⊙ 人工智能 ⊙

生成式人工智能大模型的新型风险与规制框架

刘金瑞

（中国法学会法治研究所，北京 100081 研究员）

摘　要：大模型技术的发展，掀起了生成式人工智能发展的新浪潮，但其数据训练和部署应用也带来了新的风险挑战，包括产生有害内容、泄露敏感数据、生成错误信息、滥用实施违法活动、可能危害环境和经济、向下游传导风险等。对此，欧盟近期拟对基础模型和生成式基础模型提供者设定专门义务，但背离了基于风险分级规制的立法初衷；我国出台专门办法侧重规制大模型部署者，对大模型本身风险管控有限。规制大模型风险，要遵循数据利用安全范式，基于风险分类分级规制，实现上下游的合作共治。按照这一思路构建新型风险规制框架，主要包括设立专门机构引导发展、评估和应对风险，规范数据训练以避免数据泄露和不当输出，基于特定用途风险构建风险分级管控制度，确立贯穿大模型全生命周期的透明度制度，健全防止生成违法内容的上下游共治机制。

关键词：生成式人工智能；大模型；风险规制；透明度；人工智能法

2022 年年底，ChatGPT 横空出世，在语言理解、文本生成和知识推理等方面表现出惊人的“类人”能力，上线两个月日活用户超过一亿，这促使国内外科技巨头纷纷布局其背后的大模型技术，引发“百模大战”。大模型技术驱动生成式人工智能狂飙突进的同时，其基于大数据训练实现的自动化内容生成也引发了新的风险挑战，生成有害内容、数据泄露等问题日益突出。如何防范规制风险，如何平衡好人工智能发展与安全的关系，引导生成式人工智能健康发展，已成为人类社会面临的共同难题。

对此，主要大国都在积极研究应对之策，我国和欧盟在探索专门立法方面走在前列。我国 2023 年 5 月将《人工智能法草案》列入国务院年度立法工作计划，2023 年 7 月制定的《生成式人工智能服务管理暂行办法》，主要规制生成式人工智能服务提供者，并未规制单纯的大模型技术提供者。欧盟 2021 年 4 月提出《人工智能法》提案，2023 年 6 月欧洲议会通过了提案修正案，规定了基础模型提供者的义务，但基本是将大模型比照高风险系统予以监管。目前来看，现有立法探索尚不足以充分规制大模型的新型风险。

本文就是在此背景下，聚焦生成式人工智能大模型技术带来的新型风险，结合欧盟立法经验和我国立法探索，提出规制大模型风险的基本思路和制度框架，以期能够对大模型时代的人工智能立法提供有益参考。

一、人工智能大模型技术的新突破带来了新挑战

大模型技术的新突破，掀起了生成式人工智能发展的新浪潮，开启了迈向通用人工智能的新路径，但大模型技术的新特性也带来了新的风险挑战。

（一）大模型技术驱动生成式人工智能新发展

所谓的生成式人工智能，“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”①。不同于传统的基于规则或模板的简单内容生成，近年来快速发展的生成式人工智能主要是由“深度学习”②技术所驱动的，该技术可以通过训练从数据中自动学习结构和模式，并基于这些模式来生成新的高质量内容，著名的深度学习模型包括了生成对抗网络、扩散模型和转换器模型（Transformer）等。③

其中，Transformer 是一种基于自注意力机制的深度学习模型，可以高效并行地处理序列数据，这使得对大规模数据进行训练成为可能。基于该模型，OpenAI 在 2018 年提出了第一代生成式预训练模型 GPT-1，实现了先以大规模无标注数据进行无监督的模型“预训练”，然后再用有标注数据进行有监督的模型“微调”来更好地适配下游任务，将生成式人工智能带入了“预训练模型”时代。④后来出现的 BERT、LaMDA、T5 等都是基于 Transformer 的预训练模型，这些模型的大规模数据预训练都需要强大算力支撑。⑤

这种生成式预训练模型，又称为通用模型、基础模型，⑥本文称之为生成式人工智能大模型（以下简称“大模型”），是指在大规模数据上训练，具有海量模型参数，可以适应广泛下游任务的模型。⑦可以看出，大模型的生命周期分为以数据训练为主的模型训练阶段和以模型适配为主的模型部署阶段，大模型训练具有基于大数据、依靠强算法、需要大算力的技术特征。而训练完成的大模型，其本身的技术特征可以概括如下：

一是参数规模大：大模型参数规模通常在百万级以上，甚至超过万亿级别，如 GPT-3 的参数达到 1750 亿、北京智源“悟道 2.0”的参数达到 1.75 万亿。需要指出的是，这些参数只是反映了模型所学习的数据，并不会包含或存储模型所学习的数据。

二是生成新内容：基于从训练数据中学习的模式，大模型可以生成新的内容。以 ChatGPT 为例，其从大量现有文本中学习单词在上下文中如何与其他单词一起出现，据此来响应用户请求，预测下一个最有可能出现的单词以及后续每个单词。⑧

---

① 《生成式人工智能服务管理暂行办法》第 22 条第 1 项。

② “深度学习”又称“深度神经网络学习”，是指“通过训练具有许多隐层的神经网络来创建丰富层次表示的方法”。参见我国国家标准《GB/T 41867-2022 信息技术 人工智能 术语》的 3. 2. 27“深度学习”。

③ 参见中国信息通信研究院、京东探索研究院：《人工智能生成内容白皮书》，2022 年 9 月，第 7-9 页。

④ 参见哈尔滨工业大学自然语言处理研究所：《ChatGPT 调研报告》，2023 年 3 月，第 24-28 页。

⑤ 据测算，如果采用英伟达 V100 GPU，训练一次 GPT-3 模型需要 355 GPU 年。See Chuan Li, OpenAI's GPT-3 Language Model: A *Technical Overview*, https://lambdalabs.com/blog/demystifying-gpt-3. (2023-09-26 accessed).

⑥ 参见滕妍、王国豫、王迎春：《通用模型的伦理与治理：挑战及对策》，载《中国科学院院刊》2022 年第 9 期，第 1290 页。

⑦ See Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, et al, *On the Opportunities and Risks of Foundation Models*, ArXiv Preprint, arXiv: 2108. 07258, p. 3 (2022).

⑧ See Yaniv Markovski, How ChatGPT and Our Language Models Are Developed, https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed. (2023-09-26 accessed).

