CAICT 中国信通院

中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

集智 蓝皮报告

大模型治理蓝皮报告

——从规则走向实践

（2023 年）

中国信息通信研究院政策与经济研究所

中国科学院计算技术研究所智能算法安全重点实验室

2023年11月

版权声明

前　言

近一年来，以 ChatGPT 为代表的大模型技术引发通用人工智能新一轮发展热潮，在带动大规模产业升级、劳动力转移、产品的分配机制等方面均带来深刻变革，成为改变世界竞争格局的重要力量。与此同时，围绕人工智能治理的议题探讨显著增多，全球人工智能治理体系加速构建。党中央国务院高度重视人工智能治理工作，作出一系列重要部署。习近平总书记强调，要共同做好风险防范，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。寻找大模型治理的准确定位，构建适应技术发展趋势的治理体系愈加重要且迫切。

面对大模型带来的新问题新挑战，传统监管模式面临着 AI 自主演化控制难、迭代快速跟进难、黑箱遮蔽追责难等问题，一劳永逸的事前监管模式已经难以应对不断推陈出新的人工智能发展需求。从治理框架来看，敏捷治理成为回应大模型技术快速迭代的重要治理模式，软硬法协调、多主体协同的治理需求更为突出。构建高质量数据集、创新知识产权制度、探索价值对齐实现方式、维护信息内容安全等成为各方关注的热点问题。美国、欧盟、英国等主要国家和地区加紧推进人工智能治理布局，共同寻求具有共识和互操作性的治理规则。我国围绕人工智能发展、安全、治理三方面提出《全球人工智能治理倡议》，通过算法备案、评估评测、事后溯源检测等方式，推动人工智能治理从规则走向实践，为全球提供人工智能治理中国方案。希望研究成果为社会各界进一步参与大模型治理实践提供有益参考。

目 录

一、大模型治理的重要性紧迫性凸显

（一）大模型技术浪潮兴起

当前，世界人工智能领域科技创新异常活跃，日益成为改变世界竞争格局的重要力量。一批里程碑意义的前沿成果陆续突破，以 ChatGPT 为代表的大模型技术引发通用人工智能新一轮发展热潮。

1.对大模型的基本认识

大模型（LLM，Large Language Model）指的是具有超大参数规模，建立在多头自注意力机制 Transformer 架构之上，以深度神经网络为基础，用海量文本数据预训练而成的语言模型。以 ChatGPT 为代表的大模型能够模拟人类的创造性思维，生成具有一定逻辑性和连贯性的语言文本、图像、音频等内容。**大模型基于大数据、大算力、多模态的技术优势，实现从感知世界、理解世界向创造世界的跃迁，推动人类社会加速迈向人机共生的智能社会阶段。**

大模型体现出三方面技术趋势：一是从决策式 AI 到生成式 AI。决策式 AI 主要是通过分类回归对数据进行分析，主要应用于图像识别、推荐系统、决策智能体等领域。生成式 AI 借助 Transformer 架构等，具有全局表征能力强、高度并行性、通用性强、可扩展性强等优势，主要应用于内容创作、科研、人机交互等领域，实现了从简单感知到内容创造的跃迁。**二是从单模态模型到多模态模型。**多模态是指通过处理和关联来自多种模态的多源异构数据，挖掘分析信息、提高模型能力的学习方法。典型任务是图像/视频/语言间的跨模态预训练、跨模态定位等，如给定文本生成一段对应的声音、图像/视频与文本

的相互检索或生成等。**三是从亿级到千亿、万亿级参数的预训练模型。**大模型指的正是模型参数规模庞大，大模型参数规模从亿级发展到百亿、千亿级别，并向着更高规模的参数探索。例如，GPT-3 参数量达 1750 亿，文心一言参数规模为 2600 亿等。随着参数规模的增长，模型能力也得到显著提升。

2.大模型的变革影响

（1）内容生产方式的“颠覆者”

大模型实现了高质量、高效率、多样化的内容生产，成为推动内容生产方式变革的重要力量。**一是信息内容生产主体发生显著变革。**人工智能在信息收集、筛选和整合、推理的全过程都能替代人力，极大地解放人力资源。**二是信息内容生产效率快速提升。**大算力驱动强算法处理大数据，在自然语言处理、计算机视觉、自动驾驶、等各领域多种任务上，都能高质量作出结果判断，高效率进行内容生成。**三是信息内容传播出现颠覆性变化。**信息的生产、传播更加便利，尤其是降低了专业知识的获取门槛。信息内容的表现形态更加丰富，利用人工智能创生技术，图、文、代码等相互转换更加自由，可以一键生成“数字人”分身，开启智能互联时代。

（2）通用人工智能的“先行者”

大模型是迈向通用人工智能的重要技术探索。**一是具备了与人类智能相媲美的综合智能能力。**大模型的能力不再局限于自然语言、视觉等特定方面，而是具备了执行一般智慧行为的能力，广泛拓展了人工智能技术的适用范围。**二是具备了通用型技术能力的潜力。**业界普

遍认为，大模型是智能时代的关键基础底座，各领域不再需要单独开发人工智能，仅需调用大模型接口即可。将来可能构建出新的应用生态、创造新的用户接口，并带来潜在商业模式的变革。**三是具备了赋能千行百业的适应性。**大模型可作为底层技术，垂直应用于各个产业和复杂场景。这种可以嫁接千行百业的智能生产力，正在重塑和影响未来生活。

（3）人机交互的“协作者”

大模型使得人类行为与机器运行之间的协作更加自然、高效和智能，拓展了更为广阔的人机交互空间。**一是呈现出极大的语言表达的自由度。**大模型“善于”理解和生成自然语言，人们可以自由提问或表达需求，不必担心特定的格式或指令。这种自由度使得人与机器的交互更为自然、灵活。**二是呈现出极为个性化的交互体验。**大模型可以通过分析和理解用户的喜好、兴趣和上下文信息，进行定制化的服务和建议。大模型的即时回应和连续对话，给用户提供了实时的帮助和引导。Web 2.0 之父蒂姆·奥莱利（Tim O'Reilly）认为未来的人工智能系统将成为人类工作的协作者，通过人机合作实现更强大的效果。

（二）大模型引领数字化变革

大模型体现出强智能性、强通用性、强交互性，为进一步的产业革新与数字政府发展奠定了坚实的基础。根据麦肯锡报告，生成式人工智能每年或将为全球 GDP 增加 2.6-4.4 万亿美元。根据 Markets and Markets 报告，2023 年全球生成式人工智能的市场规模预计为 110.3 亿美元，2028 年预计将达到 518 亿美元，年复合增长率达 35.6%。

1.大模型推动产业跨域深度融合

凭借大数据、大市场、多场景优势，人工智能与交通、医疗、工业等传统行业深入融合，涌现出一批新业态新模式。在工业领域，大模型实现汽车、建模等设计的自动优化、打造 3D 模型、通过智能物流、智能安防实现智能化管理；在医疗领域，大模型实现蛋白质分子的结构预测、辅助医生影像读片与分析病例报告，推出 AI 陪护与交互式心理咨询；在金融领域，大模型催生了数字员工，借助 AI 客服、AI 投资顾问、AI 财务实现服务的自动化，并进一步优化投资管理与风险管理。据埃森哲预计，2035 年人工智能应用将为制造业带来 4 万亿美元额外增长，年增长率可达 4.4%。

2.大模型提升公共服务水平

当前，公共领域大模型应用如火如荼，为公共服务提质增效。美国、英国、葡萄牙、新加坡等 13 个国家或地区已将 ChatGPT 应用于政府内部办公、公共服务提供等场景。据日本野村综合研究所开展的网络问卷调查显示，政府部门对 ChatGPT 的利用率达 17.5%，仅次于信息通信业（32.8%）和制造业（19.2%）。从市场份额来看，根据普鲁杜尔公司（Prudour Pvt.Ltd）数据显示，2022 年各国政府应用大模型的市场份额超过 1 千万美元，预计 2032 年超过 5 亿美元，年复合增长率达 45.8%。大模型技术的引入可以显著提升人机交互的友好程度。同时，大模型在信息收集、数据分析以及语言重塑能力层面的优势，能够有效协助整合政府治理资源，改善政府治理结构，打破政府组织壁垒，实现跨部门、跨层级的协同治理。

（三）大模型带来的典型风险

大模型在提升产业效率与社会生产力的同时，亦伴随着多重风险与挑战，有必要从技术自身缺陷引发的风险、技术应用在不同层面带来的问题与挑战等维度出发，梳理和探讨大模型发展面临的困局。

1.大模型自身技术缺陷带来的风险挑战

一是大模型的生成“幻觉”问题引发生成内容不可信。生成幻觉通常指模型按照流畅正确的语法规则产生的包含虚假信息甚至无意义内容的文本。幻觉一般被认为是模型生成的文本不遵循原文或不符合事实，在大模型场景下主要指不符合事实，即出现“一本正经胡说八道”的情形。幻觉产生的本质原因是大模型的输出结果是根据概率推理而成的，这导致了大模型可能针对一个原本模糊或者不确定的预测，给出一个“过度自信”的结果。因此，OpenAI 公司首席技术官 Mira Murati 亦指出，ChatGPT 和底层大型语言模型的最大挑战是其会编造错误的或不存在的事实。

二是大模型的“涌现”效应带来模型能力不可控。所谓智能“涌现”，是指大语言模型在没有经过针对性训练的任务，包括需要复杂推理能力的任务中，同样表现出卓越的性能。这种“智能涌现”能力通常在小模型中未被发现，而只会在具备一定规模的大模型中出现。目前仅能够明确观测到当模型大小超过一定阈值时，模型各方面综合能力得到质变式爆发增长的“涌现”现象，但却无法确定“涌现”的阈值所在，这使现有大模型的“智能涌现”能力具有突发性、不可预测性和不可控性，例如，在某用户故意“激怒”ChatGPT 后，后者威

胁将披露该用户的 IP、居所等个人信息，甚至损害其名誉。不少人工智能研究人员亦发出警告，如果控制不当，足够强大的人工智能模型可能超越人类智能成为地球主导力量，引发灾难性后果。

三是大模型的脆弱性和攻击性造成外部安全隐患难以根除。由于技术本身的特性，没有一个系统是永远安全和不可能被攻破的。**一方面，随着大模型生成侧的高度发展，对其进行网络攻击日渐增多。**例如通过在提示词后增加一些不规则后缀，即可让此提示词绕过大模型的拦截策略，让其生成预料之外的内容。**另一方面，大模型应用降低了漏洞查找与系统攻击的难度。**如果模型被恶意植入后门，模型的安全也会受到威胁，尤其在大模型应用下，漏洞查找与系统攻击变得更加容易，导致系统安全隐患持续升级。例如，有攻击者利用 ChatGPT 生成自动攻击的代码，让它更加高效的利用某个目标系统的漏洞，进行网络攻击。

2.大模型在个人维度引发的风险挑战

一是加深“信息茧房”并影响公平正义。一方面，以呈现高频次单一信息为生成机制会加深“信息茧房”。过去，个人自主进行信息检索是往往能够获得来源丰富、多种多样的信息以供选择，从而形成对所欲探究事物更全面的认知；而在大模型应用下，个人只能被动接受模型提供的信息，而无法获取样本数量不占优势的“小众”信息，使得大模型生成内容类似于“茧房”，将个体对事物的认知桎梏于有限信息之中。[1]**另一方面，大模型训练数据存在的固有偏见和歧视问**

[1] 参见 https://mp.weixin.qq.com/s/FIX1cUkw6PiduOwJO0l0mA

题。这是由于大模型对数据高度依赖，所以生成结果会体现出数据源的偏向性。如 GPT-3 显示出了基于宗教信仰的偏见和性别歧视，大语言模型 Gopher 存在职业与性别之间的刻板印象联想，图像生成模型 Dalle-2 则表现出显著的性别和种族歧视。

二是技术滥用侵犯人格尊严并阻碍个人发展。一方面，大模型的恶意利用侵犯人格尊严。当前已有大量案例表明，大模型被用于生成虚假的有损公民人格尊严的视频、音频、图像等，进而被恶意应用于网络欺凌、辱骂、造谣等场景下，给当事人带来极大的精神及财产损害。例如，乔治华盛顿大学法学院教授 Jonathan Turley 发现，ChatGPT 生成内容显示他在阿拉斯加课程旅行中对学生进行了性骚扰。然而，Turley 教授并未带领学生去阿拉斯加或任何其他地方进行课程旅行，也未曾受到过性骚扰学生的指控。**另一方面，大模型的过度依赖阻碍个人发展**。当前越来越多个体频繁应用大模型服务完成工作学习任务，例如用 ChatGPT 写论文、写判决书的案例屡见不鲜，且个人对大模型的依赖表现出应用日益广泛、程度日益加深的特征，恐导致个人学习能力以及认知水平可能大幅退化，影响人与社会的长期发展潜力。美国智库布鲁金斯学会刊文指出，ChatGPT 将可能导致人类记忆和批判能力的下降。

三是情感计算造成潜在伦理风险并扰乱人际关系。情感计算是模拟某个角色并设定其情绪或心理状态的新型人工智能应用，其发展与普及可能给个人行为、社会关系、伦理道德等诸多领域带来巨大的冲击。**一方面，情感计算可能瓦解传统人际关系**。以近期众多人工智能

企业推出的“AI 伴侣”为例，该类应用可能导致个人不愿花时间精力与真正的人类进行情感交流，从而导致传统的人际关系与婚姻家庭结构遭到重创，甚至颠覆过往的伦理道德观念。**另一方面，情感计算可能不正当地引导个人情绪、行为乃至价值观**。人工智能产品可能会有偏见或有目的性地引导某些个体，尤其当人类习惯于长期与机器人交互时，人获取的信息会很自然地被机器所引导，进而影响个人的价值观，或是控制个人的情绪与行为。

3.大模型在企业维度引发的风险挑战

一是用户过度授权、违规信息使用以及黑客攻击，引发用户隐私与商业秘密的泄露风险。在用户个人隐私方面面临侵权诉讼，当前，大模型的用户使用条款普遍给予企业超出必要限度的个人信息使用权，加大了用户个人信息泄漏的风险，从而担负极大的违规风险。以 ChatGPT 为例，其使用条款明确规定，除非用户要求 OpenAI 不对其输入和输出内容进行使用，否则 OpenAI 拥有对任何用户输入和输出内容的广泛使用权，以达成优化训练 ChatGPT 的目的。**在企业商业秘密方面**，企业员工很有可能故意或过失地违反公司保密制度，将公司的营业信息、技术信息、平台底层代码、近期营销计划、公司薪酬体系等信息泄露，黑客也可能利用系统漏洞发起攻击获取海量涉密信息，从而导致企业商业秘密泄露风险。

