2023

SFC 南财智库

观韬中茂律师事务所
Guantao Law Firm

中国AI治理的独立思考

——生成式人工智能发展与监管白皮书

指导单位
南方财经全媒体集团

主办单位
21世纪经济报道

联合出品
南财合规科技研究院、观韬中茂律师事务所

PREFACE 前言

1956 年，在美国汉诺威小镇的达特茅斯学院，“人工智能”的概念被首次提出。此后 60 余年，人工智能从屏幕上显示的代码逐渐转化成实践应用。但人工智能的规模化商用并非坦途，概念的火热一直以来未能助推技术突破与商业应用。

时间来到 2022 年，生成式 AI 发展为人工智能发展再注入一针强心剂。先是 Dall-E2、Midjourney、Stable Diffusion 等文生图应用引起广泛关注，接着 ChatGPT 横空出世，被视为通用人工智能的起点和强人工智能的拐点。作为里程碑式的技术进步，ChatGPT 将引发新一轮人工智能革命。人工智能发展似乎找到了自己的主流叙事。

不过，技术创新的同时也带来了监管难题。如何平衡发展与安全，中国正在摸索自己的 AI 治理路径。在此环境下，身处其中的行业创新者、技术使用者，以及作为受众的社会公众，又该如何理解生成式 AI 发展的现状与前景，应对可能的风险与挑战？在此背景下，本白皮书将通过分析生成式 AI 的发展现状、政策导向、实操中面临的风险，以及各国的监管路径，以期为未来的 AI 治理提供有益思路。

CONTENT 目录

第一部分

发展：生成式 AI 治理的第一视角

2022 年 11 月，OpenAI 推出的聊天生成预训练转换器（ChatGPT）的爆火，带来了人工智能的“iPhone 时刻”。

该产品以强大的文字处理和人机交互功能迅速风靡全球。数据显示，发布五天内其用户量就达到了 100 万，并在短短 2 个月内用户量破亿，取得现象级战绩。

以 ChatGPT 等大语言模型为标志的生成式 AI 的成功，带来了新的范式革命和广阔的商业前景，资本市场持续高涨的热情也足以彰显它的价值。不过，一个硬币总有正反两面，生成式 AI 技术在为经济社会发展带来新机遇的同时，也引发了诸多舆论争议，带来了虚假信息传播、个人信息权益侵害、数据安全、偏见和歧视等问题。

事实上，全球正在进入“生成式 AI 革命风暴”，随之掀起的是新一轮的 AI 监管潮。

欧盟领先一步，《人工智能法案》进入最终谈判阶段。

系列迹象表明，美国政府最近也在紧锣密鼓地推进监管工作：6 月 20 日，美国总统拜登就会见了 AI 专家和研究人员，讨论如何管理 AI 在就业、儿童权益、偏见和成见以及信息方面带来的机

会和风险。美国政府正在考虑为这项快速发展的技术制定具有法律约束力的规定。

中国人工智能法草案也将提请全国人大常委会审议。在 4 月发布《生成式人工智能服务管理办法（征求意见稿）》并向社会公众征求意见后，7 月 13 日，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、广电总局正式发布《生成式人工智能服务管理暂行办法》（以下简称《办法》）。至此，中国率先交出一份答卷，也可以视为“生成式 AI 时代”下中国关于 AI 治理的独立思考。

《办法》彰显了中国对以生成式 AI 为代表的通用人工智能产业治理的基本态度[1]。与征求意见稿相比，《办法》不再以风险防控为主要导向，业内似乎可以打消此前的“踩刹车”顾虑。正式文件中大幅增加了促发展的内容，同时划定了底线。网信办在就《办法》答记者问时也提出，《办法》坚持目标导向和问题导向，明确了促进生成式 AI 技术发展的具体措施，规定了生成式 AI 服务的基本规范[2]。

1.1 生成式 AI 相关概念

2022 年被认为是“生成式 AI 元年”，Dall-E2、Midjourney、Stable Diffusion 等文生图应用的出现引起广泛关注；12 月底，ChatGPT 的横空出世更是使得生成式 AI 的风头一时无两。

自 1956 年概念的首次提出至今，人工智能技术已发展超 60 年。然而，时至今日，其仍未实现大规模应用。生成式 AI 的出现标志着人工智能进入了新纪元，机器开始能够模拟人类的创造性思维，并有望促进社会生产力的大幅跃进。