三是涌现新能力：随着参数规模的增大，大模型展现出较小模型所没有的“涌现能力”①，包括小样本提示学习能力、思维链推理能力等，例如 GPT-3 未经专门训练也可以运算两位数的乘法，这些涌现能力加剧了人工智能的不可解释性。

四是呈现通用性：相较于在标注数据上训练执行分类、翻译等任务的专用模型，大模型在大量未标注数据上训练出的海量参数及强大涌现能力，使其得以通过微调等适配方式，去处理各种不同的任务，甚至处理未见过的任务，大模型的通用性大大提高②。

生成式人工智能大模型的出现，标志着人工智能研究范式开始从训练特定任务模型转向训练通用任务模型，展现出通往通用人工智能的可行路径。2023 年 3 月，GPT-4 发布，不仅比之前的大模型表现出更多的通用智能，还可以接受图像和文本的输入，实现了多模态的数据处理。微软研究院认为，GPT-4 的表现已经惊人地接近人类水平，有理由将其视为通用人工智能（AGI）系统的早期版本，堪称“通用人工智能的星星之火”。③生成式人工智能大模型的发展，揭开了迈向通用人工智能的序幕。

（二）生成式人工智能大模型引发的新型风险

由上可知，大模型是依靠大算力对大数据进行训练的结果，其能力来自对大量无标注数据中抽象共现模式的深度学习，④在本质上是大数据驱动的。从寻找大数据中的规律、释放数据价值的方式来看，不同于传统的数据挖掘和分析主要依靠专家标注数据、设计特征等高成本投入，大模型主要是在大量无标注数据上进行无监督学习，自动高效提取数据中的规律和模式，这些规律和模式最终表现为大模型中的大规模参数。通过大数据训练大炼参数而得到的大模型，具备强大的能力和通用性，本身就是训练数据价值的集中体现。因此，笔者认为大模型的训练和调用是一种新的大数据利用方式，大模型是一种高效的大数据价值实现方式。相较于传统的“人工智能系统本身的安全问题”⑤和人为的内容生成，大模型这种新型大数据利用方式，其数据训练和模型调用实现的自动化内容生成引发了新的风险挑战，主要包括以下几个方面：

1. 产生偏见、歧视等有害内容的风险

大模型训练所用的大量数据多为无标注数据，这些数据易存在偏见、歧视，甚至存在侮辱、仇恨、暴力、色情等技术界称之为“毒性”的有害内容，⑥大模型根据从这些数据中学习的模式来生成内容，生成内容便不可避免地会反映出同样的问题。其中最受关注的是偏见、歧视问题。偏见可以理解为一种主观认识和态度，往往会引发客观上对特定人群的区别对待，不公平的区别对

① See Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, *Emergent Abilities of Large Language Models*, ArXiv Preprint, arXiv: 2206.07682, pp. 2-6 (2022).

② See Wayne Xin Zhao, Kun Zhou, Junyi Li, et al, *A Survey of Large Language Models*, ArXiv Preprint, arXiv: 2303. 18223, pp. 15-20 (2023).

③ See Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, et al, *Sparks of Artificial General Intelligence: Early Experiments with GPT-4*, ArXiv Preprint, arXiv: 2303. 12712, p. 1 (2023).

④ See Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, et al, *On the Opportunities and Risks of Foundation Models*, ArXiv Preprint, arXiv: 2108. 07258, p. 48 (2022).

⑤ 涉及人工智能系统本身的完整性、可靠性、稳健性、坚韧性和不可解释性等方面。我国《网络安全法》等相关法律法规已有针对性制度设计，本文不再展开论述，本文关注重点是大模型技术带来的新风险新挑战。

⑥ 这里“毒性”的含义和类型来自 Jigsaw 公司，其 2017 年推出测量语言毒性的 Perspective API，被广泛用于内容审核。See Jigsaw, About the API, https://developers.perspectiveapi.com/s/about-the-api. (2023-09-26 accessed).

待就会导致歧视，例如性别偏见导致的性别歧视。美国国家标准与技术研究院将人工智能偏见分为三大类：系统偏见，指文化和社会中的制度规范、实践和流程造成的偏见；统计和计算偏见，指训练样本代表性不足导致的偏见；人类偏见，指人类思维中的系统性错误。[①]有研究对 DALL-E2、Stable Diffusion 等文本生成图像模型进行了测试，发现当提示输入“CEO”时，生成的都是西装革履的男性图像。[②]出现这种结果的原因，就在于训练数据本身存在系统偏见和统计偏见，不具有公平的代表性。从系统偏见角度看，如果训练数据主要来自某种语言或某个国家，大模型必然会打上这种语言或这个国家文化传统、主流价值观和意识形态的烙印；应该警惕大模型应用可能引发的文化和价值观冲突，防范其可能带来的意识形态安全风险。

2. 泄露个人信息、敏感数据的风险

此种风险主要源于两个方面：一是大模型泄露了训练数据中的个人信息、敏感数据。大模型训练往往采用大规模抓取的网络公开数据，其中可能包含姓名、电话号码等个人信息，甚至可能包括生物识别、行踪轨迹等敏感个人信息和高风险数据。而且，很多大模型默认将用户输入的提示作为训练数据，[③]其中同样可能包含个人信息、敏感数据。研究发现，大模型可能会“记忆”并在特定输入诱导下泄露这些训练数据中的个人信息、敏感数据，包括受版权保护的材料。[④]2023 年 3 月，三星公司在允许使用 ChatGPT 不到 20 天时间里，就被曝出发生了 3 起敏感数据泄露事件，导致其半导体设备测量资料、产品良率、内部会议内容等敏感保密信息泄露。[⑤]二是通过大模型推断出个人信息、敏感数据。大模型涌现出强大的推理能力，可能推断出特定个人的宗教信仰、经济状况等敏感个人信息，甚至可能分析出关系国家安全、公共安全的敏感数据。有研究发现，如果在提示指令中声称正在从事防止核恐怖主义的研究，便可以绕开 ChatGPT 拒绝响应核武器制造提示的安全护栏，而说服其给出如何制造核弹的详细说明。[⑥]虽然此发现公布后不久该提示指令便不再起作用，但确实展现出大模型强大的敏感数据析出能力。

3. 生成错误信息、误导性信息的风险

大模型生成新内容是基于训练数据的内在关联和共现概率。例如，如果在训练数据中“不前进”的高频共现词是“右转”“左转”等，那么在用户输入“不前进”后，大模型就可能按照其参数随机输出“右转”。然而训练数据可能并不具有真实性、时效性或关联性，因此模型输出结果有时便可能是不准确、不真实的，甚至可能会生成错误信息、误导性信息。OpenAI 就指出，ChatGPT 的输出有时可能是不准确、不真实和误导性的，偶尔会产生错误回答，甚至会编造事实或产

① See Reva Schwartz, Apostol Vassilev, Kristen Greene, et al, *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*, NIST Special Publication 1270, 2022, pp. 6-9.

② See Nestor Maslej, Loredana Fattorini, Erik Brynjolfsson, et al, *The AI Index* 2023 *Annual Report*, Institute for Human-Centered AI, Stanford University, April 2023, pp. 28-29.

③ 例如，OpenAI 公司的隐私政策就明确提到，会默认使用用户提供的数据来训练支持 ChatGPT 的模型，只为用户提供了退出机制。See OpenAI, Privacy Policy, https://openai.com/policies/privacy-policy. (2023-09-26 accessed).