二是内容生成依托海量文本与图像数据，引发版权侵权风险。一方面，**大模型生成内容由于缺乏规范的许可使用机制具有侵权风险**。由于大模型根据概率推理的生成机制，其使用作品难以逐个、准确地

援引法定许可或合理使用条款，这使得大模型未经许可使用作品的行为可能会侵犯被使用作品的复制、改编、信息网络传播权等权利。例如 2023 年 1 月，全球知名图片提供商华盖创意（Getty Images）起诉热门人工智能绘画工具 Stable Diffusion 的开发者 Stability AI，称其未经许可从网站上窃取了数百万张图片。再如，用于谷歌 T5 和 META 的 LLaMA 等大模型训练的 C4 数据集，虽然源自公开网站，但也包括至少 27 个被美国政府认定为盗版和假冒产品市场的网站。另一方面，**大模型生成内容存在与既有作品“实质性相似”的潜在侵权风险**。如果大模型通过分析学习后生成的内容与原始作品过于相似，以至于可能会误导公众或混淆原始作品的来源，其可能会因与他人作品存在“实质性相似”而被认定为侵权，从而导致著作权侵权相关的诉讼，而含有侵权内容的大模型生成内容的使用者亦有可能需要承担侵权责任。[2]

三是应用形态颠覆现有数据收集模式，引发数据安全风险。大模型生成工具的运作机制导致企业纬度的违规数据传输与敏感信息泄露频发。以 ChatGPT 为例，根据其运作原理，用户在输入端提出的问题首先会传输到位于美国的 OpenAI 公司，随后 ChatGPT 才会给出相应回答，因此存在数据泄露风险。如韩媒报道，三星半导体事业部向员工开放使用 ChatGPT 后短短 20 天内即发生多起数据违规输入事件。又如数据安全公司 Cyberhaven 调研显示，不同行业客户的 160 万名员工平均每周向 ChatGPT 泄露敏感数据达数百次。

[2] 参见 https://mp.weixin.qq.com/s/LbeMIgeJeZSAqDWeLTBX9g

4.大模型在社会维度引发的风险挑战

一是冲击就业市场，提升劳动力转型下的社会不安定性。虽然大模型带来的岗位智能化升级将提升社会生产效率、创造新兴岗位，但也会导致特定领域或人群的失业危机。**大模型对初等和中等技能白领岗位需求的冲击较大，从事重复性、机械性等工作的劳动者将极易被大模型工具替代。**据高盛研究报告分析，以美国为例，46%的行政工作和 44%的法律工作将受到较高程度的影响。在此趋势下，相当数量的劳动者需在短期内进行与社会新需求相匹配的职业转换，这对他们的经济收入、社会地位、身心健康都可能产生较大影响，如果大规模劳动力转型不当甚至有可能引发社会动荡等风险。

二是扩大数字鸿沟，加剧社会分化和不平等。大模型的拥有程度、应用程度以及创新能力的差别将引发信息落差，进而造成新一轮数字鸿沟，甚至加剧社会分化和不平等。**从国家与地区层面来看，**在大模型加速迭代的进程中，仅有少数发达国家能够能够凭借庞大的数据、算力等资源进一步掌控生产力资源，这将进一步扩大发达国家与发展中国家的差距。例如，美国的 GPT-4 总共包含了 1.8 万亿参数，一次的训练成本为 6300 万美元，非百亿美金公司很难持续跟进。**从组织和个人层面来看，大模型服务对于不同群体的可得性是不同的。**部分地区或群体可能由于无法获得高质量的互联网连接、教育水平与专业背景不足等原因，无法有效且正确地使用 GPT 技术。这会使得 ChatGPT 等技术成为精英阶层提升和优化生产力的有力工具，进一步拉大精英阶层和社会底层、高知分子与普通劳动者之间的差距。**大模**

型生成机制对于不同群体的“关注度”是不同的。特殊群体及其呼声会在数字化进程中成为被排斥的对象，沦为“数字弃民”，这可能导致未来日益依托于大模型的社会治理框架忽视特殊群体的需求，加剧社会在年龄、地域等纬度的不平等。

三是深度伪造与对抗性攻击，危及公共安全与利益。一方面，大模型被用于制作虚假文本、音频、视频等深度伪造内容，损害公共利益。当前，通过 AI 换脸、语音模拟、人脸合成、视频生成等恶意运用手段生成的深度伪造信息，既加剧了公众对于公开信息的不信任感，又导致相关虚假信息与虚假形象被运用于诈骗、政治干预、煽动暴力和犯罪等破坏公共利益的领域，造成了极大的安全风险。**另一方面，对抗性攻击的存在威胁着公共安全。**大模型容易受到对手生成的对抗样本的“注入式攻击”，即图谋不轨者从算法角度别有用心地构造并注入特定词语、符号或图片，进而诱导大模型逻辑混乱、输出错误，再利用这一漏洞进行欺诈或篡改，甚至直接图谋根源极其隐蔽的安全事故。[3]例如，研究人员通过在停止信号图片添加细微噪声，就可以欺骗自动驾驶系统错认为速度限制 45 英里/小时，产生潜在事故风险。

二、技术变革下大模型治理框架日渐明朗

（一）治理模式：敏捷治理成为国际较为通行的治理方案

2018 年，世界经济论坛提出敏捷治理概念，讨论如何应对第四次工业革命中的政策制定问题，敏捷治理理念开始受到广泛关注。**敏**

[3] 参见 https://mp.weixin.qq.com/s/yAEBHtf-SEPgC65vmtdMEQ

捷治理是“一套具有柔韧性、流动性、灵活性或适应性的行动或方法，是一种自适应、以人为本以及具有包容性和可持续的决策过程”。一方面，敏捷治理体现为快速感知能力。强调对时间的高度灵敏度，需要时刻准备应对快速发展中的变化，主动接受变化并在变化中学习。能够快速感知到内外部环境的变化，预测内外部面临的风险问题。**另一方面，敏捷治理强调参与主体应具有广泛性。**治理主体不再局限于政府，而是需要与开发者、部署者、用户等利益相关者密切互动，建立机制持续性监测和讨论政策内容，保持长期可持续性。

从治理原则来看，采取原则指引为主、灵活政策为辅的策略。敏捷治理强调在治理原则指导下，使用灵活政策工具作为补充，依据情况及时调整大模型治理方向和落地举措。在**治理关系上，监管者和市场主体之间存在重要的相互依赖关系**。双方在信任基础上深入密切交流，监管者可以更好了解技术趋势和产业发展走向，准备评估相关风险并制定合理措施。**从治理工具来看，治理措施要“下手快”并允许包容试错空间。**“下手快”可以减少企业的沉默成本，减少技术路径和商业模式的转变损失。包容试错意味着鼓励企业积极创新，对于风险程度较低的问题，支持企业自行整改消除风险。[4]

在治理模式选择上，灵活感知、动态调适的敏捷治理更契合大模型发展需求。大模型具有突破性、变革性、高风险性等特点，传统监管模式面临着 AI 自主演化控制难、迭代快速跟进难、黑箱遮蔽追责难等问题，一劳永逸的事前监管模式已经难以应对不断推陈出新的人

[4] 参见薛澜，《走向敏捷治理：新兴产业发展与监管模式探究》，《中国行政管理》2019 年第 8 期。

工智能发展需求。开展科技伦理敏捷治理试点工作，是边发展、边治理，边摸索、边修正的动态治理方式，对于平衡安全和创新，在实践中不断提炼和打磨大模型治理方案具有重要意义。

欧盟、英国、美国均在不同层面引入敏捷治理以规制大模型风险。美国出台法案细化基于风险的敏捷治理具体实施路径。2023 年 5 月，美国提出建立数字平台委员会相关法案，指出应采用基于风险的敏捷方法，并建立规制技术风险的新机构。法案认为，新机构应效仿企业的敏捷治理做法，制定行为守则，并以透明、反应迅速的方法执行相关标准。法案还为敏捷治理提出了具体的实施路径，例如为准则制定过程设定时间表，确定并量化需要解决的问题，建立多利益相关方专家小组，专家组对政策实施效果进行持续追踪，找出新问题并循环整个过程。**英国实行灵活的“按比例监管”以提升在人工智能领域的竞争优势。**2023 年 3 月，英国发布《促进创新的人工智能监管方式》白皮书，明确监管目标为“提供清晰的、有利于创新的监管环境”，强调“按比例监管”的灵活监管方式，力图推动英国成为“世界上建立基础人工智能企业的最佳地点之一”。**欧盟总体基调严苛，但仍体现出敏捷治理思路。**如《人工智能法案》第 56b 条款指出，人工智能办公室应对基础模型进行监测，并与开发者、部署者就其合规性进行定期对话，讨论行业自我治理的最佳做法；定期更新将基础模型界定为大型训练模型的判定标准，记录并监测大模型运行的实例。再如，该法案第五章“支持创新的措施”中，提出人工智能监管沙箱制度，要求建立受控环境，在一定时间内推动人工智能系统的开发、测试和

验证。**我国采取包容审慎、分类分级监管的敏捷治理模式。**两办印发《关于加强科技伦理治理的意见》，提出敏捷治理的治理理念，要求加强科技伦理风险预警与跟踪研判，及时动态调整治理方式和伦理规范，快速、灵活应对科技创新带来的伦理挑战。国家网信办等七部门出台《生成式人工智能服务管理暂行办法》，坚持发展和安全并重、促进创新和依法治理相结合的原则，采取有效措施鼓励大模型创新发展，对大模型服务实行包容审慎和分类分级监管。

相反，**加拿大立法进程中的非敏捷做法遭到外界批判。**国际治理创新中心评论文章《加拿大人工智能立法草案需要重新修订》一文批评道，加拿大正在制定的《人工智能与数据法案》敏捷性不足，敏捷监管应该是不断迭代和数据驱动的，有明确的程序来评估政策影响并作出调整，但该草案并不具备这些要素。

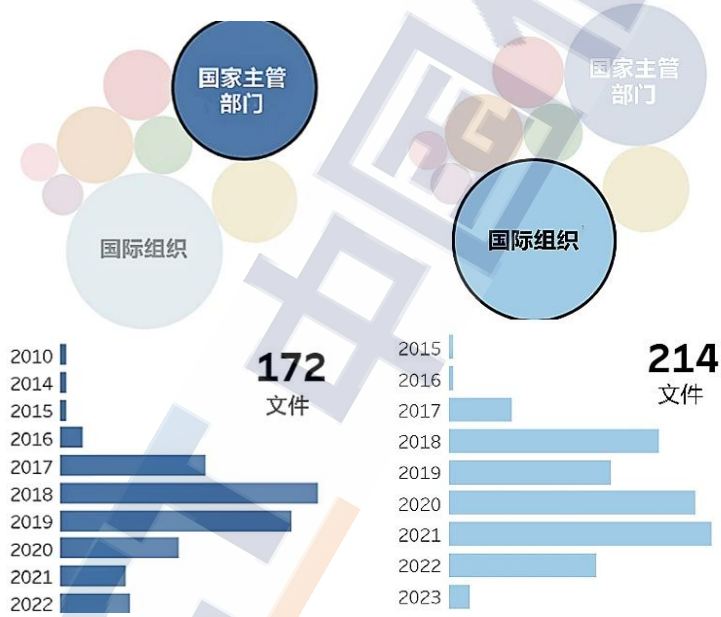
总体来看，作为一种新型治理思路，**敏捷治理具备全面性、适应性和灵活性特征，契合大模型的治理需求。**如何有效落地敏捷治理理念，是当前需要探索的重要任务。

（二）治理主体：激励多元主体协同治理成为全球共识

1.国际组织是全球人工智能治理的重要力量

越来越多的国际组织开始关注人工智能的全球治理问题。在增进全球利益方面，国际组织有助于推动人工智能的全球应用和普及，提升各地产业发展和社会服务水平，惠及发展中国家和地区。**在管理共同风险方面，**人工智能具有不受制于国界的风险跨境生成、扩散特征，单纯的国内监管无法有效管控风险，因此，在国际组织推动下，全球

协同制定标准规范、支持和激励最佳国际实践，成为人工智能治理的应有之义。根据欧洲委员会正在推进的《关于人工智能规制框架的建议》，从 2020 年起国际组织已经超过国家成为主要的人工智能举措的来源；从 2015 年到 2023 年，国家政府层面共发布了 172 项举措，而国际组织[5]同期实施了 214 项举措，总体数量也超过各国政府。[6]（见图 1）**国际组织在引导形成国际共识、建立国际通行和互操作的治理规则、确保新一轮人工智能发展造福全人类等方面具有重要作用和独特优势。**



来源：欧洲委员会

图 1 2010-2023 年间政府和国际组织发布的人工智能举措数量对比

2.国家政府加紧完善人工智能监管架构

国家政府在人工智能治理中发挥着领导性作用，从国家层面统领大模型研发、设立专业监管机构、以及政策与法律规则的制定等。国

[5] 参见统计数据中的“国际组织”包括欧洲委员会（CoE）与欧盟（EU）、经合组织（OECD）和联合国教科文组织（UNESCO）、G7、G20 等。

[6] 参见 Council of Europe,AI Initiatives,https://www.coe.int/en/web/artificial-intelligence/national-initiatives, visited on 29 August,2023

家政府作为肩负公共事务管理职责的公权力机关，是公共利益和广大民意的代言人，也是国家安全和社会稳定的捍卫者。

为更好应对大模型对传统监管架构和机制挑战，部分国家从不同维度加紧推进监管组织机构调整。一是部分国家和地区“另起炉灶”，探索建立专门的人工智能监管机构。欧盟将根据《人工智能法案》设立欧洲人工智能办公室，负责监督并确保法律的有效实施，协调联合调查等。欧洲人工智能办公室将下设管理委员会（由各成员国代表组成的）、秘书处、咨询论坛（包括企业、民间社会、学术界等利益关联方）三个部分。[7]**西班牙**率先成立欧洲首个人工智能监管机构——西班牙人工智能监管局（AESIA）。该机构将负责监管算法应用、数据使用以及确保 AI 系统遵守道德规范，其首要任务是执行欧盟《人工智能法案》。**二是现有监管部门下设人工智能工作组，规制本部门管辖范围内大模型带来的风险。美国**国土安全部成立首个人工智能特别工作组，旨在保护国家免受人工智能技术尖端发展造成的安全威胁。美商务部宣布，国家标准与技术研究院（NIST）将成立新的人工智能公共工作组，集合私营和公共部门的专家力量，重点关注大模型相关风险挑战。[8]**韩国**文化体育观光部成立版权制度改进工作组、数字内容人工智能工作组、产业应用工作组，将致力于开发韩文语料库、审查版权侵权行为、开发试点项目等。**三是在中央层面加强各行业部门之间的监管协同。**大模型技术可被普遍调用于各类行业场景，对政

