有

限的风险和轻微的风险四类，并对每类人工智能系统施以不同的监管要求。

具体而言，《人工智能法案》禁止具有不可接受风险的人工智能系统投入市场或投入使用；有限风险的人工智能系统的提供者、部署者应履行特定的透明度义务；轻微风险的人工智能系统的使用不受限制；高风险人工智能系统则是《人工智能法案》的重点关注对象。

就高风险人工智能系统而言，《人工智能法案》构建了涵盖事前、事中和事后的全生命周期监管体系：

高风险人工智能系统投放市场或投入使用前，其提供者应确保人工智能系统具有相应的风险管理体系、使用符合《人工智能法案》第 10 条规定的质量标准的训练、验证和测试数据集、附有证明该人工智能系统符合《人工智能法案》合规要求的技术文档、具有自动记录日志的能力、设置人为监督措施等。同时，高风险人工智能系统提供者还应建立质量管理体系，以确保其自身能够遵循《人工智能法案》对其施加的合规要求。

高风险人工智能系统投放市场或投入使用时，其提供者应当开展基本权利影响评估与符合性评估、制定欧盟符合性声明、加贴 CE 标识，并在欧盟委员会维护的欧盟数据库中对高风险人工智能系统进行备案。

高风险人工智能系统投放市场或投入使用后，其提供者应持续监测人工智能系统的风险活动，以便评估其系统在上市后是否持续符合《人工智能法案》的相关要求；采取纠正措施例如召回、停用，以应对人工智能系统已不符合监管要求而带来的风险；如发生严重事件，提供者还应向严重事件发生地的成员国市场监管管理机关进行报告。此外，成员国的国家市场监督管理机关、欧盟设立的人工智能办公室等机构有权依据《人工智能法案》履行相应的职责，因此，高风险人工智能系统提供者等市场主体还负有配合主管机关执法的义务。

（二）新增对通用人工智能模型及系统的监管要求

图 3：通用人工智能模型分类

随着 ChatGPT 等 AI 应用在世界范围内掀起热议，欧盟理事会于 2022 年 12 月 6 日就《人工智能法案》达成共同立场时即新增了有关通用型 AI 模型（General Purpose AI Models）的合规要求。最新通过的《人工智能法案》也顺应了对通用人工智能模型及其系统进行规制的趋势，并延续了“以风险为进路”的思路，将通用人工智能模型进一步区分为是否存在系统性风险（Systemic Risk），并对具有系统性风险的通用人工智能模型提供者提出更高水平的监管要求。

简而言之，通用人工智能模型提供者均应履行《人工智能法案》第 53 条规定的透明性义务，包括编制技术文档、提供相关信息至通用人工智能模型的下游人工智能系统提供者等。如果相关通用人工智能模型被认定为存在系统性风险，其提供者还应遵循《人工智能法案》第 55 条等相关要求，按照标准化协议和工具对模型进行评估从而识别和减轻存在的系统性风险、报告事件等。

（三）确立人工智能监管沙盒制度，激励人工智能创新，保护中小企业发展

一方面，人工智能的技术革新是数字社会发展的必然趋势，基于维护个人的安全、健康、民主等基本权利之目的而对人工智能进行监管不应过分扼制人工智能产业的发展。另一方面，新生的人工智能监管要求可能在短时间内拉高了企业的合规成本，尤其是对于小微企业（Small and Medium Enterprises, SMEs）而言，高昂的合规成本可能影响其参与人工智能市场的竞争。基于此，欧盟建立人工智能监管沙盒（AI Regulatory Sandboxes）制度，以推动人工智能生态系统的发展，促进和加快人工智能系统、特别是小微企业的人工智能系统进入市场。

《人工智能法案》第 57 条规定，欧盟各成员国应至少建立一个国家级别的人工智能监管沙盒，使得潜在的高风险人工智能系统提供者可以在可控的环境内进行研发、培训和测试，接受相关主管机关提供的指导与便利。企业可申请参与人工智能监管沙盒，且主管机关设置的申请、选拔流程应当便于小微企业参与；如企业是小微企业，还可免费进入监管沙盒。