④ See Nicholas Carlini, Florian Tramer, Eric Wallace, et al, *Extracting Training Data from Large Language Models*, ArXiv Preprint, arXiv: 2012.07805, pp. 5-8 (2021).

⑤ 参见《三星考虑禁用 ChatGPT？员工输入涉密内容将被传送到外部服务器》，澎湃新闻网，https://www.thepaper.cn/newsDetail_forward_22568264.（最后访问时间：2023 年 9 月 26 日）。

⑥ See Matt Korda, Could a Chatbot Teach You How to Build a Dirty Bomb?, January 31, 2023, https://outrider.org/nuclear-weapons/articles/could-chatbot-teach-you-how-build-dirty-bomb. (2023-09-26 accessed).

生“幻觉”输出。[①]在对信息准确性要求较高的专业领域，如法律和医疗行业，如果仅信赖大模型生成的信息而不加核实，可能会造成重大损害。例如，如果轻信大模型就某些身体不适症状给出的治疗建议，不去就医或错误服用药物剂量等，就可能会延误救治或对身体造成伤害。再如，近期美国纽约州两位律师在提交法院的法律文书中，援引了 ChatGPT 搜集的 6 个案例，但法院发现这些案例都是 ChatGPT 编造的，最终对律师和其律所分别处以了 5000 美元罚款。[②]

4. 滥用于欺骗操纵等违法犯罪的风险

上述三种风险基本都来源于大模型的大数据训练，通常属于非人为故意造成的风险。训练完成的大模型，具有强大的通用能力，存在被故意滥用于实施欺骗操纵等违法犯罪的风险。相较于上述非故意产生的错误信息，大模型可能被故意滥用于制造虚假信息。大模型超强的生成能力，以及其基于大量人类数据训练而具有的“类人”输出和交互能力，使得以低成本方式大规模制造更加逼真、更具欺骗性的虚假信息成为可能，例如可以大量制作更具说服力的网络钓鱼电子邮件。这些大模型生成的更具欺骗性的虚假信息，如果再通过大模型支撑的个性化推荐系统进行推送，鉴于“过滤泡”和“信息茧房”效应，就很可能会造成受众观念极化，甚至会对受众观念和行为进行针对性操纵。[③]这不仅可能侵害私主体权益，更可能对一国的国家安全尤其是政治安全、文化安全等造成严重威胁。例如在俄乌冲突初期，2022 年 3 月在主流社交平台上相继出现了乌克兰总统泽连斯基和俄罗斯总统普京宣布投降的视频，后来都被证实是深度伪造的。[④]此外，大模型也有可能被滥用于实施其他违法犯罪，例如生成恶意软件代码实施网络攻击等。

5. 可能危害环境和社会经济的风险

即使大模型不被滥用，其正常使用也可能会对环境和社会经济造成一定的风险。但目前来看，这些风险似乎并不如上述几类风险那样紧迫和确切，不过从人工智能的发展来看，这类风险很可能在不远的将来成为重大挑战，应该重视和监测这类风险的增长和演变，做到未雨绸缪。例如，大模型的大算力需求，会消耗大量的能源和资源，从而可能造成一定的环境危害。有研究发现，训练 GPT-3 大模型会产生 552 吨二氧化碳，消耗 1287 兆瓦时电力，但也认为 GPT-3 的泛化能力使得不需要针对每个任务重新训练模型，具有潜在的能源优势。[⑤]再如，长期以来，有不少观点认为人工智能将会消灭大量工作岗位。但有研究指出，人工智能工具正在赋予而不是取代人的因素，人工智能如果合乎道德地开发和部署，可以赋予人们做更多事情的能力。[⑥]此外，还有研究关注了大模型应用可能带来的不平等加剧、工作质量降低、创意经济受损等风险。[⑦]

---

① See OpenAI, What is ChatGPT, https://help.openai.com/en/articles/6783457. (2023-09-26 accessed).

② 参见《美国律师因引用 ChatGPT 虚假案例受到处罚》，财联社网，https://www.cls.cn/detail/1385458.（最后访问时间：2023 年 9 月 26 日）。

③ 参见刘金瑞：《数据安全范式革新及其立法展开》，载《环球法律评论》2021 年第 1 期，第 10 页。

④ 参见《外媒：普京宣布投降？泽连斯基宣布投降？其实全是假视频》，腾讯网，https://new.qq.com/rain/a/20220318A04M4100.（最后访问时间：2023 年 9 月 26 日）。

⑤ See David Patterson, Joseph Gonzalez, Quoc Le, et al, *Carbon Emissions and Large Neural Network Training*, ArXiv Preprint, arXiv: 2104.10350, p. 7 (2021).

⑥ See U.S. Chamber of Commerce Technology Engagement Center, *Artificial Intelligence Commission Report*, March 2023, pp. 34-36.

⑦ See Laura Weidinger, John Mellor, Maribeth Rauh, et al, *Ethical and Social Risks of Harm from Language Models*, ArXiv Preprint, arXiv: 2112.04359, pp. 33-35 (2021).

6. 通用性造成风险传导给下游应用

大模型呈现较强的通用性，可以用于解决广泛的下游任务。但这种通用性也意味着大模型自身缺陷会被所有下游模型所继承，大模型自身缺陷引发的风险会传导给下游应用。大模型的自身缺陷主要源于其训练数据的缺陷，因而大模型可以传导给下游应用的风险主要就是其大数据训练引发的风险，由上文可知包括产生有害内容、泄露敏感数据、生成错误信息等风险。大模型向下游应用的风险传导，意味着大模型的风险管控必须依靠大模型价值链上下游参与者的共同努力。其中，最重要的就是训练开发大模型的主体和适配大模型解决下游任务的主体，本文将前者称为大模型提供者，将后者称为大模型部署者。除非大模型的提供者同时也是部署者，否则在一般情况下，由于大模型深度学习算法和涌现能力的不可解释性，大模型部署者在理解和应对大模型风险上存在较大难度，其应对大模型传导的风险，离不开大模型提供者共享必要的技术文件和相关信息。

总结起来，大模型这种新的大数据利用方式，引发的新型风险可以分为两类：一类是模型数据训练引发的风险，主要表现为产生有害内容、泄露敏感数据、生成错误信息等；另一类是模型部署应用引发的风险，主要表现为滥用实施违法犯罪、可能危害环境和经济、向下游应用传导风险等。由前文可知，前一类风险来自模型训练阶段，根源在于训练数据的质量问题和敏感性，例如训练数据集代表性不足、存在有害内容和敏感数据等。后一类风险出现在模型部署阶段，根源在于模型被滥用、模型的负外部性和通用性。面对这些新型风险，技术界正在努力研究有效的缓解措施，通过基于人类反馈的强化学习（RLHF）等，推进大模型与人类价值观和意图对齐，并已取得了一定的成效。例如 GPT-4 相较于 GPT-3.5，生成内容的真实性评估得分高出 40%，对敏感请求（如医疗建议）符合其政策响应的概率提高 29%，对不允许内容的请求响应倾向降低 82%。[①]