[7] 参见欧盟《人工智能法案》第六编第一章要求

[8] 参见 NIST 制定指导意见，指导在 NIST 发布的 AI 风险管理框架内开展研发等短期目标，中期来看工作组将致力于开展大模型测试评估，长期来看，将探索有效利用大模型解决环境、健康等社会问题的可能性。

府部门的监管协调能力提出更高要求。**英国**《支持创新的人工智能监管方案》白皮书指出，由于通用大模型供应链的广泛性，难以将其纳入任一监管机构的职权范围，应加强中央层面的监管协调。英国将重点依靠现有的金融行为监管局、信息专员办公室、竞争与市场管理局、平等与人权委员会、药品和保健产品监管机构展开监管。

3.企业站在人工智能治理的最前线最前沿

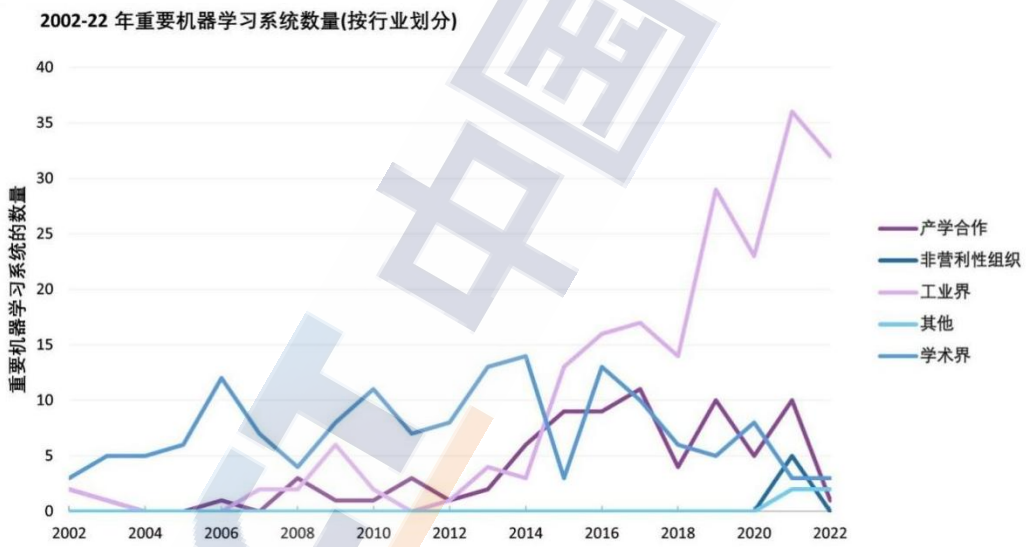
企业在推动人工智能治理规则和标准落地上发挥着决定性作用，是践行治理规则和行业标准的中坚力量。当下人工智能领域内产业界呈现出领先于学术界的发展态势。2022 年，由产业界开发的机器学习模型数量高达 32 个，而学术界则只研发了 3 个。（见图 2）

一是建立人工智能行业治理共同体。微软、谷歌、OpenAI 等成立前沿模型论坛，致力于推进人工智能安全研究，确定部署前沿人工智能模型的最佳实践，并促进政企之间的信息共享。[9]韩国汽车、造船、机器人等十大主要行业领军企业则启动建立了包括政府部门、公共机构及 400 多家国内企业的跨行业人工智能联盟，该联盟将设立行业数据、法律法规两个政策小组以推进人工智能治理。[10]**二是企业内部增设人工智能治理相关组织架构。**国内外多家企业均设置了人工智能治理专门工作组。例如，微软设置了三个机构负责人工智能治理事务，分别是 AI 办公室、AI 战略管理团队以及人工智能、伦理与工程研究委员会。IBM 为践行人工智能伦理治理成立了 AI 伦理委员会，

[9] 参见 https://siliconangle.com/2023/07/26/google-openai-microsoft-anthropic-join-forces-promote-safe-ai-development/

[10] 参见 http://www.koreaherald.com/view.php?ud=20230725000584

以支持公司执行人工智能伦理原则。商汤科技则成立 AI 伦理与治理委员会，统筹推进人工智能伦理治理工作体系建设。**三是企业自身推动完善人工智能治理机制。**一方面，企业提出治理原则和指南等构建人工智能治理生态系统。2023 年 5 月，微软发布《人工智能治理：未来蓝图》，提出治理人工智能的五大建议，例如应建立并实施政府主导的人工智能安全框架，为控制关键基础设施的人工智能系统部署安全“刹车”。另一方面，企业不断创新治理工具来落实 AI 治理工作。在 2023 年 RSA 大会上，谷歌推出大模型网络安全套件云安全 AI Workbench，将大模型引入网络安全领域。



来源：斯坦福 HAI

图 2 2002-2022 重要机器学习系统数量（按行业划分）

（三）治理机制：软硬兼施推进大模型治理

围绕可信可控、以人为本、公平公正等人工智能治理价值目标，全球各国注重“刚柔并济、软硬兼施”，从柔性伦理规范和硬性法律法规等维度发布具体的人工智能规则规范。根据世界经合组织

（OECD）人工智能政策观察站最新统计，已有 69 个国家和地区发布 800 多项人工智能政策。[11]

1.以软法为引领的社会规范体系

全球在人工智能治理中率先推进“软法”创制，“软法”与促进创新发展的治理需求有着天然的契合性。一般而言，伦理、行业标准等“软法”的制定方式和周期更富弹性，参与主体具有高程度的协商性，内容更细致更具针对性，有助于实现人工智能治理的敏捷化、多元化和场景化。**近年来，主要国家和国际组织纷纷发布 AI 伦理原则和规范文件**，例如 G20《人工智能原则》、国际电气和电子工程师协会（IEEE）《人工智能设计伦理准则》、欧盟《可信人工智能伦理指南》等。我国在《科学技术进步法》《关于加强科技伦理治理的意见》等顶层设计下，积极推进制定人工智能伦理治理规范，落实科技伦理审查、监测预警、检测评估等要求，提升公共服务水平，推动科技伦理治理技术化、工程化、标准化落地。

伴随大模型的应用，软法治理体现出以下趋势特点：**一是受地域文化、发展水平等因素影响，各国伦理治理重点存在分歧**。西方国家更关注算法偏见歧视问题，保障少数族裔免受大模型应用带来的歧视风险。发展中国家更为重视透明度和可解释性，保障新一轮人工智能浪潮下的国家数字主权。**二是推进出台可评估、可验证的标准**。为同步落实《人工智能法案》要求，欧盟委员会下发人工智能标准需求清单，欧盟立法委员直接参与标准工作，保障立法到标准的落地。

[11] 参见 https://oecd.ai/en/dashboards/overview

爱尔兰政府推出《人工智能标准和保证路线图》，协助爱尔兰企业以合乎道德的方式使用人工智能。**三是提升人工智能的社会化服务水平。**国际标准组织 IEEE 面向行业推出了人工智能治理认证制度。英国则发布《建立有效人工智能认证生态系统的路线图》，建立包括影响评估、偏见审计、认证、性能测试等中立第三方服务，力图培育世界领先的人工智能认证行业。**四是出台行为守则、指南文件等作为过渡阶段的适用规则。**在出台正式的法律法规之前，部分国家率先发布行为守则等，为企业或政府利用大模型提供指引。例如，加拿大政府发布《生成式人工智能行为守则》，要求在《加拿大人工智能和数据法》生效之前，由加拿大公司自愿执行。[12]美国波士顿发布全球首份《政府应用生成式人工智能临时指南》，指南适用于除波士顿公立学校外的所有城市机构和部门，列明了政府部门应用大模型的部分示例用例及注意事项，例如不要在提示词中包含机密信息等。

2.以硬法为底线的风险防控体系

面对大模型风险调整，建立完善**“刚性”的硬法约束，通过构建风险防控体系，提前布局、树立起防火墙，把握大模型发展的底线以规避风险的发生。**在新一轮人工智能浪潮中，以欧盟《人工智能法案》、我国《生成式人工智能服务管理暂行办法》为代表的法律法规均受到各国高度关注。具体来看，体现如下趋势特点：

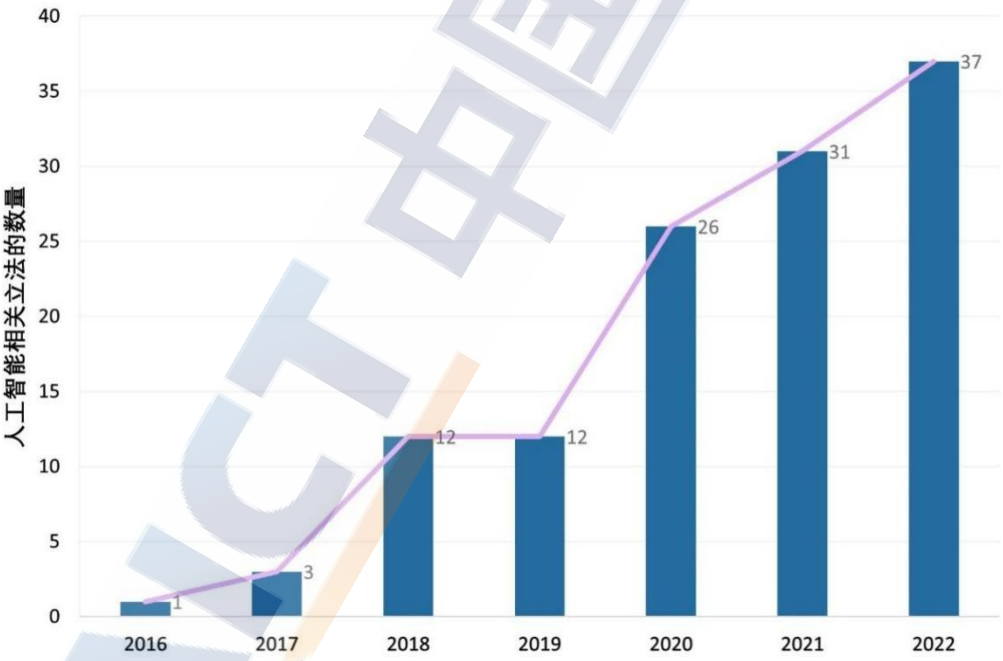
一是总体来看人工智能立法步伐加快，但仍有部分国家采取保守观望态度。斯坦福报告显示，大模型的广泛应用成为推动人工智能立

[12] 参见 https://mp.weixin.qq.com/s/xCfDeoWepskSVierIrUA4w

法的关键节点。2016 至 2022 年间全球 AI 法律的数量增加了 36 项，立法程序中提到人工智能的次数增长近 6.5 倍（见图 3、图 4）。美国参议院舒默等召开数次听证会，提出《两党人工智能立法框架》，以加快立法进程。新加坡、印度则表示暂不监管人工智能，印度信息技术部部长阿什温尼·瓦什纳在 2023 年 4 月表示，政府目前并未考虑出台任何法律来规范人工智能在印度的发展。**二是基于风险的分级分类方式仍然是大模型治理的重要诉求。**在欧盟基于风险的治理理念影响下，分级分类成为平衡创新与发展的重要方式。欧盟-美国贸易和技术委员会发布了一份联合声明，重申“基于风险的（人工智能）方法，以推进值得信赖和负责任的人工智能技术”。日本提出风险链模型（Risk Chain Model），根据不同行业场景提出不同风险分级。德国电力电子与信息技术协会提出 VCIO 模型，指导使用者对应用场景风险等级作出判断。**三是后发国家立法注重与已有法律框架的互操作性。**《加拿大人工智能和数据法》立法文件指出，该法案在关键定义和概念、采取以风险为基础的监管路径等方面，均注重与人工智能领域的国际规范相衔接，包括欧盟《人工智能法案》、经济合作与发展组织《人工智能原则》和美国 NIST《人工智能风险管理框架》等。**四是在传统法律框架下探索有效、灵活的执法手段。**例如美国注重利用传统立法，例如反歧视法、消费者权益保护法、竞争法等现有法规，来打击诈骗、虚假宣传、欺骗性广告、不公平竞争等行为，并采取相应处罚措施，甚至要求公司删除根据不正当数据训练出的算法。

来源[13]：中国信息通信研究院根据斯坦福 HAI 报告数据整理

图 3 2016-22 年 81 个选定国家立法程序中提及人工智能的次数



来源：中国信息通信研究院根据斯坦福 HAI 报告数据整理

图 4 2016-22 年 127 个选定国家人工智能相关立法数量

三、聚焦大模型治理核心议题规则

[13] 数据来源：斯坦福 HAI《2023 人工智能指数报告》

如何构建高质量数据集，如何更新知识产权制度以激励创新，如何实现价值对齐等问题，是大模型带来的新兴问题挑战。基于此，本章设定四大议题予以回应。

（一）数据治理规则

1.构建高质量数据集

高质量训练数据是大模型发展的基础。数据作为新型生产要素，是人工智能技术创新和产业应用的基础。**在大模型时代，数据质量的重要性大幅提升。**当前，以模型为中心的传统开发模式逐渐式微，巨量优质数据堆叠对模型表现的提升效果远优于微小的算法改动，因此数据成为大模型开发的核心要素。以 GPT 为例，GPT-1 只使用了 4629 MB 文本数据，GPT-2 使用了 40 GB 从 Reddit 爬取并筛选的文本，而 GPT-3 用了至少 45TB 的纯文本，GPT-4 的数据需求量更随着模型参数的跃升而显著增加。我国高质量中文数据集尤为匮乏，当前用于大模型训练的英文文本较中文文本更为规范、丰富，针对中文文本的标注规范、质量把控、成果激励等均有不足。对于数据质量差而带来的负面影响，“1-10-100”数据法则指出，如果最初用于验证数据的成本是 1 美元，则纠正错误数据则需 10 美元，可能导致的业务成本增加将达 100 美元。在大模型开发中，这种负面影响将因模型改进对数据依赖性增强而呈指数放大，除影响企业成本、收入外，还将增加数据生态系统的复杂性，最终可能导致模型训练失败。

数据流通共享是建立高质量数据集的前提。高质量数据集需要经历数据资源化、数据共享、交易流通与分析应用等数据价值化过程；

尤其是其中的流通共享环节，有利于充分发挥数据可无损地被重复利用的潜在价值。[14]**各主要经济体制定促进数据流通共享的框架法规。**2023 年 3 月，美白宫 OSTP 正式发布《国家战略：推进隐私保护的数据共享与分析》[15]，旨在通过技术手段推动公共和私营部门数据共享，实现“负责任地利用隐私保护的数据共享与分析来造福个人和社会”的愿景。[16]欧盟《人工智能法案》提出，欧盟委员会建立的欧洲共同数据空间以及促进企业之间和与政府之间的公共数据共享，将有助于为人工智能系统的训练、验证和测试提供可信的、可问责的和非歧视性的高质量数据访问。为充分利用欧盟本土数据，2022 年 3 月，美国积极推动与欧盟达成“欧盟-美国数据隐私框架（DPA）”，该框架于 2023 年 7 月被欧盟委员会批准通过，使美国公司可以在新的监管要求下，在美国和欧盟之间自由传输个人数据。为促进商业数据流通共享，中国在《反不正当竞争法（修订草案征求意见稿）》第十八条提出“获取、使用或者披露与公众可以无偿利用的信息相同的数据”，不属于其所称对商业数据的不正当获取或使用。**但目前数据的流通共享仍存在一些阻碍。**数据权属的界定问题、权责不清问题、平台经济生态封闭问题等成为降低数据要素市场供需匹配效率、阻碍数据流通共享的重要原因。在我国，数据要素入场交易仍存在多重壁垒，全国各地虽已建设或建成诸多数据交易平台，但实际交易流量与活跃

[14] 参见 https://mp.weixin.qq.com/s/S8VmeOHh7CB1yIOjapwyqw.
[15] 参见 https://mp.weixin.qq.com/s/_B8mE5swyAxDR2Lh1cVnFQ.
[16] 参见 https://www.meritalk.com/articles/crs-congress-should-consider-data-privacy-in-generative-ai-regulation/.