生成式 AI 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。

基于监督学习的方法差异，机器学习领域可以分为判别式（discriminative）和生成式（generative）两种典型模型：判别式模型是对条件概率建模，学习不同类别之间的最优边界，从而完成分类任务；生成式模型则面向类建立模型，计算基于类的联合概率，然后根据贝叶斯公式分别计算条件概率，进而根据输入数据预测类别。生成式 AI 更强调学习归纳后的演绎创造，通过模仿式、缝合式的生成创作，不断判别和进化，从而产生全新的内容，其本质是“创造未知世界”。

掀起本轮 AI 技术浪潮的正是后者。生成式 AI 技术以革新产业的面貌席卷了科技界，驱动了生产流程升级转型。

1 [钱玉娟，参与制定者解读生成式 AI 新规：“发展被放到更重要的位置”，经济观察报，https://www.8btc.com/article/6826106]

2 [国家互联网信息办公室有关负责人就《生成式人工智能服务管理暂行办法》答记者问，http://www.cac.gov.cn/2023-07/13/c_1690898326863363.htm。]

根据技术实现及应用路径，生成式 AI 又可以进一步细分为数据层、算力层、模型层和商业化应用层。

数据层包括数据提供、数据分析以及标注等环节。生成式 AI 产品的诞生需要依托海量的数据资源。公开资料显示，GPT-3 模型的基础来自 8000 亿个单词的语料库（或 45TB 的文本数据），包含了 1750 亿个参数。“8000 亿”是 ChatGPT 的训练数据，“1750 亿”是它从这些训练数据中所学习、沉淀下来的内容。

算力层是 AI 训练的基础设施，包括数据中心、服务器，以及高性能的 AI 芯片。据华泰证券测算，训练一次 ChatGPT 模型（13 亿参数）需要的算力约 27.5PFlop/s-day，即以 1 万亿次每秒的速度进行计算，需要耗时 27.5 天；而 ChatGPT 单月运营需要算力约 4874.4PFlop/s-day[3]。

模型层位于生成式 AI 的中游，是生成式 AI 得以实现的关键环节。生成式 AI 的成型得益于 2014 年以来 GAN（生成对抗网络模型）、Transformer（转换器模型）、Diffusion（扩散模型）等模型的发展与融合。例如，基于多种底层的 AI 技术，OpenAI 公司经过多次迭代的 GPT-4 模型，谷歌对标 GPT-4 的大模型 PaLM 2 等，通过开放 API 调用，可以赋能各类垂直应用厂商。

商业化应用层则涵盖文本、音频、图片、影片的生成等，是产业链的最下游，但也是 AI 能否大规模应用、能否真正创造价值的关键。

根据 PitchBook 的统计数据，2022 年投资圈向生成式 AI 公司共投入 13.7 亿美元（折合人民币约 93.69 亿元），融资事件发生 78 起，接近此前 5 年的总和。据风投公司 NFX 统计，目前已有 550 家生成式 AI 公司相继入局，共筹集近 140 亿美元的资金[4]。互联网及传统 AI 大厂将从云服务、预训练模型提供等的基础层、中间层入手，创业公司则聚焦在下游的垂直应用。多数公司刚刚完成标准化产品输出，开始进入商业化探索的初级阶段[5]。

1.2 生成式 AI 发展现状

在2023年过去的几个月里，算力基础设施看涨，各大公司抢滩大模型，类GPT商用化加速探索，整个产业链市场快速打开。

据东吴证券预计，AIGC 在内容生成中的渗透率将快速提升，应用规模迅疾扩增，预计 2030 年 AIGC 市场规模将超过万亿元人民币。

---

3 [郭美婷、蔡姝越：AI 契约论④：AIGC 产业链发展车驰船快，风险暗礁“涌现”，21 世纪经济报道，https://m.21jingji.com/article/20230531/herald/42b574c1fc7895ac1ab3246dc2789c16.html]

4 [https://www.nfx.com/post/generative-ai-tech-market-map]

5 [汉能投资，生成式 AI 开辟人工智能新纪元，AIGC 赛道风起云涌（上），https://www.36kr.com/p/2223151983822208]