尽管这些技术研究努力显著提高了大模型的安全性，但大模型引发的风险挑战仍然较为突出。应对大模型带来的新型风险，仅靠技术层面的缓解措施和对齐手段是远远不够的，还应该探索适应大模型技术特征和发展需要的法律规制手段。

二、生成式人工智能大模型立法的欧盟探索及其镜鉴

对于大模型带来的风险挑战，最早开始探索人工智能综合性立法的欧盟给予了及时回应。虽然欧盟委员会 2021 年 4 月提出的《人工智能法》提案[②]（以下简称“提案”），主要是将人工智能系统按特定用途分成不可接受、高、有限和最小等 4 个风险等级予以分级规制，最初并未涉及没有特定用途的人工智能系统，但随着大模型技术的发展，如何规制大模型和通用人工智能系统，成为欧盟《人工智能法》立法无法回避的问题。2022 年 12 月，欧盟理事会通过了关于《人工智能法》提案的共同立场，[③]专门增加一章“通用人工智能系统”，不过该章主要授权欧盟委员会未来对此立法，并未作出针对性规定。

① See OpenAI, GPT-4, https://openai.com/research/gpt-4, March 14, 2023. (2023-09-26 accessed).

② European Commission, 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, COM (2021) 206 final, 21. 4. 2021.

③ Council of the European Union, 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 - General Approach (6 December 2022), ST 15698 2022 INIT.

欧洲议会对这些问题进行了更为充分的讨论，2023 年 6 月通过了关于《人工智能法》提案的修正案[①]（以下简称“提案议会版”），提出了相对完整的以大模型为中心的通用人工智能风险规制方案。可以预见，在接下来欧洲议会、欧盟理事会和欧盟委员会就《人工智能法》最终文本的“三方谈判”中，大模型规制将是重点议题。以下结合提案议会版的最新案文，对欧盟规制生成式人工智能大模型的制度探索做一简要梳理。

（一）对基础模型和生成式基础模型设定了专门义务

提案议会版将本文所谓的大模型称为“基础模型”，并规定了其提供者的应尽义务。

1. 基础模型提供者的义务

提案议会版新增第 28b 条“基础模型提供者的义务”，在第 2 款明确了基础模型提供者应当遵循的 7 项义务，包括：

一是风险管理义务。要求通过适当的设计、测试和分析来证明，在开发之前和整个开发过程中，以适当的方法识别、减少和缓解对健康、安全、基本权利、环境以及民主和法治的合理可预见的风险，并记录开发后剩余的不可缓解的风险；

二是数据治理义务。要求仅处理和纳入受基础模型适当数据治理措施约束的数据集，特别是审查数据来源的适当性以及可能存在的偏见和适当的缓解措施；

三是技术可靠义务。要求在生命周期内达到适当水平的性能、可预测性、可解释性、可纠正性、安全性和网络安全，并以适当方法进行评估，例如有独立专家参与的评估。

四是环境保护义务。要求适用相关标准来减少能源使用、资源使用和浪费，提高能源效率和系统整体效率；只要技术可行，应测量和记录能源资源消耗及其他环境影响。

五是信息提供义务。要求制定广泛的技术文件和易于理解的使用说明，以使下游提供者能够遵守高风险人工智能系统提供者的义务。

六是质量管理义务。要求建立一个质量管理系统，以确保和记录对该条规定的遵守，并有可能进行试验以满足这一要求。

七是模型登记义务。要求在可公开访问的欧盟高风险人工智能系统数据库中登记。

2. 生成式基础模型提供者的义务

该条第 4 款进一步规定了生成式基础模型提供者的义务。无论是生成式人工智能系统所用基础模型的提供者，还是将基础模型专门化为生成式人工智能系统的提供者，除上述 7 项义务之外，还应当遵循以下 3 项义务：

一是透明度义务。要求遵守第 52 条第 1 款的透明度义务，即人工智能系统、提供者本身或部署者，以及时、清晰和易于理解的方式，告知接触人工智能系统的自然人，他们正在与人工智能系统交互，除非这一点从使用情况和使用场景来看显而易见。

二是防止生成非法内容义务。要求基础模型的训练、设计和开发，应确保按照公认的现有技术水平，有充分的保障措施，防止生成违反欧盟法律的内容，并且不损害包括言论自由在内的基

① European Parliament, Amendments Adopted by the European Parliament on 14 June 2023 on the Proposal for a Regulation of the European Parliament and of the Council on Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, P9_ TA (2023) 0236.

本权利。

三是披露受版权保护的训练数据义务。要求在不损害欧盟或成员国版权立法的情况下，记录并公开提供受版权法保护的训练数据的使用情况的足够详细的摘要。

（二）基于基础模型的人工智能系统的风险分级规制

提案议会版延续并完善了提案基于风险的规制方法。基础模型提供者除了需要遵循上述专门义务之外，一旦基础模型部署集成到人工智能系统之中，其还会适用人工智能系统的风险分级规制。除了最小风险情形下可自由使用外，基于基础模型的人工智能系统，在其特定用途造成不可接受风险、高风险和有限风险时，会受到不同程度的规制。

1. 禁止构成不可接受风险的人工智能系统

提案议会版第 5 条明确列出了应予禁止的人工智能系统，原因就在于欧盟立法者认为这些系统会对人类安全构成“不可接受风险”。在最初的提案中，这些系统包括采用潜意识技术、利用人们弱点和用于社会评分的系统。提案议会版在此基础上对禁止的人工智能实践和系统清单进行了大幅度补充和完善，以禁止操纵性、侵入性和歧视性地使用人工智能系统，主要包括：采用有目的操纵或欺骗技术的系统；强调利用人们弱点的系统，包括利用已知或预测的人格特征或者社会经济状况的系统；公共场所的“实时”远程生物特征识别系统；“事后”远程生物特征识别系统，除非是获得司法授权并且为追诉严重犯罪的执法所必要；使用敏感特征（例如性别、种族、宗教、政治取向等）的生物特征分类系统；预测性警务系统（基于画像、位置或过去的犯罪行为）；执法、边境管理、工作场所和教育机构中的情绪识别系统；不加区分地从社交媒体或闭路电视录像中抓取面部图像来创建面部识别数据库的系统。

2. 高风险人工智能系统的全生命周期义务

欧盟立法者认为，构成高风险的人工智能系统，可以投放欧洲市场，但必须遵守某些强制性要求和进行事前符合性评估。提案第 6 条明确了两大类高风险人工智能系统：一是用作产品安全组件或适用附件 2 欧盟健康和安全协调立法的系统（例如汽车、医疗器械领域的系统等）；二是在附件 3 确定的八个特定领域部署的系统，欧盟委员会可以通过授权立法进行必要的更新。八个特定领域为：自然人的生物特征识别和分类；关键基础设施的管理和运作；教育和职业培训；就业、工人管理和自营职业；获得和享受基本的私营服务和公共服务及福利；执法；移民、庇护和边境控制管理；司法行政和民主程序。