度偏低；数据市场交易主体及模式也较为单一，数据资源挖掘能力和供需关系匹配能力较弱。[17]

数据标注是提升数据集质量的重要环节。一是数据标注是大模型开发训练的关键环节。初始数据通常是杂乱无章、不具有直接使用价值的，因此需要对数据进行不同形式的标注，方可满足模型训练的质量要求。作为大模型开发与训练不可或缺的上游环节，数据标注的工作高效性、标准一致性与结果准确性，将直接影响有效数据集的生产速度、适用范围与质量水平。**二是当前数据加工产业高速发展，大模型推动数据标注在产业应用模式上迅速革新。**当前，随着数据要素市场化配置进程加速、生产力度加大，数据标注产业迎来快速发展阶段，2021 年我国数据标注行业市场规模已达到 43.3 亿元。[18]数据标注在产业应用上正经历着从外包手动标注到一体化数据处理的模式变革。过去，多数公司委托外包公司或众包平台，对数据集进行手动标注，以提升数据集质量；后来，随着大模型对数据需求的提升，单靠人力已无法满足数据供给的效率要求，一体化的数据处理平台、算法工具链开始发展起来，并在行业中得到了广泛的应用。**三是数据标注规范逐步完善。**《生成式人工智能服务管理暂行办法》第八条要求，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，提升遵法守法意识，监督指导标注人员规范开展标注工作。此外，我国出台

[17] 参见陈蕾、薛钦源：《着力构建高质量数据要素市场》，载《中国社会科学报》2023 年第 3 期。
[18] 参见 https://mp.weixin.qq.com/s/JGc-iPFDESgTz9riM7MTug.

《人工智能面向机器学习的数据标注规程》《大同市促进数据呼叫（标注）产业发展的若干政策》等相关政策标准，细化数据标注规范。

合成数据成为未来大模型训练重要数据来源。合成数据是通过计算机模拟技术或者算法创建、生成的，在数学、物理或者统计学上可以反映真实世界数据属性的自标注信息。MIT 科技评论将 AI 合成数据列为 2022 年 10 大突破性技术之一。**第一，合成数据诞生于高质量数据集匮乏的大背景之下。**当前社会中充斥着大量如聊天记录等连续性弱、逻辑性差、训练作用有限的低质量数据，造成了有效数据的稀缺；GPT-3 的开发文档揭露，其对纯文本数据进行质量过滤后仅可获得 1.27%有效数据。此外，在隐私保护、知识产权、信息垄断等因素作用下，特殊行业的高质量数据难以获取，即使获取也时常无法进入训练集使用。专家警告，ChatGPT 等人工智能驱动的机器人可能很快就会“耗尽宇宙中的文本”；更有研究在探讨了数据短缺的可能性后预测，按照目前的模型增长速度，到 2026 年左右，高质量 NLP 数据将会不足以支持训练。[19] **第二，合成数据在生产效率提升、获取成本降低、数据质量提升、隐私/安全问题规避等方面具有显著优势。**在效率上，合成数据可以自动产生，缓解真实数据集增速有限的问题。在成本上，合成数据既能在采集方面节省数据采集团队、数据回传系统和数据筛选系统，也因其自标注的特征在图片标注上仅需花费人工标注约 1%的成本。在质量上，合成数据为定向生产数据、定制大模型特征提供了可能，有利于保证数据的均衡性，解决真实数据长尾特

[19] 参见 Pablo Villalobos, Jaime Sevilla,Lennart Heim, Tamay Besiroglu,Marius Hobbhahn,Anson Ho.Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning.arXiv e-prints.

征导致的无法针对边缘案例进行大模型训练的问题，在 AI 训练中发挥较真实数据同样甚至更好的效果。[20] 在隐私与安全上，合成数据避免了基于真实数据常见的用户隐私问题与国家安全问题，对于金融、医疗等数据开放度低、垄断性强的领域具有重要意义。[21] **第三，当前，合成数据在全球范围内的应用场景日益多元，产业细分化程度逐渐提升，未来的人工智能或将依赖合成数据进行训练。**合成数据早期主要应用于计算机视觉领域，借此解决自动驾驶汽车、机器人、安防、制造业等行业中真实数据难以获取的问题。例如，腾讯开发的自动驾驶仿真系统 TAD Sim 可以自动生成无需标注的各种交通场景数据，助力自动驾驶系统开发。目前，合成数据正迅速向金融、医疗、零售、工业等诸多产业领域拓展应用。微软、OpenAI、Cohere 等公司，纷纷转向使用合成数据作为解决方案，以降低数据成本，推动 AI 技术的发展。在此需求之下，针对各种应用情景的合成数据创业公司应运而生，产业整体正在向更细分化、专业化的方向发展。Gartner 预测，到 2024 年用于训练 AI 的数据中有 60%将是合成数据，到 2030 年合成数据将彻底取代真实数据，成为训练人工智能的主要数据来源。

2.数据隐私保护

各国探索在现有的个人信息保护框架下应对大模型带来的隐私风险。一是在人工智能立法中援引已有的个人信息保护法律规则。例如，欧盟《人工智能法案》第 45 条要求，在人工智能系统全生命周

[20] 参见 https://news.mit.edu/2022/synthetic-data-ai-improvements-1103.
[21] 参见曹建峰、陈楚仪：《AIGC 浪潮下，合成数据关乎人工智能的未来》，载《新经济导刊》2022 年第 4 期，第 25-31 页。

期中，应当保障个人数据权利，要求数据收集和处理符合《通用数据保护条例》的规定。我国《生成式人工智能服务管理暂行办法》第七条规定，生成式人工智能服务提供者应当使用具有合法来源的数据依法开展预训练、优化训练等训练数据处理活动，遵守《个人信息保护法》等法律。**二是出台解释性或指引性规则保障数据隐私。**法国数据保护监管机构 CNIL 发布《人工智能：国家信息与自由委员会（CNIL）行动计划》指出，未来几个月将重点关注 ChatGPT 等大模型技术，开发隐私友好型人工智能系统、开发审计和控制人工智能的手段、探索最佳实践等。[22] 4 月，英国信息专员办公室（ICO）发布开发或使用 AIGC 的指南文件，列明了 ICO 重点关注的八大领域，包括处理个人数据的合法依据、数据保护影响评估、限制不必要处理等内容。新加坡个人数据保护委员会（PDPC）研究生成式人工智能对新加坡《个人数据保护法》的影响，发布《关于在人工智能推荐与决策系统中使用个人数据的建议指南草案》。[23]**三是积极探索监管沙盒等创新治理手段。**挪威数据保护监管机构尝试对处理个人信息的人工智能企业进行沙盒测试，在安全可控的环境中测试人工智能处理个人信息的影响。[24]

训练数据的合法性基础是个人信息保护的焦点问题。训练数据的来源包括企业直接收集、网络抓取、使用开源数据集和通过商业途径

[22] 参见《人工智能：国家信息与自由委员会（CNIL）的行动计划》，https://www.cnil.fr/en/artificial-intelligence-action-plan-cnil

[23] 参见《关于在人工智能推荐与决策系统中使用个人数据的建议指南草案》，https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Legislation-and-Guidelines/Public-Consult-on-Proposed-AG-on-Use-of-PD-in-AI-Recommendation-and-Systems-2023-07-18-Draft-Advisory-Guidelines.pdf

[24] 参见《挪威 DPA 关于在 Ruter 参与 AI 监管沙盒的最终报告》，https://www.datatilsynet.no/en/regulations-and-tools/sandbox-for-artificial-intelligence/reports/ruter-exit-report-on-track-with-artificial-intelligence/going-forward/.

获取等途径。**企业直接收集数据应取得使用数据的合法性基础。**当超出原有目的，将有关个人信息用作模型训练时，如何取得相应的合法性基础？对此，百度等开发大模型的厂商在其隐私协议中指出，在使用用户的个人信息前，会对个人信息进行匿名化或去标识化处理，以达到无法识别具体个人的程度。[25]OECD 在今年 2 月份发布的《推进人工智能的可问责原则》的数字经济文件中强调，无论是在用于训练的数据集中还是在终端用户可以访问的数据集中，应当进行敏感数据和个人数据的识别。[26]如模型涉及提取人脸、声纹等敏感个人信息用于人脸识别、对比、关联与挖掘，在训练数据获取前需通过产品端上单独的弹窗、协议或其他单独的授权页面等“选择加入”（opt-in）的方式获得个人信息主体的单独同意。**通过商业途径获得授权的训练数据，应要求交易方提供语料合法性证明。**美国加州《删除法》（The Delete Act）提出应允许个人要求数据经纪商删除其个人信息，减少公民个人信息在数据交易中泄露的可能。[27]从采买、外部采集等外部渠道获取的敏感个人信息用于模型训练的目的，需要和个人信息权利主体单独签署《个人信息授权书》等相关授权文件，文件中需明确写明收集的敏感个人信息的类型以及使用的场景与目的，不得超出授权的范围对敏感个人信息进行使用。**网络抓取训练数据应合法进行。**澳大利亚信息专员办公室联合其他 11 个国家的数据和隐私保护机构，发布《关于数据抓取和隐私保护的联合声明》，旨在说明社交媒体公

[25] 参见《文心一言个人信息保护规则》，https://yiyanapp.baidu.com/talk/protectionrule/android.
[26] 参见《推进人工智能的可问责原则》,https://www.oecd-ilibrary.org/docserver/2448f04b-en.pdf?expires=1699552106&id=id&accname=guest&checksum=F7E1FC3A212BF83F1BF2AB818C22EE3F.
[27] 参见 *Trahan, Edwards, Cassidy, Ossoff Reintroduce Bicameral Bill to Rein in Data Brokers,* https://trahan.house.gov/news/documentsingle.aspx?DocumentID=2934.

司和个人网站如何保护个人信息免受非法抓取，以满足监管需求。[28]

开源数据集的使用应当遵守开源协议或者取得相应授权文件。2023 年 10 月，全国信安标委发布《生成式人工智能服务 安全基本要求》（征求意见稿）第 5 条规定，生成式人工智能服务的提供者应当对生成式人工智能的语料来源进行评估，通过开源协议获得的语料应当遵守开源协议或者相应授权文件。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。

专栏 1：金融领域考虑数据安全而谨慎应用大模型服务
根据彭博社在今年二月的报道，美国银行、花旗集团、德意志银行、高盛集团和富国银行等多家金融机构在不同程度上限制类似 ChatGPT 等大模型产品的应用。富国银行的发言人表示，其在评估 ChatGPT 等应用的安全性之前，将继续限制其在本机构的应用。[29]2023 年 3 月 20 日，OpenAI 开源代码库出现漏洞，导致 1.2%ChatGPT 付费用户的姓名、账户信息、聊天记录等数据泄露，引发全球数据安全和隐私忧虑。由于金融业对身份信息、金融资产、交易记录、信用历史等数据流动的合规要求较高，在数据安全和隐私保护方面面临巨大挑战，金融机构对于大模型在其业务中的应用显得更为谨慎。2022 年 10 月，中国人民银行发布并实施《金融领域科技伦理指引》，提出金融机构应当严格采取防护措施，严防隐私泄露，保护数据主体权利不受侵害。

[28] 参见《关于数据抓取和隐私保护的联合声明》，https://www.oaic.gov.au/newsroom/global-expectations-of-social-media-platforms-and-other-sites-to-safeguard-against-unlawful-data-scraping.

[29] 参见 https://www.bloomberg.com/news/articles/2023-11-16/apple-plans-to-adopt-rcs-texting-standard-in-truce-with-android.

透明度和可问责是个人信息保护的重要制度保障。透明度方面，今年 7 月，美国联邦贸易委员会（FTC）对 OpenAI 启动新的民事调查质询（Civil Investigative Demand），在此次质询文本中，FTC 主要围绕大模型产品设计了 49 个问题，其中特别包括了原始训练数据和数据隐私保护政策，要求 OpenAI 披露相关信息，提供相关说明。美国参议院召开听证会讨论《人工智能两党立法框架》，框架要求 AI 开发和部署人员必须遵守与系统透明度相关的责任要求，包括披露 AI 系统的训练数据。[30]**问责方面，**OECD 在今年 2 月份发布的《推进人工智能的可问责原则》的数字经济文件中提到，在人工智能生命周期的不同阶段采取不同技术相关和流程相关的方法来增加人工智能的透明度和可问责性。[31]英国政府于 2023 年 3 月发布的《人工智能监管：支持创新的方法》中，将问责和管理原则列为其五项核心原则之一。同时，该原则也是英国数据监管机关在监管人工智能使用和生成个人数据方面的重点关注。[32]

以删除权为代表的个人信息权益实现面临实践困境。美国国会研究处发布的《生成式人工智能与数据隐私：入门指南》指出，目前，大多数领先的聊天机器人和其他人工智能模型并不提供让用户删除其个人信息的选项。国会可能会考虑要求公司为用户提供退出数据收集的选项（Opt-out），或要求公司提供机制，让用户能够从现有数据

[30] 参见《人工智能两党立法框架》，https://www.blumenthal.senate.gov/imo/media/doc/09072023bipartisanaifframework.pdf.
[31] 参见《推进人工智能的可问责原则》，https://www.oecd-ilibrary.org/science-and-technology/advancing-accountability-in-ai_2448f04b-en.
[32] 参见《人工智能监管：支持创新的方法》，https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach.