此外，为了保护中小企业的利益，《人工智能法案》第 62 条还规定了针对中小企业的保护措施。例如，在合适的情形下，当地主管机关还应针对小微企业开展有关《人工智能法案》的培训，提高相关企业对人工智能监管要求的认识，并促进小微企业参与人工智能系统相关标准化规则的制定。

（对于《人工智能法案》更为详细的监管规则，请见本书中《路未央，花已遍芳——欧盟 < 人工智能法案 > 主要监管及激励措施评述》与《全球人工智能治理大变局之欧盟人

工智能治理监管框架评述及启示》等文章。）

四、《人工智能法案》时代的企业合规方向

可以预见，《人工智能法案》一经生效，将对适用企业，尤其是提供不可接受风险、高风险人工智能系统的企业产生重大的实质影响。尽管存在诸多争议且落地可操作性有待观察，但作为全球首部综合性人工智能立法，《人工智能法案》有潜力凭借其广泛适用性取得事实层面和法律层面的“布鲁塞尔效应”。

（一）尽快评估自身法律适用性并留意配合法案的分阶段实施时间框架

按计划，《人工智能法案》将于在官方公报上公布 20 天后生效，并在生效 24 个月后完全适用。但考虑到市场主体对于超前监管规则抑制产业发展等广泛担忧，《人工智能法案》的不同规则将分阶段实施，并预计于 2027 年全面实施。

具体而言，不可接受风险的人工智能系统应在《人工智能法案》生效 6 个月后完全禁止；通用人工智能模型的治理规则将在《人工智能法案》生效 12 个月后执行；高风险人工智能系统的合规要求则预留了 36 个月的合规整改期限。因此，位于欧盟境外的组织应审慎评估其人工智能业务是否可能受到《人工智能法案》的辐射，以确认是否需开展相关合规工作。

（二）面对高额罚则，审慎、加快开展落实合规方案

如此后相关企业需遵循《人工智能法案》开展业务，则应尽快开展合规工作，避免未来自身因违反法案规定而被处以最高达 3500 万欧元或全球年营业收入 7% 数额的罚款。

主要作者

（按拼音 / 字母排序）

戴梦皓

daimenghao@cn.kwm.com

方禹

fangyu3@cn.kwm.com

景云峰

jingyunfeng@cn.kwm.com

刘迎

liuying3@cn.kwm.com

楼仙英

cecilia.lou@cn.kwm.com

宁宣凤

susan.ning@cn.kwm.com

钱琪欣

qianqixin@cn.kwm.com

瞿淼

qumiao@cn.kwm.com

宋海燕

seagull.song@cn.kwm.com

孙及

sunji@cn.kwm.com

唐丽子

tanglizi@cn.kwm.com

王哲峰

wangzhefeng@cn.kwm.com

吴涵

wuhan@cn.kwm.com

张逸瑞

zhangyirui@cn.kwm.com

赵新华

atticus.zhao@cn.kwm.com

设计排版：张园园

编辑：王静静

本出版物中，凡提及“香港”“澳门”“台湾”，将分别被诠释为“中国香港特别行政区”“中国澳门特别行政区”“中国台湾地区”。

如您对本系列丛书有任何问题，请与编辑人员联系，

电邮地址：publication@cn.kwm.com

金杜是一家总部位于亚洲的国际化律师事务所。作为在中国内地、香港特别行政区、澳大利亚、美国、日本和新加坡等重要法域拥有执业能力的国际化律师事务所，金杜在全球最具活力的经济区域都拥有相当的规模和法律资源优势。我们面向全球，为客户锁定机遇，助力他们在亚洲和世界其他区域释放全部发展潜能。

我们始终以伙伴的合作模式为客户提供服务，不止步于满足客户所需，更关注实现客户目标的方式。我们不断突破已取得的成就，在重塑法律市场的同时，打造超越客户预期的律师事务所。

金杜公众号微信平台 KWM_China