1.2.1 算力

在 AI 大模型时代，AI 领域的“军备竞赛”正从过去算法和数据层面的竞争，转变为底层算力的竞争。

算力是大模型成本结构中最大的一块，GPU 的性能，决定了这个新兴行业的步调。但是，GPU 性能提升的速度，已经落后于大模型训练和推理需求的增长[6]。

GPU 是训练模型与加速推理的关键算力硬件。大模型拔高了对数据中心带宽、数据存储的门槛。云服务商会采购各类硬件，辅以冷却系统与运维服务，构建灵活、可扩展的 IaaS 平台，按需为客户提供算力。

据介绍，大约每隔 4 个月，AI 计算需求就会翻倍。广发证券分析师测算，随着国内生成对话式 AI 产品加速推广，在乐观假设下，国内 AI 大模型在训练与推理阶段或产生相当于 1.1 万台至 3.8 万台高端 AI 服务器的算力需求。以英伟达单片 A100 GPU 产品售价 10 万元、AI 加速卡价值量占服务器整机约 70% 计算，则对应 126 亿元至 434 亿元增量 AI 服务器市场规模[7]。

1.2.2 大模型

大模型已经成为本轮生成式 AI 竞赛的杀手锏。各个科技公司加码大模型，上演“百模大战”。

《中国人工智能大模型地图研究报告》显示，截至今年 5 月 28 日，中国 10 亿参数规模以上的大模型已发布 79 个。而美国和中国就占全球已发布大模型总量的 80% 以上[8]。

谷歌基于 PaLM 2 推出聊天机器人巴德（Bard），据介绍 Bard 支持 40 多种语言，同时还拥有图像分析功能；微软日前宣布将 GPT-4 导入全新 Bing 搜索引擎和 Microsoft 365 Copilot，亚马逊也通过发布泰坦（Titan）以宣布加入战局。

中国大模型不断涌现，既有实力雄厚的互联网平台企业：百度、阿里、华为等互联网公司发布 “文心一言”、“通义千问”及华为盘古大模型等；也有人工智能新秀，比如澜舟科技的孟子 GPT、智谱 AI 的 ChatGLM、科大讯飞的星火大模型等。

一批高校、科研院所也相继入局，清华大学发布大模型 GLM-130B、复旦大学则发布大语言模型 MOSS；上海人工智能实验室发布天气预报大模型“风乌”、北京智源人工智能研究院发布“悟道 3.0”等。

---

6 [未尽研究、启明创投：《2023 生成式 AI 报告》，https://www.qimingvc.com/sites/default/files/State_of_Generative_AI_2023.pdf]

7 [彭思雨，AI 大模型迎来风口 算力需求爆发，中国证券报，http://ex.chinadaily.com.cn/exchange/partners/82/rss/channel/cn/columns/sz8srm/stories/WS6487bb87a310dbde06d230f3.html]

8 [科技部发布《中国人工智能大模型地图研究报告》，http://lib.ia.ac.cn/news/newsdetail/68630]

大模型整理

	公司名称	大模型	发布时间	应用场景
互联网公司	华为	盘古大模型	2021年4月 首次公布	物流监控、药物研发、气象预测、铁路巡检、矿山、金融、政务等
	达摩院	八卦炉	2022年	
	腾讯	混元大模型	2022年4月 首次对外披露，尚未正式发布	支持腾讯众多产品和业务，以及广告营销、金融风控、交互翻译、数智人客服等场景
	阿里	通义大模型	2022年9月2日	办公学习助手、电商跨模态搜索、开放域人机对话、AI辅助设计、法律文书学习、医疗文本理解等
	百度	文心大模型	2023年3月16日	办公助手、商业营销、社交文娱、金融、能源、制造、传媒、城市等
	字节跳动	飞书"My AI"	2023年4月11日	应用于飞书APP，主要功能包括总结会议纪要、分析经营数据、文本续写等
	知乎 (与面壁科技合作)	知海图AI	2023年4月13日	问答社区场景，如热榜摘要、搜索聚合、面壁露卡
	昆仑万维	天工大模型	2023年4月17日	智能问答、聊天互动、文本生成、AI作曲、游戏设计等
	麒麟合盛	天燕大模型 AiLMe	2023年4月18日	文本交互、图像生成、视频创作、语音识别与合成等场景
	360	360智脑	2023年5月18日	办公助手、AI搜索、数字员工、智能安防、税务、企业服务等场景
	京东	言犀大模型	2023年7月13日	在线咨询机器人、智能语音外呼、智能语音应答等，暂时应用于京东内部业务
	网易	玉言	尚未发布	语音助手、智能客服、互联网搜索、游戏创作、新闻传媒等领域
	蚂蚁	贞仪大模型	正在研发	
AI公司	MiniMax	自研大模型	2021年12月	办公助手、游戏创作与陪玩、广告营销、教育助手、代码生成等
	竹间智能	大模型训练调优平台EmotiBrain	2022年6月18日	可应用于办公、营销等场景，为金融、通信、制造、互联网等企业提供大模型服务
	澜舟科技	孟子GPT	2023年3月14日	银行客服、智能投研、行业搜索、AI写作与绘画、智能营销等
	智谱AI	ChatGLM	2023年3月28日 公布ChatGLM-6B	应用于代码生成、内容创作（文生图）、智能问答等场景
	商汤科技	日日新大模型 SenseNova	2023年4月10日	代码开发、智慧医疗、商业营销、AI数字人、智能汽车等，目前已在金融、医疗、电网巡检、实景建设、珠宝等行业落地
	毫末智行	自动驾驶生成式大模型DriveGPT（雪湖·海若）	2023年4月11日	DriveGPT运用在车端后，可实现城市NOH、捷径推荐、智能陪练、场景脱困等场景功能。目前的合作伙伴有：长城旗下新能源品牌汽车"欧拉""摩卡"等
	出门问问	序列猴子	2023年4月20日	办公助手、商业营销、社交文娱、金融、能源、制造、传媒、城市等