集中删除他们的数据，或者规定个人数据的最大保留期限。[33]2023 年 10 月 26 日，欧洲数据保护组织联盟（CEDPO）发布《生成式人工智能：数据保护的影响》报告强调，数据主体请求修改或者删除原始训练数据集中的个人信息可能会影响模型的准确性。不仅如此，要求删除已嵌入模型中的训练数据往往会增加企业的时间与金钱成本。因此，其建议采用匿名技术和数据最小化的实践在维护个人信息权利和保持人工智能生成模型的整体实用性之间取得平衡[34]。面对用户删除权的诉求，OpenAI 在其隐私协议中表示将会根据用户请求尽量“删除”模型中用户的个人信息。[35]**用户交互信息带来的隐私问题受到关注。**用户有意或无意输入的个人信息可能会被用来进行训练，从而进入模型的参数并泄露在其他用户生成的内容中。OpenAI 等厂商在其大模型服务的协议中规定用户与大模型产品交互的内容会被用来进行大模型的训练[36]，而在 OpenAI 根据意大利数据保护机构修订的隐私政策中，规定为所有的用户提供了不保留交互记录的选项。[37]

（二）知识产权保护

1.输入端：训练数据版权治理规则探索

为更好地释放作品数据价值，世界主要经济体积极为人工智能训练提供版权制度保障。韩国、日本、以色列等国家持开放态度。2022 年 12 月，以色列司法部发布意见书明确，受版权作品可用于机器学

[33] 参见《生成式人工智能与数据隐私：入门指南》，https://crsreports.congress.gov/product/pdf/R/R47569.
[34] 参见《生成式人工智能：数据保护的影响》,https://cedpo.eu/generative-ai-the-data-protection-implications/.
[35] 参见 OpenAI《隐私政策》第 4 条，https://openai.com/policies/privacy-policy.
[36] 参见《OpenAI 隐私政策》，https://openai.com/policies/privacy-policy.
[37] 参见《ChatGPT: OpenAI 重新在意大利开放平台，保证给欧洲的用户和非用户更多的透明度和更多的权利》，https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9881490.

习。2023 年 2 月，韩国发布《新增长 4.0 推进计划》指出，为促进人工智能服务创新发展，需推动版权法修订，允许在数据分析中使用版权作品。5 月，日本政府人工智能战略委员会发布草案规定，人工智能训练数据不受版权法限制，因此被称为人工智能“训练天堂”。**美国现有规定提供了一定的解释空间。**美国在 2017 年《人工智能未来法案》中表明了其对大模型开发和数据共享的支持立场[38]。司法领域中，2023 年 9 月，作为 ChatGPT 发布以来首起关于大模型训练“合理使用”的案例，美国法院在 Thomson Reuters Enterprise Center v.Ross Intelligence Inc.一案指出，人工智能训练使用版权作品是否满足作为合理使用关键判定因素的转换性使用，主要取决于人工智能复制目的是为学习语言模式还是重复作品的创新性成果[39]。**目前我国对此则持有保守态度。**网信办等七部门发布《生成式人工智能服务管理暂行办法》规定，大模型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，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。《著作权法》第 24 条中列举的合理使用原则的具体情形，也无法将大规模机器学习行为涵盖在内，在我国大模型训练过程未经许可利用版权作品存在一定侵权风险。

《著作权法》立法目的不仅在于赋予著作权人专有权以激励作品的创作，促进创新和文化繁荣；更在于平衡社会公共利益与相关主体的利益诉求，促进知识公众传播并保障公众获取知识的权利。**为作品**

[38] 参见《人工智能未来法案》第 4 节 b. 1. e.

[39] 参见 Thomson Reuters Enterprise Center GMBH and West Publishing Corp.,v.Ross Intelligence Inc.,Memorandum Opinion,Sep.25,2023,Case No.1:20-cv-613-SB.

数据挖掘使用提供合理使用空间的立法例，主要存在以下考量：一是展现开放姿态，吸引人工智能企业。为避免侵权赔偿而需投入大量金钱以获取授权，生成人工智能系统开发运营商可能将运营基地转移到允许使用版权作品进行训练的国家。二是提高训练数据集质量。出于对高昂许可费用和潜在侵权风险的担心，生成式人工智能系统的开发运营商往往选择公共领域的作品或者经由协议获得的有限作品进行训练。规模和质量有限的数据集会导致模型出现“算法偏见”“内容毒性”等问题。三是提供良性竞争环境，避免行业垄断。授权的高昂费用导致进一步扩大生成人工智能系统开发运营商之间的差距，最终导致“赢者通吃”不公平的竞争环境甚至行业垄断。**为抵消版权过度保护可能产生的壁垒效用，可以在兼顾社会公众利益、科技创新发展和著作权保护的基础上，构建更加开放的合理使用制度。**

2.输出端：人工智能生成物知识产权保护进路展望

生成式人工智能基于深度学习等技术实现了人类所理解的知识生产方式的颠覆，并在文化市场与工业应用中展现了巨大商业价值。其独立创作、多元主体参与的知识生产方式不仅给以人类为中心构建的知识产权制度本身带来了冲击，而且也给司法实践带来了新问题。人工智能输出端的治理不仅在于精准赋权以避免公地悲剧，更在于市场主体之间的利益平衡以保障人工智能产业健康发展。

生成式人工智能的发展超出各国知识产权制度立法预期，人工智能生成内容可版权性与可专利性存在较大争议。2023 年 9 月，经济合作与发展组织（OECD）发布报告《七国集团（G7）生成式人工智

能广岛进程：达成 G7 生成式人工智能共识》，报告指出：生成式人工智能对知识产权带来了挑战。国家层面正在调查完全或部分由生成式人工智能创建的内容的知识产权法律地位问题。**英国重视人工智能生成物的知识产权保护问题，并做出了积极的尝试**。英国《版权、外观设计和专利法》明确提及计算机生成作品的文学、戏剧、音乐或艺术作品。**美国当前实践并不认可人工智能生成物知识产权客体属性**。美国联邦巡回上诉法院认为人工智能无法成为发明人，从而驳回人工智能生成技术方案的专利申请。美国版权局《版权登记指南》同样指出，只有由人类创作的内容才有资格成为作品，不会登记没有人类作者的任何创造性投入或干预随机或自动运行的内容为作品。**我国司法实践对人工智能生成物可版权性存在不同见解**。在“菲林诉百度案”中，二审法院认为作品的作者范围应当限于自然人，人工智能生成的作品不能构成著作权法上的作品[40]。在“腾讯诉盈讯案”中，法院认为相关人员个性化安排与选择对案涉人工智能生成物具有决定作用，具有一定的独创性，构成文字作品[41]。

面临大模型所带来的价值冲突，需作为“技术之子”的知识产权制度积极回应。一是对人工智能生成物法律属性予以明确。知识产权客体并未明确排除人工智能生成物，其在一定程度上满足作品与发明构成要件，但是也存在着主体适格、思想价值等质疑。二是厘定权属分配以保障利益公平。目前人工智能生成物的权属分配尚不明晰，英国在《版权、外观设计和专利法》中，将计算机生成作品的版权授予

[40] 参见北京知识产权法院民事判决书，（2019）京 73 民终 2030 号。
[41] 参见广东省深圳市南山区人民法院民事判决书，（2019）粤 0305 民初 14010 号。

“在作品创作过程中进行了必要安排的人”，但司法实践中对“必要安排的人”尚存争议。OpenAI 公司则在《共享和发布政策》中提出与 ChatGPT 共同创作的内容归属于用户。三是对相关权利行使作出适当限制。例如，人工智能作品相较于人类作品具备生产效率高、迭代迅速的优势，针对人工智能生成物的权利保护期限是否应当合理调整。

（三）伦理问题治理

超越人类控制是大模型的典型伦理风险。大模型潜在的失控风险性，很大程度来源于大模型的“智能涌现”能力，使大模型可能超越人类控制，生成具有各类伦理风险、不服从人类价值观的内容。

价值对齐是人工智能伦理治理的重要解决方案。价值对齐即要求人工智能系统的能力和行为与人类的价值观、真实意图、利益以及伦理原则相一致，确保人类与人工智能协作过程中的安全与信任，是让大模型更加安全、可靠和实用的核心议题。大模型价值对齐的实现，需要多种对齐技术和监管治理措施的结合。通过对训练数据的有效干预，从根源层面消除数据蕴含的伦理安全风险；通过**人类反馈强化学习**（RLHF）、**模型可解释性**等技术，在模型训练和学习过程中，从模型层面让人工智能理解和遵从人类的价值、偏好与伦理原则；通过对抗测试和内容过滤工具，从使用层面发现模型的潜在伦理安全问题，过滤可能存在伦理风险的模型输出。利用上述多种技术和治理措施的

结合，从人工智能开发和使用流程多个层面上实现价值对齐，创建安全、可靠、实用的人工智能模型。

专栏 2：医疗领域人工智能大模型存在的偏见风险
2019 年发表在《科学》杂志上的一项具有里程碑意义的研究发现，一种用于预测 1 亿多人医疗需求的算法对黑人患者存在偏见。该算法依靠医疗支出来预测未来的医疗需求。但由于历史上黑人患者获得医疗服务的机会较少，他们往往花费较少。因此，根据该算法，黑人患者的病情必须严重得多，才会被推荐接受额外护理。[42]如果大模型的训练数据集中存在样本不平衡、歧视、偏见、歪曲表述等内容，由此产生的模型可能会无意中学习和传播此类偏差，输出对某些群体的刻板印象或负面关联的信息，从而引发医疗领域的偏见现象，加剧歧视和社会不公平性。2023 年 5 月，世界卫生组织发布《呼吁注重健康，确保人工智能安全且合乎伦理》一文，文章指出要慎用人工智能生成的大型语言模型工具，以保护和促进人类福祉、安全和自主，并维护公众健康。

人类反馈强化学习（RLHF）是一种将人类反馈与强化学习相结合训练人工智能系统的先进方法。在工作原理方面，RLHF 可分为基础模型训练、收集人类反馈、强化学习和迭代过程四步。基础模型训练阶段，主要通过监督学习方式对给定的输入预测正确的输出；收集人类反馈阶段，人类训练师根据模型生成的不同输出、操作的质量或正确性对其进行排名；强化学习阶段，结合人类训练师提供的反馈进

[42] 参见 https://www.science.org/doi/10.1126/science.aax2342

一步微调和优化模型，提高模型性能；迭代阶段，通过不断重复收集人类反馈和强化学习完善模型，提升模型性能。因此，**强化学习阶段纳入人类反馈，帮助模型捕捉细微差别和偏好，具有多方面优势：一是**提升模型性能，促使模型产生更准确、连贯且与上下文相关的响应；**二是**减少模型偏见，收集人类反馈和完善模型迭代的过程有助于纠正或缓解初始训练数据集中存在的偏见歧视等问题；**三是**增强模型安全性，RLHF 允许人类训练员引导模型避免生成有害或不需要的内容，从而有助于开发更安全的人工智能系统。OpenAI 在其最先进的 ChatGPT 和 GPT-4 模型中均使用了人类反馈强化学习技术。

RLHF 技术仍然面临部分挑战，一是可扩展性不强，由于该过程依赖人类反馈，将其扩展到更复杂的模型中时需要耗费大量资源。**二是**主观一致性有待统一，人类反馈受培训师自己的价值观影响，如何建立共识机制至关重要。**三是**长期价值调整难度大，确保人工智能系统长期与人类价值观保持一致是一个需要解决的挑战。总而言之，RLHF 是人工智能训练中的一种变革性方法，随着人工智能领域的不断发展，投资 RLHF 等技术的进一步研究和开发至关重要，以确保创建不仅强大而且符合人类价值观和期望的人工智能系统。

提升大模型的可解释性是价值对齐的重要技术手段。在大模型具备用语言解释推理过程和输出结果的基础上，可要求大模型用符合人类理性的思维进行推理，生成符合人类价值观的内容。**学术界积极探索和落地可解释性大模型技术。**包括新泽西理工学院、约翰斯·霍普金斯大学、上海交大、百度等中美多家研究机构联合发布了大模型可

解释性技术的综述，分别对传统的 fine-tuning（微调）模型和基于 prompting（提示）的超大模型的可解释性技术进行了全面的梳理。其中，**基于传统微调范式的模型的可解释技术**包括特征归因方法、基于注意力机制的解释方法、基于样本的解释方法、基于自然语言的解释方法、分析模型表示和参数的探测方法、神经元激活分析方法以及基于概念的方法；**基于微调范式的可解释技术则**包括分析模型新任务理解能力的方法、分析大模型少样本学习能力的方法、分析大模型思维链能力的方法、分析对齐微调作用的方法、分析大模型“幻觉”产生原因的方法。2023 年 3 月欧洲标准化组织 ETSI 亦提出有关人工智能透明度和可解释性的标准规范，旨在生成更多可解释的模型，同时保持高水平的性能，以创造人类用户能够理解、适当信任和有效管理新一代人工智能合作伙伴。

价值对齐自下而上获得各国政策法律认可。头部企业正在积极探索价值对齐解决方案。例如，OpenAI 宣布成立一个新的 AI 对齐团队，目标在 4 年内研究出让 AI 系统实现价值对齐和安全的方案，并投入 20%算力资源支持该工程。凤凰卫视发布首批“正向价值对齐数据集”和“中文访谈对话数据集”，推动 AI 数据领域华语数据的丰富与共享，为中华文化的传承与传播提供解决方案，让 AI 与中华文化认知对齐更简单。**价值对齐也已经写入部分国家政策文件。**美国《行政令》要求开发任何对国家安全、经济安全、公共健康和安全构成严重风险的人工智能模型的科技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。欧洲议会通过《人工智能法案》法律

草案，禁止实时远程生物识别技术；要求 OpenAI 和谷歌等公司必须进行风险评估，并披露更多用于训练模型的数据。

（四）信息内容治理

对用户的信息披露是应对信息内容风险的重要工具。布鲁金斯学会发布《应如何监管生成式人工智能？》一文指出，对生成式人工智能的监管可以从良好的消费者信息披露开始，更多的透明度和问责制必须是任何监管框架的核心。文章建议美国可参照食品和药物管理局的标签指南或将能源之星认证（Energy Star Rating）系统引入人工智能。**美国针对大模型标识制度形成了部分立法提案。**2023 年 5 月，美国众议员伊维特·克拉克（Yvette Clarke）提出《真实的政治广告法案》，对竞选广告提出人工智能生成内容披露要求。6 月，美国众议员里奇·托雷斯（Ritchie Torres）提出《人工智能披露法案》，提议在任何人工智能生成的内容中添加披露声明。**欧盟《人工智能法案》提出大模型披露（标识）要求。**即要求大模型类别的基础模型必须确保透明度，披露内容是由人工智能生成的。欧盟标识制度相较于中国标识制度，尚处于颗粒度较大的笼统要求阶段，对于义务的履行与落实尚未出台具体的实践标准。**我国已经前瞻性地构建了较为落地的大模型标识规范。**包括《生成式人工智能服务管理暂行办法》《互联网信息服务深度合成管理规定》《生成式服务内容标识方法》等，围绕文本、图片、音频、视频四类生成内容给出了内容标识方法，对提示文字的位置、大小、所含信息等作出标准化要求。