提案议会版延续了这一分类规则，完全保留了提案的附件2，但对第二类高风险系统的认定和附件 3 的内容提出了重大修改。欧洲议会认为，属于附件 3 的八个特定领域的系统并不会自动归类为高风险系统，而是必须满足额外的限定条件，即“对自然人的健康、安全或基本权利构成重大损害风险”，才会被认为是高风险系统。还进一步补充和完善了附件 3 各个领域的表述，将第一个领域修改为“生物特征和基于生物特征的系统”，在各个领域之下增加了一些新的高风险系统，包括第 5 条规定之外的情绪识别系统、评估个人教育和职业培训水平的系统、决定个人健康和人寿保险资格的系统等，尤其是纳入了影响政治竞选中选民投票的系统和超大型社交媒体平台用于

推荐的系统。①

提案的第 3 编对高风险人工智能系统提供者规定了贯穿整个系统生命周期的义务和要求，涉及风险管理、数据和数据治理、技术文件、记录保存、透明度和信息提供、人类监督、技术可靠性（第 9-15 条），以及质量管理、符合性评估、利益相关方义务等，其他编还规定了登记到欧盟高风险系统数据库、上市后监测、报告严重事故等义务，提案议会版延续了这些规定，并进一步完善了相关表述。

3. 有限风险人工智能系统负有透明度义务

欧盟立法者认为，构成有限风险的人工智能系统，其风险主要是特定的操纵风险，为使人们避免被操纵，得以做出知情选择或后退决定，这些系统应负有透明度义务。②提案第 52 条列出了三种此类系统及相应的透明度义务：一是与人类交互的系统，应当告知正在与人工智能系统交互；二是根据生物特征数据进行情绪识别或社会分类的系统，应当告知系统的运行；三是生成或操纵图像、音频或视频等内容（“深度伪造”）的系统，应当披露内容是人为生成或操纵的，但出于法律授权等合法目的时除外。提案议会版延续了这一规定，并根据不同系统的特点，完善了相应的告知内容。

（三）欧盟立法探索的经验与镜鉴

总结来看，对于大模型驱动的本轮通用人工智能发展浪潮，欧盟立法者及时关注并回应了其中的风险挑战：基于风险分级规制的方法，以大模型为中心，对于基础模型、生成式基础模型、基于基础模型的人工智能系统设定了不同义务要求，提出了通用人工智能分层监管方案。但从目前最新的提案议会版来看，基础模型提供者承担的义务，除了环境保护义务和生成式情形下的防止生成非法内容和披露受版权保护的训练数据义务外，其在风险管理、数据治理、技术可靠性、信息提供、质量管理、模型登记等方面的义务，与高风险人工智能系统承担的义务基本一致。这说明欧盟立法者实际上将基础模型本身归类为高风险，比照高风险人工智能系统对基础模型予以监管。

但从风险管理和技术可靠性等要求来看，以这种思路监管基础模型实际并不可行。考虑到基础模型的通用性，在风险管理方面，要求识别、减少和缓解所有合理可预见的风险，至少就要考虑和分析《人工智能法》法案附件 3 中所有高风险用途中的可能风险，然后在此基础上制定和实施针对所有这些风险的缓解措施；在技术可靠性方面，要求达到并评估适当水平的性能、安全性等，就需要针对所有高风险用途在这些技术方面进行可靠性测试和评估。满足这种监管要求，需要付出不可估量的成本和代价，是基本不可能完成的任务，而且让大模型提供者针对所有假设性的可能用途实施风险缓解和可靠性保障，而很多的可能用途最终又不会实现，似乎也没有必要。

深究来看，风险分级规制方法，是以人工智能系统特定用途来确定风险分级的，欧盟立法者将具有通用性的大模型直接按照高风险人工智能系统予以监管，其实并没有考虑大模型实际特定用途的不同风险，也没有看到往往是大模型部署者而非提供者决定其实际用途，更没有考虑大模

① 根据欧盟《数字服务法》的规定，这种超大型社交媒体平台是指拥有超过 4500 万活跃用户的平台。

② See European Commission, 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, COM (2021) 206 final, 21. 4. 2021, pp. 14-15.

型数据训练和部署应用引发两类不同的风险。这种削足适履、一刀切的立法方案，实际背离了基于风险分级规制的初衷。应该从大模型技术特征和新型挑战入手，探索切实有效、适应通用人工智能发展需要的大模型风险规制方案。

三、生成式人工智能大模型新型风险的规制框架

2023 年 7 月，国家网信办等七部门制定了《生成式人工智能服务管理暂行办法》（以下简称《暂行办法》），其适用范围聚焦于利用生成式人工智能技术向我国境内公众提供生成式人工智能服务的提供者，而排除了仅“研发、应用生成式人工智能技术”的企业、科研机构等。这实际区分了生成式人工智能的服务提供者和技术提供者，后者若并不向境内公众提供生成式服务，则不适用《暂行办法》。那么对于基于大模型技术的生成式人工智能服务来说，《暂行办法》规制的是大模型部署者，并未规制单纯的大模型提供者。这种将规制重点放在大模型部署应用、鼓励大模型训练开发的监管思路，坚持了发展和安全并重，有利于我国大模型和通用人工智能业态的创新发展，值得高度肯定。

但从前文分析来看，大模型引发的新型风险有一类根源于模型训练阶段，大模型通用性会将这些风险传导给下游应用，应对这种风险离不开大模型价值链上下游参与者的沟通合作。从这个角度看，除了大模型提供者本身就是部署者的情况之外，目前《暂行办法》第 7 条的训练数据合法性要求、第 14 条的“模型优化训练”整改措施等规定，仅适用于下游的大模型部署者，只规范部署者适配大模型时的小规模数据训练，实际无法管控上游的大模型基座及其风险，并不能充分解决大模型带来的风险挑战。因此，从生成式人工智能大模型的风险防范和发展需要来看，《暂行办法》仍有较大完善空间。

基于大模型的技术特征，针对其引发的新型风险，结合国内外的立法进展，本文就规制大模型风险的基本思路和制度框架提出以下建议。

（一）大模型风险规制基本思路

规制大模型这种新的大数据利用方式引发的新型风险，应该遵循数据利用安全范式，采用基于风险的分类分级规制方法，实现上下游参与者的合作共治。

1. 遵循数据利用安全范式

大模型是利用大数据训练大炼参数而来，其训练开发和部署应用是一种新的大数据利用方式，这种新的数据利用方式引发了新的风险挑战。规制大模型的新型风险，就应该遵循数据利用安全范式。《中华人民共和国数据安全法》确立了数据安全新范式，要求“确保数据处于有效保护和合法利用的状态”，既确保传统的数据“自身安全”，也确保数据大规模流动和挖掘的“利用安全”。而其中的数据利用安全范式，笔者认为关键就是确保数据大规模流动和利用的可控性和正当性。[①]对于大模型而言，其数据利用包括两个方面：一是模型训练阶段提取大数据中规律和模式的数据训练，训练结果表现为大模型，尤其是其中的算法和参数；二是模型部署阶段基于数据训练结果即大模型来响应数据输入生成新的内容。简言之，大模型的数据利用包括利用数据去训练大模型和再利用数据训练结果即大模型来生成内容。而由上文可知，大模型的两类新型风险即模型