	公司名称	大模型	发布时间	应用场景
	第四范式	式说	2023年4月26日	企业级办公助手，可实现AI时代数据资源治理、大规模分布式资源管理与调度、应用统一管理及应用快速集成
	科大讯飞	星火大模型	2023年5月6日	办公助手、代码开发、广告营销、交互客服、企业管理、数字员工、工业设备检测和管理等
科研院所	清华大学	GLM-130B	2022年7月	
	复旦大学 邱锡鹏教授团队	对话式大语言模型MOSS	2023年2月20日	MOSS对话系统，具备搜索引擎、文生图、计算器、方程求解等功能
	上海人工智能实验室	天气预报大模型"风乌"	2023年4月7日	主要应用于气象预测领域
	北京智源人工智能研究院	悟道3.0	2023年6月9日	流畅的文本对话、多种语言生成、文图生成、多步可控人脸编辑、代码写作
	武汉人工智能研究院、中国科学院自动化研究所和华为联合研发	多模态大模型"紫东太初"	2023年6月16日 发布紫东太初2.0	手语教学、法律咨询、交通出行、医疗机器人、医学影像判读、短视频内容审核、交通违规图像研读、音乐理解和生成等行业场景领域
	达观数据	曹植	2023年7月7日	通用文本写作、企业申报材料自动生成、金融报告AIGC智能写作，可应用于金融、银行、证券、保险、政府、能源、医药等多领域
	鹏城实验室	鹏城·脑海	正在研发	
教育行业	网易有道	子曰	2023年5月5日	AI虚拟人口语教练、中文作文批改DEMO、有道AI学习机等
	学而思	MathGPT	尚未发布 预计在年内推出	以数学领域的解题和讲题算法为核心
车企	理想汽车	MindGPT	2023年6月19日	具备智能汽车场景，支持声纹识别、内容识别、方言识别、出行规划，AI绘画、AI计算等功能，将应用于车载AI助手"理想同学"
	蔚来、小鹏、长城汽车、奇瑞新能源等车企已在申请GPT商标 未来将应用于智能驾驶场景			

在教育和新能源汽车行业，一些企业选择布局与原有业务相适应的大模型。如网易有道为教育场景自研的类 ChatGPT 模型“子曰”，学而思的自研数学大模型 MathGPT 也预计在年内推出。新能源车企如理想汽车已经发布了自研 MindGPT，将应用于车载 AI 助手“理想同学”，而蔚来、小鹏等车企也已在申请 GPT 商标。在医疗领域，上海联通、华山医院联合开发的 Uni-talk、医联“MedGPT”、云知声的“山海”等也相继登场。

处于核心的模型层，目前可分为通用大模型和垂直大模型。通用大模型能够处理多种任务和应用于不同领域，是资金、资源雄厚的科技巨头优选；行业大模型则是针对特定领域或任务进行优化设计的模型，基于自建模型或利用通用大模型，引入行业语料进行模型深度训练，以提升对特定行业 / 领域应用场景的支撑能力。