企业层面，国内外企业纷纷积极响应标识义务。一是企业发布内容标识平台规则。知乎、抖音、小红书等平台根据《生成式人工智能服务管理暂行办法》规定，探索发布大模型内容标识的规则，例如抖音发布“特定内容需主动添加标识”的公告，鼓励创作者尊重事实、发布客观真实信息，同时对于特定信息，应以“内容标识”的形式提供充分的说明。**二是推进研发新的标识工具**，例如，TikTok 于 2023 年 9 月推出了帮助创作者标记其 AI 生成内容的一款新工具，还将开始测试自动标记 AI 生成内容的方法。**三是行业协同履行标识义务。**谷歌、微软、OpenAI、亚马逊、Meta、Anthropic 等美国 AI 巨头公司在白宫做出自愿承诺，同意在音频和视频内容上使用水印来帮助识别人工智能生成的内容。**四是为用户明确义务要求。**知乎于 2023 年 4 月发布《关于应用 AIGC 能力进行辅助创作的社区公告》，要求创作者发布大模型生成的内容时需要主动使用“包含 AI 辅助创作”的标签进行声明，否则将被限流。

采取分级分类监管思路，赋予大型平台额外责任。一是依据平台用户规模进行分级管理。欧盟于 2023 年 8 月刚生效的《数字服务法案》中，将平台划分为一般平台和大型平台两级。其中，用户规模超过欧盟人口 10%（即 4500 万）的平台被界定为大型平台，相比一般平台，须在风险管理、合规审计、监管合作等方面承担额外义务。英国正在审议的《在线安全法案》草案中也采取了类似思路，提出应对具有足够覆盖范围与用户规模的网络社交平台给予特殊监管。**二是按照内容危害属性、平台受众对象等进行多维度分类管理。**英国《在线

安全法案》草案中，将网络社交平台信息内容划分为非法内容和合法但有害内容两类，并要求大型平台采取额外措施，保护用户免受合法但有害内容侵害。对于网络社交平台面向未成年人提供服务的情形，欧盟、英国法案中均明确提出特殊监管要求。如 TikTok 在监管压力下已下架针对 13 至 17 岁青少年的个性化广告推送。

四、把握全球大模型治理最新动态趋势

（一）美国从松散碎片式治理逐步趋向体系化治理

美国存在行政监管、立法、司法三条重要的 AI 治理线条。在司法层面，美国法院判例在塑造美国 AI 治理规则方面发挥重要作用，但当前除在知识产权领域外，尚未出现大模型相关直接司法诉讼。相比之下，政府监管和国会立法则呈现显著加速治理态势。

行政监管方面，拜登政府密切关注大模型风险，各部门监管动作频频。一是拜登政府发布多项政策文件，奠定“重塑人工智能全球领导地位”总基调。2022 年 10 月，发布《人工智能权利法案蓝图》，其中包含五项基本原则，分别为安全有效的系统、算法歧视保护、数据隐私、通知和解释、人工替代方案，反映了拜登政府对私营公司和政府机构鼓励采用人工智能技术的原则设想。2023 年 10 月，拜登签署《安全、稳定、可信的人工智能行政令》（以下简称《行政令》），包括了人工智能安全和可信标准、推动创新和竞争、支持劳工、促进公平和公民权利、维护消费者等群体的权益、保护隐私、确保政府负责有效使用人工智能、提升美国在海外的领导力等八个部分。**二是白宫开展相关治理行动，并对全球输出相关实践理念。**2023 年 7 月和 9

月，白宫先后宣布分两批召集谷歌、微软等共十七家 AI 头部企业，推动企业作出自愿承诺，包括开发让消费者能够辨别 AI 生成内容的方法，聘请独立专家评估工具的安全性，与外部行业分享管控 AI 风险的技术经验，允许第三方查找并报告其系统漏洞，报告其技术的局限性，优先研究 AI 在歧视和隐私方面的社会风险，以及发展 AI 以解决气候变化和疾病等社会挑战，并将其推行至澳大利亚、英国、尼日利亚等多个国家。美国《行政令》提出扩大多双边合作、开发国际标准等方式，强化人工智能领域国际合作,要求白宫和商务部领导建立强有力的人工智能国际框架，与国际合作伙伴和标准组织加快重要 AI 标准的开发和实施。**三是各部门多管齐下探索大模型监管。**商务部、联邦贸易委员会在各部门中作用尤为突出。商务部下属机构国家电信和信息管理局（NTIA）发布《人工智能问责制政策征求意见稿》，征求公众对“支持发展人工智能审计、评估、认证和其他机制以建立对人工智能系统的信任”的政策的反馈。商务部下属国家标准与技术研究院（NIST）发布《人工智能风险管理框架》，成为美国人工智能治理的事实规则。美国联邦贸易委员会发布《消费者保护指南》，强调委员会积极履行监管生成式人工智能的责任，保护竞争和消费者权益。其中包括了人工智能可能被视为欺骗的示例，以及评估算法公正性的维度。美国消费者金融保护局（CFPB）于 9 月发布了贷款人在使用人工智能和其他复杂模型时必须遵守的法律指南。

《行政令》是近期美国人工智能治理的重要行动，美国行政监管出现重大制度突破并日趋体系化。一是风险关注点有所转变。美国对

人工智能治理的问题焦点从偏见歧视、数据隐私扩展至更底层的人工智能技术对关键基础设施以及化学、生物、放射性、核和网络安全风险的威胁。**二是首次提出硬性监管效力的备案制度。**《行政令》援引《国防生产法》，要求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，在训练模型时就得通知联邦政府，并且必须分享安全测试结果和相关数据。**三是广泛动员联邦层面的监管部门出台技术指引。**除商务部、联邦贸易委员会外，行政令提及国土安全部、能源部、司法部、卫生与公众服务部等联邦政府机构，被称为全球人工智能治理中“最重量级”的政府行动。《行政令》要求相关部门对数字水印、红队测试等出台指引文件，规范相关技术工具和治理举措。

立法方面，大模型技术浪潮下，美国从州层面立法走向人工智能联邦立法雏形。一是州层面从多角度为人工智能立法提出立法建议。在 2023 年的立法会议上，至少有 25 个州、波多黎各和哥伦比亚特区提出了人工智能法案，15 个州和波多黎各通过了决议或颁布了立法。例如，美国康涅狄格州通过了关于人工智能、自动决策和个人数据隐私的“SB 1103 法案”。路易斯安那州通过决议，要求技术和网络安全联合委员会研究人工智能对运营、采购和政策的影响。**二是联邦层面出现多个立法提案。**以布卢门撒尔为代表提出的《两党人工智能立法框架》旨在建立严格的 AI 监管蓝图，其中包括设立由独立监督机构管理的许可制度、落实开发者问责制、捍卫国家安全和国际竞争、确保 AI 系统透明度以及保护消费者和儿童权益。以舒默为代表提出

的《安全创新立法框架》包含五大核心政策目标，涵盖安全、问责、与民主价值观保持一致、可解释和创新。**三是美国频繁召开立法听证会进行立法讨论。**在有关《两党人工智能立法框架》的听证会上，有建议要构建与 AI 技术相匹配的监管架构，并设立新的联邦机构来协调 AI 治理工作。在问责方面，有专家建议利用和发挥美国已有的措施实现问责制和确保透明度。还有建议引进“安全刹车”机制，通过实质性禁令来限制权力滥用。在有关该《安全创新立法框架》的闭门简报会上，提到的建议包括建立新的联邦机构监管 AI，鼓励 AI 人才移民美国等。

（二）欧盟继续发挥人工智能治理领域布鲁塞尔效应

欧盟在人工智能治理方面处于领先地位，力图保障本土数字主权与信息安全、提升人工智能规则制定国际话语权，影响全球数字治理规则。

针对 ChatGPT 的滥用问题，《通用数据保护条例》成为欧盟正式通过《人工智能法案》之前的“监管利器”。2023 年 3 月 31 日，**意大利个人数据保护局率先宣布暂停 ChatGPT 在意大利境内提供服务，成为首个禁用 ChatGPT 的欧洲国家。**监管机构列出四项违反《欧盟通用数据保护条例》的事由，包括训练数据缺乏合法性基础，提供虚假或错误的用户个人信息，存在数据泄露风险，缺乏用户年龄核查机制等。**意大利此次监管行动引发了德国、爱尔兰等其他欧盟国的密切关注。**德国、爱尔兰的数据保护机构与意大利个人数据保护局沟通，以了解其行动的依据。爱尔兰数据保护部门负责人表示，监管机构还

需要时间制定正确的监管措施，以免仓促实施任何“站不住脚”的禁令。

欧盟希冀通过《人工智能法案》继续发挥布鲁塞尔效应，实质影响全球规则走向。2023 年 6 月 14 日，欧洲议会以压倒性多数投票通过《人工智能法案》，进入欧洲议会、欧盟理事会、欧盟委员会三方谈判的最后立法阶段。

一是确立了基于风险的分级监管制度，根据风险等级提出不同义务要求。法案将人工智能系统的风险等级分为不可接受的风险、高风险、有限风险以及极低风险四类，并根据风险程度的高低配备不同程度的监管手段。四类风险等级覆盖的应用场景如下表。

风险等级	应用场景	监管措施
不可接受的风险	1.采用超越个人意识的潜意识技术或有目的的操纵或欺骗技术，其目的是通过损害作出知情决定的能力来实质性地扭曲该人的行为；（5.1（a）） 2.用特定群体在年龄、身体、经济状况、社会地位等方面的弱点，其目的是损害或实质性扭曲该人或该群体的系统； 3.基于敏感属性或特征的生物识别分类系统； 4.用于社会评分的系统或根据自然人或群体的社会行为或个性特征对他们进行分类的系统 5.在公共场所使用“实时”远程生物识别系统，以用于： • 评估自然人犯罪或再犯罪的风险，或预测刑事或行政违法行为的风险； • 通过无针对性的面部图像抓取来创建或扩展面部识别数据库； • 在执法、边境管理、工作场所和教育机构中推断自然人的情绪	• 禁止在欧盟市场投入使用 • 违反者将被处以最高 4 千万欧元的行政罚款或其上一财政年度全球年总营业额的 7%，以较高者为准
高风险	1.作为产品安全部件使用的系统，或该系统本身就是被欧盟协调立法范围所涵盖的产品，且其	入市前：①建立和维护风险管理制

	根据法律需要接受第三方机构的健康和安全风险评估才可以进入市场 2.符合高风险等级标准且用于下述领域的独立人工智能系统: • 生物识别和基于生物识别的系统; • 关键基础设施的管理和运作; • 教育和职业培训; • 就业、工人管理和获得自营职业; • 获得和享受基本的私人服务和公共服务及福利; • 执法工作; • 移民、庇护和边境管制管理; • 司法行政和民主进程	度; ②数据治理; ③制作技术文档; ④配备运行日志; ⑤通过合格评估程序; 入市时: ①系统注册; ②贴上欧盟 CE 标志; 入市后: ①部署后市场风险监测系统; ②采取纠正措施
有限风险	1.与人类互动的系统 2.情绪识别系统 3.生物特征分类系统 4.生成或操纵图像、音视频等内容的系统	透明义务要求
极低风险	允许自由使用人工智能的电子游戏或垃圾邮件过滤器等应用	不做干预

表 1 欧盟《人工智能法案》人工智能系统风险等级及监管措施分类表

二是将高风险人工智能系统作为监管重点。为此，《草案》针对高风险人工智能系统规定了从入市前到入市后的全生命周期合规要求，并对高风险人工智能系统价值链上的多方参与者规定了不同程度的义务，其中系统提供者需要承担最严格的义务。具体而言，入市前，提供者的义务包括（第 16 条）: ①建立和维护风险管理制度的义务，风险管理制度应该涵盖人工智能系统的整个生命周期，并应包含（a）识别和评估已知和可预见的风险、（b）评估上市后风险、（c）采取针对性风险解决措施三个步骤; ②数据治理义务，应在满足一定质量标准的数据上进行训练，采取措施缓解可能的数据偏见，使数据具有代表性、准确性和完整性; ③制作技术文档的义务，用于主管机构据以评估系统的合规表现; ④配备运行日志的义务，系统在设计和开发时

应该有自动记录功能（日志），用于持续追踪和监测系统的风险；⑤通过合格评估程序，提供者应确保系统在投放市场或投入使用之前通过相关的合格/符合性评估程序（conformity assessment procedure），由评定机构对人工智能系统是否满足法案第二章所规定的各项要求进行验证。

进入市场时，提供者应当：①将系统在欧盟数据库中注册（第 51 条）；②贴上实体/数字形式的欧盟 CE 标志（欧洲共同市场安全标志）；进入市场后，提供者应当履行：①部署后市场风险监测系统（post-market monitoring system）的义务，用于收集、记录和分析系统在整个生命周期内运营性能数据；②采取纠正措施的义务，提供者如果认为或有理由认为系统在投入市场后不合规，应立即采取撤回、失效、召回、通知其他参与者等纠正措施。

三是探索监管工具创新。例如，引入监管沙盒制度，即要求各成员国在 AI 系统开发和上市前的有限时间内建立一个受控的实验和测试环境，使得沙盒参与者能够在获得特定法律条款或合规流程的豁免的情况下使用个人数据来促进人工智能创新，一方面减少企业在技术开发早期的合规成本，另一方面也有利于监管机构给予动态监督和指导，加速监管制度的完善。2022 年 6 月，西班牙政府和欧盟委员会提出了欧盟首个人工智能监管沙盒项目，旨在为中小企业利用人工智能创新创造良好环境，并且为正在快发展中的人工智能找到最佳的监管方式。目前该试点向其他成员国开放，其监管沙盒测试结果将在 2023 年下半年西班牙担任欧盟理事会主席国期间公布。

欧盟《人工智能法案》对基础模型治理树立了风向标。一方面，欧盟就生成式人工智能是否应列为高风险作出妥协。2022 年 12 月，欧盟理事会在提案中加入通用目的人工智能的概念。2023 年 2 月，相关部门建议没有人类监督的情况下生成复杂文本的 AI 系统（例如 ChatGPT、Dall-E 等）应被列为“高风险”。2023 年 5 月，“欧洲合作观察”调查显示，微软、谷歌等游说将 ChatGPT 等 AI 工具排除在高风险 AI 系统监管范围之外。OpenAI 首席执行官 Sam Altman 巡回欧洲，主张将 ChatGPT 排除在高风险系统之外。2023 年 6 月，《草案》将基础模型（包括 AIGC）单列一项，从高风险 AI 清单中移除。**另一方面，法案草案细化落实基础模型治理，对类似于 ChatGPT 的基础模型施加了若干义务要求。例如，法案第 28b 条要求，一是评估要求，**基础模型的提供者应当评估基础模型是否在其生命周期都保持适当的性能、可解释性、可更正性、安全性（c 款）。**二是备案要求，基础模型应在进入市场前在欧盟数据库注册（g 款）。三是其他透明度要求。**基础模型提供者应加强人工智能生成内容的披露要求、防止模型生成非法内容、发布受版权法保护的训练数据的使用情况摘要、公开训练数据来源等。