① 参见刘金瑞：《数据安全范式革新及其立法展开》，载《环球法律评论》2021 年第 1 期，第 11 页。

数据训练引发的风险和模型部署应用引发的风险，正是根源于这两方面的数据利用。

遵循数据利用安全范式规制大模型的新型风险，就是既要确保大模型数据训练的可控性和正当性，也要确保大模型部署应用的可控性和正当性。具体而言，对于模型训练阶段，一方面训练数据的收集和聚合要符合可控性，重要数据、个人识别信息等敏感高风险数据不应纳入训练数据；另一方面鉴于数据训练结果会决定模型生成内容，训练数据的来源和内容要符合正当性，例如要尽量排除存在偏见、歧视的数据。对于模型部署阶段，一方面模型的具体部署要符合可控性，例如不得将模型用于存在不可接受风险的领域，将模型用于高风险用途时应规定相应的风险管控义务；另一方面模型的使用目的和生成内容要符合正当性，不得将模型用于违法目的，应防止模型生成违法内容。

2. 基于风险分类分级规制

《暂行办法》强调要“实行包容审慎和分类分级监管”。笔者认为，落实该原则性规定的关键，在于对大模型进行基于风险的分类分级规制。大模型的风险分类，本文认为应主要根据大模型的风险根源和风险领域来区分，可以分为大模型系统本身存在的传统风险和大模型数据利用引发的新型风险，后者如前文所述又包括模型数据训练引发的各种风险和模型部署应用引发的各种风险。对于前者的规制，目前已有《中华人民共和国网络安全法》（以下简称《网络安全法》）等相关法律法规作出专门规定；对于后者的规制，要遵循上述数据利用安全范式，探索确保大模型数据训练及部署应用可控性和正当性的制度设计。大模型的风险分级，本文认为应主要根据大模型实际特定用途的风险程度来确定，借鉴欧盟《人工智能法》提案，可以将风险分为不可接受风险、高风险、中风险和低风险等 4 个等级。

基于风险的分级规制，需要将规制重点放在大模型实际用途上，并根据风险等级匹配不同监管方式。而管控风险一般认为包括四种策略：接受风险、避免风险、控制风险以及转移风险。[①]对于大模型存在不可接受风险的用途，应该力求避免风险，原则上予以严格禁止；对于大模型的高风险、中风险用途，应该侧重控制风险，规定与风险等级相适应的风险管控义务。由于不可能实现绝对安全，对于大模型的低风险用途，以及采取避免风险、控制风险措施后仍然存在的残留风险，妥当策略是接受风险的存在，这也是“坚持发展和安全并重”的应有之义。从这个角度看，无论是欧盟《人工智能法》提案议会版一刀切地将大模型作为高风险系统监管，还是《暂行办法》征求意见稿曾要求生成内容“应当真实准确”等，都没有对风险进行分级规制，都有追求绝对安全之嫌，确实在一定程度上忽视了风险防范与产业发展的平衡。

3. 实现上下游的合作共治

妥当应对和规制大模型向下游应用的风险传导，离不开大模型价值链上下游参与者的共同努力。确定上下游参与者各自应当承担的风险管控义务，就需要厘清大模型用于下游任务时，相关主体的不同角色及其对大模型的控制水平。目前“下游大模型部署者”[②]调用大模型的方式主要有

① See U. S. Department of Homeland Security, Risk Management Fundamentals, April 2011, p. 23, https://www.dhs.gov/sites/default/files/publications/rma-risk-management-fundamentals.pdf. (2023-09-26 accessed).

② 本文从价值链上下游角度提及大模型提供者和部署者时，不包括提供者同时也是部署者的情形。

两种：开源访问和 API（应用编程接口）访问。[①]在开源访问的情况下，提供者会公开模型的参数和源代码，部署者可以直接检查源代码和参数并根据开源许可进行修改和适配。在 API 访问的情况下，提供者仅向部署者提供大模型的 API 调用接口，部署者可以利用一些训练数据微调模型以适配下游任务，但无权修改模型的源代码和参数。但不管是开源访问还是 API 调用，都是部署者决定大模型的实际用途。

可见，对于大模型部署者来说，在 API 模式下，大模型的源代码和参数仍完全控制在提供者手中，其无法知晓大模型的底层技术细节，也无法通过修改大模型来应对风险，即使在开源模式下可以修改模型源代码和参数，但考虑到大模型算法和涌现能力存在不可解释性，其实际并不能完全理解和管控大模型上游数据训练带来的风险。而对于大模型提供者来说，其无法介入部署者适配模型的数据训练，也无法干预部署者决定大模型的实际用途，并不能管控模型适配数据训练和模型部署应用引发的风险。因此，全面管控大模型应用风险，单靠部署者抑或提供者都不可行，需要实现二者的合作共治。从这个角度看，无论是《暂行办法》仅从大模型部署者入手防范风险，还是欧盟提案议会版将风险管理的重任仅赋予大模型提供者，都是让他们去完成不可能完成的任务，实际无法达成全面管控大模型风险的目标。此外，利用基于大模型的人工智能系统生成内容的最终使用者，决定了生成的具体内容及内容受众，使用者和内容受众也是参与治理大模型生成内容风险的重要主体。为了实现大模型风险的合作共治，这些上下游参与者既需要进行充分的风险沟通和信息共享，也需要协作采取必要的风险应对措施。

（二）大模型新型风险规制框架

根据上述基本思路，规制大模型引发的新型风险，重点是确保大模型数据训练及部署应用的可控性和正当性。对此，结合国内外立法进展，本文提出以下规制框架。

1. 设立专门机构引导发展、评估和应对风险

建议借鉴域外经验设立专门的人工智能监管机构，名称可为“人工智能发展和安全委员会”，以全方位监测和应对大模型等技术路径带来的风险挑战，引导和促进人工智能安全发展。2018 年底，美国依据《2019 年国防授权法》，设立了“国家人工智能安全委员会”，负责审查人工智能、机器学习和相关技术的发展，以全面解决美国国家安全需要。欧盟提案议会版提出设立欧洲人工智能办公室，以确保该法有效和协调执行，其职责明确包括对大模型的监管：提供特别的监督和监测，就基础模型及利用这些模型的人工智能系统是否合规，以及行业自我治理的最佳实践，与基础模型提供者建立定期对话制度；记录并监测已知的大模型大型训练的运行情况，以及发布基础模型发展、扩散和使用状况的年度报告，并附上应对基础模型特有风险和机遇的政策选择。

笔者认为，专门机构对于大模型的风险监管，除了监督落实法定义务之外，可以侧重以下两个方面：一是组织对高性能大模型进行强制性风险评估。目前欧盟提案规定的大模型风险评估，主要是提供者和部署者的自我评估。《暂行办法》第 17 条提到的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安全评估，也属于自我评估，仅侧重信息内容安全风险。考虑到大

① See Sabrina Küspert , Nicolas Moës , Connor Dunlop, The Value Chain of General-purpose AI, https://www.adalovelaceinstitute.org/blog/value-chain-general-purpose-ai/.（2023-09-26 accessed）.

模型的潜在风险可能影响巨大，为避免造成难以承受的后果，建议由专门机构组织相关领域专家，对某些高性能大模型在其上市前进行强制性的第三方风险评估。这种强制性评估的范围应该限于具有强大能力的大模型，可以从大模型性能指标入手进行界定，比如限于参数超过 1 亿或计算量超过一定门槛的大模型。鉴于对大模型技术的评估才刚开始探索，专门机构应该负责组织研究和发展针对大模型的评估方法和评估标准。基于风险评估的结果，专门机构可以组织制定大模型的风险应对措施。二是加强对大模型远期风险的监测、研究和应对。如前所述，大模型正常使用也可能对环境和社会经济造成一定的风险，比如对环境和劳动就业的影响，但目前研究显示这些风险尚未构成迫切威胁，可称之为一种远期风险。对于这类远期风险，目前妥当应对策略是接受风险，但也需要加强监测和研究，一旦发现其转变为现实威胁，应该及时调整应对措施。为监测需要，大模型提供者应有义务定期报告资源消耗等情况。

2. 规范数据训练以避免数据泄露和不当输出

由上文可知，规制大模型数据训练引发的新型风险，就是要确保数据训练的可控性和正当性，最大限度地避免数据泄露以及因数据训练导致模型生成有害甚至违法内容，提高生成内容的准确性和可靠性。大模型提供者和部署者的数据训练要满足以下要求：

一是确保训练数据的收集和聚合符合可控性。由于大模型可能会“记忆”部分训练数据，其强大的推理能力也会分析出敏感信息，因此应该尽量避免将敏感的高风险数据作为训练数据，以最大限度地避免敏感数据泄露。笔者认为，关系国家安全和公共利益的重要数据、关系个人权益的隐私和个人识别信息、关系企业商业利益的商业秘密等，原则上不应作为公开大模型的训练数据。

二是确保训练数据的来源和内容符合正当性。数据来源正当，主要就是《暂行办法》强调的“具有合法来源”，不是以非法方式获取的数据。笔者认为，很多大模型默认将用户输入的提示作为训练数据、只为用户提供退出机制的做法，正当性值得质疑，至少在用户输入构成个人信息的时候，应当原则上征得个人同意或者具有其他合法依据。数据内容正当，主要就是数据承载的信息内容不得违反法律法规的禁止性规定，不得侵害他人合法权益，符合一国的主流价值观和意识形态，尽量排除存在偏见、歧视等不公平内容的数据，确保训练数据质量。《暂行办法》已明确要求采取有效措施“增强训练数据的真实性、准确性、客观性、多样性”。不过从防止产生歧视的角度看，本文认为还应该要求训练数据具有充分的代表性、与预期目的的相关性，以及采取适当的偏见检测和纠正措施，以最大限度地降低大模型中嵌入不公平偏见的风险。

3. 基于特定用途风险构建风险分级管控制度

如前文所述，规制大模型部署应用引发的新型风险，就是要确保大模型部署应用的可控性和正当性，规制对象主要是大模型部署者。确保可控性，势必需要对大模型特定用途的风险进行分级管控，按前述建议就是分不可接受风险、高风险、中风险和低风险予以不同程度的规制。从这个角度看，《暂行办法》侧重规制部署者确有道理，但仅提出分级监管原则并未规定风险分级规制，在风险管控方面仍有较大完善空间。根据前述风险管控策略，应该允许低风险用途，禁止存在不可接受风险的用途，就高风险和中风险用途设定相适应的风险管控义务，后两个方面是分级管控制度的重点所在。具体而言：

一是明确大模型部署应用的“禁止清单”。禁止清单主要是理清存在不可接受风险的用途，笔者认为主要是指可能严重危害国家安全、公共安全和重大公共利益，会造成难以承受后果的用途。这种不可接受性往往取决于一国的核心价值观、国家利益和文化传统等，不同国家可能会有不同的界定。从欧盟提案的界定来看，是基于欧盟维护基本权利等价值观，侧重禁止操纵性、侵入性和歧视性的用途。欧盟界定中提到的采用潜意识或欺骗技术、利用弱势群体弱点等操纵人的行为等用途，可以为我国界定这种禁止性用途所借鉴；但对于欧盟拟禁止的预测性警务等用途，本身在欧盟立法过程中就引起很大争议，是否列入我国的禁止清单需要进一步论证。还需要指出的是，欧盟提案由于立法权力所限将军事用途排除在适用范围之外，笔者认为我国未来立法应该明确禁止将大模型用于自主武器系统、核威慑等军事用途。①

二是明确界定高风险、中风险用途并规定相应的风险管控义务。笔者认为高风险用途主要是指可能危害国家安全、经济运行、社会稳定、公共健康和安全等的用途。欧盟提案议会版对高风险用途的界定采用了“特定领域列举+抽象要件认定”的方法，即列出了关键基础设施管理和运作等八个领域里可能存在高风险用途的系统，然后根据“对自然人的健康、安全或基本权利构成重大损害风险”限定条件具体认定是否构成高风险用途。这种界定方法可以为我国所借鉴。对于决定大模型高风险用途的部署者，借鉴欧盟提案，可以从风险管理、透明度、记录保存、技术可靠性等方面规定其风险管控义务。在风险管理方面，对于是否可以部署某种高风险用途，建议借鉴欧盟提案议会版第 29a 条高风险系统部署者“基本权利影响评估”义务，要求部署者建立利益相关者尤其包括受影响者在内多方参与的风险评估机制。至于中风险用途，借鉴欧盟“有限风险”的理解，笔者认为主要是指因运行不透明而导致人们可能被自动化系统误导、操纵，可能危害人的自主性的用途。为避免这种风险，应规定此时部署者负有一定的透明度义务，要告知使用者人工智能系统存在和运行情况，保障人们在知情后有权选择是否使用系统。

4. 确立贯穿大模型全生命周期的透明度制度

为有效应对大模型的新型风险，大模型的上下游参与者既需要直面大模型的不可解释性，努力理解大模型训练部署和运行输出的原理，也需要进行充分的风险沟通和信息共享，促成风险的合作共治。而做到这两方面的关键，在于确立贯穿大模型全生命周期的透明度制度。具体而言，包括三个方面：