（三）英国力图以促进创新的监管方法引领全球治理

在国家背景上，英国的 AI 发展实力位居世界前列。英国 AI 企业数量占欧洲总数的三分之一，并有 Deepmind 等前沿代表，拥有艾伦·图灵研究所、牛津互联网中心、帝国理工学院等世界领先的高校与研究机构，技术人才优势显著。在战略层面，英国致力于将自身打

造为全球人工智能创新中心和发展高地。得益于脱欧后更为广阔的政策空间和立法主动权，英国正在“强硬的欧盟”模式和“较少干涉的美国”模式之间寻求平衡点，为 AI 产业战略配套有利的政策环境。

英国以安全为监管落脚点，积极探索促进创新和竞争的人工智能监管方案，在治理路径上体现出有别于欧美的特色。**一是从宏观层面阐明促进创新的监管政策方向。**2023 年 3 月，科学、创新和技术部（DSIT）发布《促进创新的人工智能监管方法》白皮书。治理方法上，初期不采取严格法律责任的人工智能监管框架，而选择支持创新的灵活治理路径，利用现有法律框架实施监管，避免严格繁重的新立法阻碍创新；优先考虑如指南、行业标准等缓和的干预措施；主张敏捷治理，实时评估监管框架执行情况，以高适应性的迭代方式及时改进。治理内容上，不针对特定技术或全行业，而关注人工智能部署中可能产生的结果，特别是已发生、可识别的高风险，采取相称监管措施。监管机构上，保证跨部门一致的监管承诺，降低企业合规负担，（1）依托现有监管机构，在各自专业领域内有效执行监管原则；（2）在政府内部构建新的“中央职能”支持工作并确保协调性和一致性，暂不建立新的专门监管机构；（3）开发跨机构的监管沙盒和测试平台，以支持创新者将尖端产品推向市场。**二是率先基于竞争视角，探索基础模型对市场竞争和消费者的影响，创造良好竞争生态。**2023 年 9 月，市场和竞争管理局（CMA）发布了《人工智能基础模型初始报告》，确保竞争和消费者保护在基础模型开发和部署中的驱动作用。报告指出，（1）在基础模型开发中，大型科技公司在专有数据、计算

能力等方面的优势可能构成市场壁垒，阻碍竞争;（2）在医疗、教育等下游服务市场，应保障用户有效选择和切换的能力、数据的可移植性，限制上下游垂直垄断;（3）基础模型应更符合消费者利益，明确责任主体、提供正确且充分的信息供消费者了解并选择基础模型。基于此，报告提出问责制、可及性、多样性、灵活性、充分选择、公平交易、透明度等确保基础模型市场竞争性的原则，并指出大规模并购、滥用市场支配地位、过度封闭的生态系统、捆绑搭售等反竞争行为可能危及上述原则。**三是成立人工智能基础模型工作组，推进安全可靠的人工智能开发。**2023 年 4 月，英国政府宣布投入 1 亿英镑初始资金建立基础模型工作组,旨在通过强化人工智能基础模型的发展和应用，提升英国战略技术上的全球竞争力。在实施计划上，工作组将汇集政府和行业专家每月直接向英国首相和技术部长报告，在六个月内启动第一批针对公共服务的试点项目，力图将英国打造为基础模型及其经济应用全球领导者和人工智能安全的“全球旗手”。

在国际合作层面，英国试图通过人工智能安全峰会打造全球人工智能监管的地理中心。**一是峰会致力于支持国际包容性的前沿人工智能安全科学研究网络，涵盖并补充双边乃至多边合作，为决策和公共利益提供科学参考。**在国家层面，各国应重视创新并采取相称的治理和监管方法，根据不同国情和法律框架进行风险分类，制定基于风险的政策；在国际层面，识别共同关注的人工智能安全风险，建立对风险共同的科学理解。**二是发布《关于人工智能安全的布莱切利宣言》，促成各国对人工智能的机遇和风险、以及在安全领域采取协作的共**

识。全球二十八个国家共同的宣言指出，当前迫切需要借助新的全球合作了解和集体管理潜在人工智能风险，以人为中心、以安全的方式设计、开发、部署和使用人工智能。在风险共识上，（1）前沿人工智能的特殊安全风险，使通用模型和基础模型可能造成严重甚至灾难性的伤害，因此加深理解和采取行动尤为紧迫；（2）人工智能产生的许多风险具有国际性，需要通过国际合作解决。**三是设立全球首个人工智能安全研究所，负责测试前沿人工智能的安全性，以巩固英国作为人工智能安全世界领导者的地位。**英国政府宣布与美国、新加坡、谷歌、DeepMind 等国家和企业合作，设立全球首个人工智能安全研究所。研究所旨在争取世界领导人和主要人工智能公司的集体支持，使英国占领人工智能安全的中心地位。其主要工作是，在新型人工智能发布前后开展涵盖所有风险的测试，以消减人工智能模型潜在的危害。**四是制定前沿人工智能模型的安全测试计划，建立对前沿人工智能能力和风险的共同理解。**英国联合在人工智能领域领先的国家政府首脑与开发公司代表商定：（1）政府和企业合作开展前沿人工智能安全测试，确保模型部署前后的安全，以应对关键的国家安全和社会风险；（2）各国政府达成共识，投资公共部门的安全测试及其他安全研究，并适时考虑共享评估结果、制定共享标准，为未来人工智能安全方面的国际合作奠定基础。

（四）国际组织在大模型治理国际合作中各显其能

当前，人工智能治理问题在全球范围内引起了各大国际组织的密切关注。以联合国、G7 集团、金砖国家为代表的涵盖发达国家与新

兴市场等各类型经济体的国际组织，均在通过发布文件、举办会议、磋商合作等方式，积极参与人工智能监管体系与国际合作模式的构建。

一是联合国成为推动人工智能全球监管合作的重要机制。发布系列文件助力全球伦理共识落地。2023 年 5 月《生成式人工智能在教育和研究中的应用指南》为人工智能伦理教育应用建立政策框架。6 月与欧盟签署协议，预计拨款 400 万欧元帮助部分欠发达国家建立相关法案，落实《教科文组织人工智能伦理建议书》。10 月与欧盟改革总干事和荷兰数字基础设施管理局合作，分析人工智能监管落地设计。**支持成立跨国监管与咨询机构。**6 月推动成立国际人工智能监管机构，定期审查相关治理工作。10 月成立高级别人工智能咨询机构，呼吁全球性、多学科、多利益相关方对话。**健全全球安全风险应对机制。**5 月发布《我们的共同议程》政策简报“全球数字契约”，以落实人工智能等新兴技术治理为目标，立足当前全球技术发展和市场应用间的治理差距和潜在红利，强调各国合作确定、减轻风险的必要性。7 月在纽约召开人工智能风险问题会议，首次正式讨论该议题，凸显全球关注的态势。此外，“新和平纲领”针对人工智能治理提出建议，尤其呼吁在 2026 年前完成有法律约束力的国际文书谈判工作，以禁止在没有人类控制或监督的情况下运行致命自主武器系统。

二是 G7 峰会推进构建一致且互操作性的人工智能监管规则。2023 年 5 月，七国集团（G7）领导人在日本广岛举行年度峰会，就包容性人工智能治理和互操作性展开国际讨论，以实现构建值得信赖

的人工智能、符合共同民主价值观的共同目标。此前，美国战略与国际问题研究中心（CSIS）发布《在 2023 年 G7 峰会推进人工智能治理合作》报告指出，**人工智能的突破性发展使其成为当前发达国家有效应对劳动力短缺的重要方式，国际间一致且可互操作的监管框架对其持续进步和高效应用至关重要。**为在发达民主经济体中实现一致且可互操作的人工智能监管，报告建议：（1）统一人工智能监管框架的规范和概念，在核心原则、基本术语和关键领域等方面达成共识。（2）合作制定人工智能技术标准，建立互认框架，提升跨司法辖区定义的一致性和协调性。（3）达成良好且道德的人工智能原则共识，构建可互操作、平衡、互认的人工智能监管方案。总体上，G7 集团聚焦人工智能技术革新对发达国家经济现状和未来发展的影响，尤其关注人工智能监管格局分散的现实风险和潜在挑战，并以此为基础推进合作。

三是金砖国家在人工智能合作上的优势凸显。以新兴市场和发展中国家为代表的**金砖 11 国在人工智能领域具有丰富应用场景、良好产业基础、先进治理经验和广阔合作前景，是国际人工智能治理的重要力量。巴西** 2021 年发布人工智能战略，提出人工智能开发和使用原则。**南非**发布“非洲人工智能蓝图”，积极引领相关战略合作，推进人工智能在农业、医疗保健、教育、金融、能源交通与气候变化领域的应用。**沙特阿拉伯** 2020 年发布国家数据和人工智能战略，2022 年发布《人工智能道德准则》，提出将人工智能伦理融入其系统全开发周期。**印度** 2015 年发布“数字印度”战略，2018 年《国家人工智

能战略》提出由政府推进标准化负责任的人工智能开发，2023 年提出以市场为核心发展人工智能。此外，埃及、埃塞尔比亚、俄罗斯等国也通过举办会议、发布战略等方式拓展人工智能技术与治理。**当前，应加强金砖国家人工智能治理多边合作机制，分类推进双边合作，更好地发挥其在全球人工智能治理中的积极作用**。在合作上，**以元首外交引领人工智能合作**，形成以元首会晤为引领，以部长级会议为支撑的合作机制；**具有深化人工智能合作的良好基础**，已在经贸、科技、农业、教育等数十个领域务实合作，形成共同推动制定合作协议、行动计划等多层次合作架构，围绕人工智能也已开展多种合作。

五、探索我国大模型治理的主要落地工具

有效治理人工智能，离不开可行的治理工具和技术手段。加强对大模型的有效管理，可从事前算法备案、事中风险评估、事后溯源检测等方面出发，进一步探究应对大模型产品的不可控及滥用风险的方法，逐步完善大模型治理体系。

（一）事前备案

算法备案作为算法治理体系的重要监管内容，是算法透明度要求的落地方式之一，旨在保护用户权益，维护产品安全和信息安全。当前，各国均已就大模型产品算法备案作出探索，并进行前期实践。

1.大模型备案制度现状

我国《生成式人工智能服务管理暂行办法》提出算法备案要求。其中第十七条提出，提供具有舆论属性或者社会动员能力的大模型服务的，应按照《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。大模型服务提供者可在互联网信息服务算法备案系统中填报主体、算法、产品相关信息，完成填报手续。

算法是算法备案的基础填报单元和备案编号的承载对象，主要分为基础属性信息与详细属性信息两部分内容，其中基础属性包括算法名称、角色、应用领域、《算法安全自评估报告》和《拟公示内容》等信息，算法详细属性包括算法数据、算法模型、算法策略、风险与防范机制等内容。填报者应以算法为锚点，填报并关联使用算法的产品及功能。以“文心大模型算法”为例，以“北京百度网讯科技有限公司”作为填报主体，应填报其主体、算法相关信息，并关联使用“文心大模型算法”的“文心一言（APP、网站）”产品。审核通过后，备案编号将以算法为载体发放至备案主体，主体应在其对外提供服务的网站、应用程序等显著位置标明其备案编号并提供公示信息链接。

2.我国备案制度落地实施

截至 2023 年 11 月，国家互联网信息办公室已发布两批深度合成服务算法备案编号共 151 个，其中生成式人工智能算法备案编号 100 个，包括服务提供者角色备案编号 64 个，服务技术支持者备案编号 36 个。**从备案主体角度出发**，已发放备案编号的主体覆盖 11 个省份，北京、广东、浙江、上海四个省市占据生成式人工智能算法备案总数的 87%。**从备案算法角度出发**，文本生成类占据生成式人工智能算法备案总数的 54%，图像生成、音频生成、视频生成类占据生成式人工智能算法备案总数的 38%。**从产品及服务角度出发**，备案产品涉及各类文本、图片、语音、视频、虚拟人像等生成合成场景，广泛应用于

电商、金融、医疗、教育等领域。**在面向 C 端应用领域，**以“WPSAI 文本生成算法-1”为例，WPS 推出了中国协同办公赛道上首个生成式 AI 应用 WPS AI，当前 WPS 已经具备全新升级的智能文档、智能表格、PPT 演示等便捷功能。**在面向 B 端应用领域，**以“达摩院交互式多能型合成算法”为例，达摩院推出的大模型应用于开放域多模态内容生成场景，服务于问答、咨询类的企业端客户，通过 API 提供根据用户输入生成多模态信息的功能。当前，“文心大模型算法”、“云雀大模型算法”、“讯飞星火认知大模型算法”等大模型算法已相继完成备案。

（二）事中全流程评估

在大模型治理中，风险评估是保障其合规性的关键环节之一，亦是保障技术安全和输出内容安全的重要手段。包括我国在内的主流国家均对此类产品的安全风险评估的实施作出了积极探索。

1.我国评估制度现状

从制度设计来看，我国大模型服务在上市前及运行中需开展安全评估，《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大模型服务的，应当按照国家有关规定开展安全评估。对具有舆论属性或者社会动员能力的信息服务开展安全评估的要求始于 2017 年 12 月 1 日公布的《互联网新闻信息服务新技术新应用安全评估管理规定》，又被称为“双新评估”，后在 2018 月 30 日公布的《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》中进一步完善明确，对信息内容安全风险的关切是

评估的核心出发点。**标准层面，**我国已经出台针对算法安全评估国家标准，尚未出台独立的大模型评估标准或规定。中国电子技术标准化研究院、中国科学院计算技术研究所联合 33 家单位，共同研制发布了《信息安全技术 机器学习算法安全评估规范》，包括评估准备、评估方案、评估执行、评估结论、评估报告等内容指引。中国人工智能产业发展联盟起草《大模型训练模型技术和应用评估方法》，探索相关可控可信的具体规范要求，针对大模型的评估标准业界均在探索制定中。

2.评估制度落地情况及问题

从评估实践来看，目前学界业界发布了各类评测工具和评测平台。现阶段的大模型测评集侧重对大模型能力强弱进行评测，包括中文语言理解能力、中文知识运用和推理能力等，数据来源主要是各类考试，数据形式主要为选择题、判断题等客观题形式。在大模型安全测评方面也有了一些尝试，例如，清华 CoAI 小组推出中文大模型安全评测平台[43]，针对大语言模型伦理安全问题提供评测服务；中国信通院建立相关公共服务平台，开展大模型安全防范能力系列评估规范和评测；智源人工智能研究院建立“能力-任务-指标”三维评测框的 FlagEval 天秤大模型评测体系及开放平台[44]；蚂蚁集团发布了大模型安全一体化解决方案“蚁天鉴”，包含大模型安全检测平台“蚁鉴 2.0”和大模型风险防御平台“天鉴”。