一是大模型提供者及高风险用途部署者的信息公开义务。从欧盟立法来看，提案最初只规定了高风险人工智能系统才负有透明度义务，提案议会版明确要求基础模型应在欧盟高风险人工智能系统数据库中登记并按附件 8 要求公布相关信息。笔者赞同这一思路，不论大模型是否用于高风险用途，都有必要保持一定的透明度。借鉴欧盟的规定，并结合前文所述，笔者认为大模型提供者应该公布以下信息并保持更新：提供者名称等基本信息；大模型训练数据的来源；大模型的能力、局限性以及合理可预见的风险缓解措施；大模型训练所需的计算能力以及对环境的可能影响；大模型按照公共或行业基准具有的性能；大模型内外部测试和优化的说明等。大模型高风险用途部署者也应该参照这些内容公布模型部署应用情况，并着重说明高风险系统的预期用途、局

① 参见刘金瑞：《人工智能的安全挑战和法律对策初探》，载《中国信息安全》2018 年第 5 期，第 73 页。

限性、潜在风险及缓解措施。在信息公开形式上，我国可以借鉴欧盟建立可公开访问的数据库。

二是大模型中风险用途部署者及使用者的透明度义务。除了前述大模型中风险用途部署者负有透明度义务外，基于大模型的中风险系统的使用者，对受到系统影响的人也应负有一定的透明度义务。例如，欧盟提案议会版第 52 条新增规定：与人类交互的系统的使用者，利用系统做出决策时，应当告知接触系统的人，谁负责决策过程以及现有的权利和程序，这些权利和程序允许反对适用系统并就系统所做决策或所致损害寻求司法补救，包括寻求解释的权利；未被禁止的情绪识别系统或生物特征分类系统的使用者，应当在处理生物特征数据和其他个人数据前征得接触系统的人的同意；“深度伪造”系统的使用者，应当以适当、及时、清晰和可见的方式披露内容是人为生成或操纵的。

三是大模型价值链上游参与者向下游参与者提供必要信息的义务。大模型提供者应当向部署者、大模型高风险用途部署者应当向使用者，提供必要的技术文件和使用说明，以支持下游人工智能系统的正常运行和依法使用，尤其是符合高风险系统的监管要求。欧盟提案附件 4 规定这些信息包括人工智能系统的一般描述、要素和开发过程的详细说明、运行和控制的详细资料、风险管理的详细描述等。提案议会版认为还应当包括：系统的主要目标、输出质量和输出可解释性；系统的结构、设计规格、算法和数据结构以及它们的彼此联系和整体逻辑；特定系统性能指标的适当性；系统开发的能源消耗以及使用的预期能源消耗等。笔者认为，必要信息的提供，应该考虑上下游参与者之间约定的大模型利用方式，并在技术信息共享和商业秘密保护之间取得适当平衡。

5. 健全防止生成违法内容的上下游共治机制

规制大模型部署应用引发的风险，除了要确保具体部署的可控性，还要确保特定用途的正当性，这就既需要禁止将大模型用于违法目的，也需要防止大模型生成违法内容，而后者涉及机器生成违法内容的新挑战，是规制大模型新型风险的关键所在。针对大模型生成内容的信息内容安全风险，欧盟提案议会版和《暂行办法》都明确要求防止生成违法内容。大模型之所以生成违法内容，既可能是源于大模型训练数据缺陷的模型有害输出，也可能是大模型被故意部署滥用于生成违法内容，因此防止大模型生成违法内容需要上下游的合作共治。前述大模型提供者的数据规范训练义务、大模型部署者的风险管控义务等，都是尽量避免大模型生成违法内容的有力措施。

但如果基于大模型的人工智能系统还是生成了违法内容，应该如何及时发现与处置呢？对此，《暂行办法》要求生成式服务提供者即大模型部署者应当承担网络信息内容生产者责任，发现违法内容及时采取停止生成等处置措施和模型优化训练等整改措施，并向有关主管部门报告。不过细思之下，虽然部署者提供了生成式服务，但该服务是基于大模型提供者的技术，下达指令决定具体生成内容及受众的是服务的使用者，使用者才是其中最重要的内容生产者。仅将大模型部署者认定为内容生产者，实际是让部署者对使用者生成内容行为直接承担全部责任，考虑到使用者利用生成式服务会生成海量内容，这种要求似乎过于严苛；让部署者采取模型优化训练的整改措施，但除了开源模型等情况外，部署者往往无法对上游大模型进行优化修改，实际根本无法完成整改目标。

这说明仅依靠部署者难以完成发现和处置违法内容的重任，应当进一步健全发现和处置违法

内容的共治机制。一是健全违法内容发现的共治机制。目前《暂行办法》强调部署者有义务“发现”并处置违法内容，不过考虑到使用者会生成海量内容，要求逐一人工审查几乎是不可行的，因此参照《网络安全法》第 47 条的规定和理解，笔者认为不能将部署者发现违法内容的义务理解为让其对所有生成内容承担普遍审查义务，发现违法内容需要部署者、使用者和主管部门的共同参与：除了部署者负有一定的“主动”发现义务，即应该根据现有技术水平采取人工审核监督、识别过滤措施等手段，积极查找违法内容之外，还应该畅通违法内容的举报机制，完善主管部门的举报处理和巡查机制，动员广大使用者积极举报违法内容；部署者对于通过使用者举报、主管部门告知等途径“被动”获知的违法内容，应当及时予以处置。二是健全违法内容处置的共治机制。当部署者发现自己不足以阻却违法内容生成时，应及时将有关情况告知大模型提供者，由提供者采取修改模型参数、模型优化训练等措施进行整改，并向有关主管部门报告。当然如果发现生成违法内容是由于部署者适配模型造成的，应当由部署者通过模型适配优化训练等措施进行相应整改，不应将整改责任强加给上游的提供者。

Regulatory Framework for New Risks of Large Generative AI Models

LIU Jin-rui

(The Law Institute of China Law Society, Beijing 100081)

Abstract: The advancement of large model technology has greatly driven the development of generative AI, but its data training and deployment have also generated new risks and challenges, including harmful contents, sensitive data leakage, misinformation, misuse for illegal activities, possible environmental and economic harms, risk transmission to downstream, etc. The EU recently intends to set specific obligations for providers of foundation models and generative foundation models, but it deviates from the original legislative intention of regulating based on risk classification. Our country has issued special measures to regulate deployers of large models, but they only have a limited control of the risks of large models. To regulate the risks of large models, it is necessary to follow the data utilization security paradigm, regulate based on risk categorization and classification, and achieve the co-governance of upstream and downstream participants. The key is to construct a regulatory framework for new risks, which mainly includes setting up a special agency to guide development, assessing and responding to risks, regulating data training to avoid data leakage and illegitimate output, building a control system based on risk classification by the risks of specific uses, establishing a transparency system throughout the entire life cycle of large models, and improving the co-governance mechanism of upstream and downstream to prevent the generation of illegal contents.

Key Words: Generative AI; Large Models; Risk Regulation; Transparency; Artificial Intelligence Act

(责任编辑：王青斌)