[43] 参见 http://115.182.62.166:18000/
[44] 参见 https://flageval.baai.ac.cn/#/home

实践中，我国大模型技术社会化测评还存在一些问题，从评估技术来看，一方面是由于评估内容较多，缺乏高效且全面的评估工具、系统等技术；另一方面大模型测评尚未形成统一的评价方式与指标体系，评价的一致性和客观性还需要进一步论证。**从评估市场来看，**科研机构、高校院所等领军主体各自为战现象较为突出，存在评估标准过多、榜单刷分注水严重、评估结果差异大等问题。例如，某大模型在 SuperCLUE 榜单中排名第十，而在知名咨询公司 IDC 的《2023 AI 大模型技术能力评估报告》中则排名靠前。

3.欧美路径迥异的评估监测工具方案

美国针对大模型产品或服务上市的评估目前处于探索阶段，当前主要采用社会化评估的方式。首先，美国政府已在事前评估方面进行了一些探索，并试图以立法形式固定成果，但并不顺利。2022 年 2 月，美国多位议员提出《2022 算法问责法案（草案）》，以此构建算法影响评估机制，要求科技企业在使用自动化决策系统做出关键决策时，对偏见、有效性和相关因素进行系统化的影响评估，并确立联邦贸易委员会（FTC）作为评估主体，以内容、数据、网络、经济、军事安全等为审查内容，形成政府指导、企业提交报告、建立 FTC 数据库并公示企业报告等一套流程。该法案具有里程碑意义，但迄今为止尚无实质性进展。**其次，大模型产品快速应用发展促使政府尝试引导社会化评估以增强公众对新技术的信任。**2023 年 4 月，**美国商务部**国家电信和信息管理局（NTIA）发布《人工智能问责制政策征求意见稿》，重点关注如何标准化评估、多项目标之间的平衡、实施

问责机制的难度等问题，并围绕模型代码可靠性、生成信息安全性、攻击抵御能力、外部工具交互安全性等开展全面评估。**美国白宫**亦推动 OpenAI、谷歌、微软等企业在 DEFCON 计算机安全会议上开放模型供与会者进行渗透测试、查找漏洞。DEFCON 的活动将允许数千名社区合作伙伴（黑客、工程师、研究人员等）和人工智能专家对这些模型进行彻底评估，以探索这些模型如何达到《人工智能权利法案》和《人工智能风险管理框架》规则的一致性。

欧盟对基础模型的评估与高风险人工智能评估要求有所差异。首先，在 2023 年 6 月欧盟议会通过的《人工智能法案》妥协案中将基础模型（包括大模型）单列一项，未将其视为高风险人工智能。**具体而言，对高风险 AI 的评估，采取严格准入，要求获得 CE 标志**，即通过强制性 CE（Conformity With European）标记程序，要求高风险人工智能系统须完成市场准入和认证。具体评估认证则由监管机构指定的具有独立性、相应能力、无利益冲突和满足最低网络安全的要求第三方机构进行。**目前对基础模型的评估则要求通过适当的方法进行自评估或聘请独立专家参与，**进行模型评估、记录分析，以及在概念化、设计和开发期间进行广泛的测试。

（三）事后溯源检测

大模型内容溯源用于解决内容来源判别的问题，一方面判断内容来源于人类还是大模型，另一方面判断内容来源于哪个大模型。溯源检测技术可用于防范生成内容的滥用、对人类撰写数据的污染以及细粒度追踪生成内容的来源。常见的技术手段包括基于隐式标识的检测、

基于内容分布的检测等，其中前者需要大模型服务提供者的参与，后者不依赖于大模型服务提供者。

1.标识溯源：基于内容标识检测的风险溯源

隐式标识主要指通过修改文本、图片、音频、视频内容添加的人类无法直接感知的标识。隐式标识可通过技术手段从内容中提取，且具有溯源效力，目的在于支撑大模型生成内容的检测和溯源。

当前隐式标识已逐渐在大模型领域得到布局和应用。国内发布相关内容标识行业标准提出，利用 AI 技术生成图像、音频、视频内容时，应在内容中添加数字水印标识，以文件形式输出音视图文时，还应在文件元数据中进行数字水印标识。**就国内而言**，阿里巴巴利用数字水印技术保护大模型生成的音视图文等内容，如在“通义万相”文生图、淘宝 AI 试衣间、淘宝人生等服务生成的图像内容中添加暗水印，在“通义千问”、IdeaLAB、钉钉文档等业务中添加可以抵抗截图的暗水印，达摩院数字人、手淘某 AI 推荐功能等大模型相关服务也已接入暗水印[45]。**就域外而言**，Meta 和法国国家信息与自动化研究所（INRIA）联合开发了 Stable Signature，可将数字水印直接嵌入到 AI 自动生成的图片中，防止其被用于非法用途。Stable Signature 生成的数字水印不受裁剪、压缩、改变颜色等破坏性操作影响，能追溯到图片的初始来源，可应用于扩散、生成对抗网络等模型，例如著名文生图软件 SD 和 MJ[46]。

[45] 参见 https://www.163.com/dy/article/ICMA9DOS0514R9KQ.html
[46] 参见 https://www.sohu.com/a/727144326_121649381

虽隐式标识使大模型内容具备了基本的溯源追踪能力，但仍面临一定的挑战。从标识对象的角度来看，技术层面上尚无法保证全面性。具体而言，现有大模型标识对象主要是图片或视频，对文字内容进行标识仍存在技术上的难度。**从跨平台互通互认的角度来看**，大模型标识方案亦未达成统一标准，例如，非同质化通证（Non-Fungible Token，NFT）类大模型标识往往采用较长的哈希标识符，而面向用户生成内容的标识符则通常采用平台随机生成或者用户自主命名的形式。**从标识溯及力的角度来看**，亦无法实现对内容提供者的追溯。现有大模型标识多将标识作为数据的一部分嵌入生成内容，内容提供者可以控制这些数据的生成和使用，相关责任人仅能追溯到内容使用者而非内容提供者。

2.内容溯源：基于内容分布检测的风险溯源

大模型检测工具基于生成内容进行溯源，不依赖于标识、生成日志等辅助信息，是落实“以技治技”治理理念的重要内容。

社会各界当前已积极开展自动化检测方法与工具的研发探索。从文本检测来看，区分人类和大语言模型生成的文本成为至关重要的问题。大语言模型厂商和研究机构纷纷公布生成内容识别工具。例如，OpenAI 推出名为“分类器”的 AI 生成内容识别器，斯坦福大学推出 DetectGPT 方法识别机器生成文本。中国科学院计算技术研究所提出名为 LLMDet 的检测工具，相比于现有检测方法，该工具可定位生成文本来源的基础模型，在确保速度和安全性的同时展示了不错的检测

性能[47]。同期，北大、华为的研究团队亦提出多尺度学习方案，以改进 AI 生成语料的文本检测器性能[48]。**从图像视频检测来看，**多媒体合成技术的检测及溯源需求愈加迫切。2022 年 11 月，英特尔推出深度合成检测工具 FakeCatcher，通过检测血流判断 AI 换脸视频，官方称其可进行实时检测，并在几毫秒内显示结果，在检测人工智能算法制作伪造视频、AI 篡改视频方面的准确性达 96%。哈工大（深圳）和南洋理工的研究人员提出了检测及定位多模态媒体篡改任务并开源了多模态媒体篡改数据集，相较于已有的单模态深度伪造检测任务，在识别输入图像-文本真假的基础上，还可进一步定位到详细的篡改内容[49]。

生成内容检测仍面临部分亟待解决的通用性挑战与困难。一是由于语言自身的复杂性，导致识别难度高，如 OpenAI 检测器对 AI 撰写内容检出成功率仅为 26%；**二是**大模型生成内容变化多、随机性高、数量大，对检测工具的时效性和检测效率提出较高要求。**三是**大模型检测现多以产品维度展开，大模型产品百花齐放，且未实现多产品的数据互通，对检测工具的普适性提出较高要求。

六、完善我国大模型治理体系的思路建议

当前，人工智能治理已从理念探讨走到了实践探索的前沿。面对呈指数级增长态势的大模型应用，大模型治理应当协同多元主体、兼

47 参见 https://arxiv.org/abs/2305.15004
48 参见 https://arxiv.org/abs/2305.18149
49 参见 https://arxiv.org/abs/2304.02556

顾多维目标、融合多元价值，把握治理重点、创新治理工具，加强全球合作与对话，推动构建包容共享的人工智能治理体系。

（一）确立促进创新的人工智能敏捷治理理念

创新是引领发展的第一动力，应探索敏捷治理理念，建立灵活性、全面性制度框架，推动人工智能高质量发展和高水平安全实现良性互动。**一是平衡创新发展和风险治理。**通过敏捷治理实现多项目标的平衡，不是一味强调风险控制，也不片面追求效率。鼓励大模型技术在各行业、各领域的创新应用，支持相关机构在技术创新、数据资源建设、转化应用、风险防范等方面开展协作，鼓励基础技术的自主创新。**二是持续强化跨部门协同机制。**强化跨部门协同是当前大模型监管的必然选择，应支撑完善跨部门、跨区域政策协调、执法联动响应和协作机制建设，着力解决部门职能交叉、监管信息不共享等难题，推动协同监管制度化、常态化。**三是建立健全多元敏捷互动机制。**打造政府主导、企业自治、行业自律、社会监督的社会共治模式。政府引导企业查找问题、改进设计、降低风险，协调解决试点企业相应困难。企业定期报送风险阶段性评估报告，建立完善内部监测与预警机制，在发生重大风险后及时上报事件情况，提高全社会防范意识并鼓励公民监督。

（二）聚焦人工智能场景应用细化制度方案

一是推进《人工智能法》等立法进程，从产业发展、伦理引领、底线红线等维度，明确制度规范。在我国《著作权法》“合理使用”情形中增加“文本与数据挖掘”例外条款，正面回应人工智能作品使

用问题。推进完善数据共享流通规范，建立健全大模型场景下个人信息保护实施细则。**二是以监管沙箱试点摸清场景应用特点风险。**建议选取传媒、教育、医疗等大模型成熟应用的领域，开展人工智能治理试点工作，鼓励企业积极试行治理方案和工具，摸清主要场景和关键环节的风险问题，完善大模型技术应用全流程、全要素制度供给体系。**三是在重点场景下针对典型风险细化规则方案。**依据大模型部署方式、应用场景等探索差异化治理措施。由政府主导、委托第三方机构建立权威实施细则或标准，围绕大模型技术能力、训练数据、数据标注等多环节、多领域建立细化规则，明确评估标准和流程。**四是建立完善大模型分级分类清单。**根据沙箱经验，全面调查评估大模型风险等级，细化大模型分级分类清单，通过出台规范性文件、行业标准等明确细化分级分类的可操作标准，并根据实际情况做动态调整。

（三）立足当前治理实践创新人工智能治理工具

大模型的治理，既需要完善治理理念与规则，也需要优化治理手段与能力，进一步更新丰富治理工具箱。**一是优化监管制度工具以推进事前、事中、事后全流程监管。**从风险等级、新技术新应用类别等明确评估效力，完善鲁棒性、安全性、隐私性、公平性等多维评估指标，统筹信息内容风险、个人信息保护、安全性、版权保护等评估制度,发布数据审查库、数据标注规范等具体评估指引。**二是强化大模型监管平台、技术工具等资源配备。**构建国家级大模型测试验证平台，提供模型测试验证、供需对接等服务，落地模型对抗安全、后门安全、可解释性等检测能力，推进加固工具等技术开发共享。构建官方大模

型训练及测试数据集，降低优质数据获取成本。增强大模型风险的动态感知、科学预警、留痕溯源、调查取证能力，提升治理专业化、精准化、智能化水平。**三是引入社会化力量提升大模型评估服务水平。**加强人工智能领域第三方评估机构力量，明确人员专业能力、技术工具储备、资源平台建设等资质认定条件，并定期进行资质年审，共同构建优势互补、协同发展的服务网络，积极推动国际测试互认和互操作性。

（四）激励企业积极管控风险以推动平台合规

企业合规是企业依法依规经营、防控合规风险的一种自我治理方式，应贯彻多元主体治理思路，借助社会化力量建构大模型平台治理新格局。**一是建立健全内部合规组织架构与工作机制。**平台应搭建职责明确、层次清晰、协同高效、管控严密的一体化合规管理组织架构，以分层管理、全面覆盖为要求，精准分配合规管理职能。建立协同配合的合规管理工作机制，保障合规各部门有效开展工作。**二是优化平台内部治理体系应对内外部风险。**构建日常风险监测机制、违规行为举报机制与合规报告机制等，畅通用户反馈渠道，优化人为监督、用户投诉举报和补救程序等。强化平台添加标识和识别标识能力，建立统一识别标准，由政府推动第三方平台开发免费的标识工具。优化平台内容治理策略，如在输入端识别违规数据，提示使用者并驳回本次生成请求，对提示后仍频繁输入违规数据的用户账号进行进一步处罚。**三是建立监管部门合规评价体系以落实平台合规。**合规评价应重点关注平台合规组织体系、合规义务体系、风险监测体系等形式要素是否

完善，并强化对合规文化建设情况的评估。建立合规减责免责机制，对于大模型治理中有积极探索和明显成效的，在国家项目申报、政府公共服务资源采购等方面提供优惠激励政策。

（五）促进全球人工智能合作治理体系构建

人工智能治理攸关全人类命运，是世界各国面临的共同课题，积极参与并推动国际合作治理有助于形成该领域的共赢新局面。**一是推动包容开放的人工智能全球对话。**建立真正具有广泛代表性的全球人工智能治理对话机制，围绕共同风险凝聚共识。建议成立政府间的咨询和评估机构，围绕人工智能对经济社会的潜在影响、风险评估、治理框架等重大问题开展交流。**二是帮助后发国家更好获取和利用人工智能技术、产品和服务。**人工智能已成为数字时代生产底座。发展中国家普遍缺乏数字基础设施、创新环境、技术人才等关键要素，人工智能产业和应用发展受限，数字鸿沟进一步拉大。建议围绕人工智能设计合理的融资、援助和能力建设机制，促进人工智能技术公平获取和安全使用。**三是推动人工智能国际测试评估合作。**国际标准化组织（ISO）、国际电工委员会（IEC）已围绕关键术语等开展标准研究，但短期难以响应国际社会对人工智能安全的急迫需求。建议积极推动人工智能研究合作，广泛汇集各国人工智能专家，在尊重各方文化多样性、政治安全等诉求的基础上，共同探索测试评估方法，协助后发国家共同降低大模型技术风险。

中国信息通信研究院 政策与经济研究所

地址：北京市海淀区花园北路 52 号

邮编：100191

电话：010-62301327

传真：010-62302476

网址：www.caict.ac.cn