由于大模型在资金、算力、语料训练集等方面存在较高门槛，垂直大模型以其成本低，部署升级灵活的优势成为新赛道；不过，垂直大模型需要专门的、行业深度训练的数据以更贴合业务；还需更好地与企业内部知识库进行配合，才能做到实时迭代更新。

1.2.3 生成式 AI 应用市场

大模型持续火热，业内更关心应用落地。只有让大模型与千行百业的具体业务场景结合，才能产生具体应用价值。目前，大模型技术比较热门的落地领域包括办公软件、社交文娱、商业营销、家庭助理和金融等。

这些领域内，少部分公司选择自主部署研发模型，更多的公司则选择接入较为成熟的大模型（类 ChatGPT 产品），以直接赋能其原有产品和服务。例如，在办公领域，微软 Microsoft 365、字节飞书“My AI”、金山 WPS 等均宣布已接入大模型。社交文娱方向，出现了 AI 搜索引擎如微软必应、谷歌 Magi，还有应用于游戏影视的英伟达 AI 智能游戏助手 GeForceRTXR.O.N.、Adobe 的 Premiere Pro 等。在家庭场景，AI 也充当起家庭管家、私人家教、智能汽车助手，如阿里就率先将 AI 大模型接入了智能音箱天猫精灵。

在商业营销方向，多种类 ChatGPT 产品涵盖智能客服、推荐算法、虚拟人直播、广告策划等具体应用场景。类 ChatGPT 产品正在逐步渗透到生产和生活的各个环节。

可以看出，第一类场景为提升生产工作效率的通用工具，通过生成式 AI 提升内容供给速度、降低内容创作门槛，从而使得人工资源能够更多地投入到高价值的工作及创作流程中，提升整体工作效率。第二类场景则是可能改变行业格局的场景应用，比如有场景的 C 端，有数据的 B 端，带来一个增量产业的崛起（如教育、医疗分诊，个性化生成，高频时效交互）等[9]。

1.3 关于生成式 AI 治理的中国思考

当欧盟正试图通过专门的《人工智能法案》来展现他们对于生成式 AI 基于风险的治理思路时，中国在 7 月发布的《办法》则体现其对于生成式 AI 治理的不同思考。“发展”正逐渐成为中国 AI 治理的第一视角。

与此前的征求意见稿相比，《办法》有较大的思路调整，“坚持目标导向和问题导向”，单设了“技术发展与治理”章节，同时也新增了不少有力措施来鼓励生成式 AI 技术发展。

其背后所反映的，正是中国对于目前生成式 AI 发展、治理的独立思考。

---

9 [招商证券，2023 年 AIGC 产业链投资机会分析 OpenAI 引领 AI 产业变革，https://www.vzkoo.com/read/2023052697991ce0aea8ebff5d0d6d6f.html]

接入类ChatGPT技术应用

应用类别	细分方向	应用名称
办公软件	协同办公平台（综合）	微软Microsoft 365 Copilot、阿里A版钉钉、字节旗下飞书“My AI”、百度旗下“如流”、金山办公WPS AI、谷歌Google Wordspace等
	AI办公工具（细分）	ChatGPT、福昕软件、上上签“哈勃”、Notion AI、GrammarlyGo、阿里通义听悟等
	AI编程工具	微软Copilot X、亚马逊Code Whisperer、华为云Code Arts Snap、硅心科技aixcoderXL、智谱AI CodeGeeX等
社交文娱	搜索引擎	微软新版Bing、谷歌Magi、360AI浏览器、360AI搜索等
	社交媒体	Snapchat旗下My AI、机械佛Hotoke AI、弥知科技KiviGPT、斯坦福团队成果rizzGPT、Minimax对话机器人生成平台Glow等
	游戏影视	Discord社区的Clyde、Adobe旗下Premiere Pro等、计算美学AI设计创意平台Yeahpix、360鸿图、英伟达AI智能游戏助手GeForceRTXR.O.N.等
商业营销	智能客服	Shopify的Shopping Assistant等
	推荐算法	Intacart的食物搜索工具、Expedia旅游 推荐机器人等
	虚拟人直播	Synthesis AI数字人方案、天娱数科的虚拟数字人“天妤”、世优科技数字人、来画智能生成平台等
	客户管理	Salesforcet的Einstein GPT 、Salesforce旗下Slack等
	广告策划	蓝标BlueFocus AIGC矩阵、有赞“加我智能”平台等
	服务机器人	穿山甲机器人“Tiamo小鱼”、猎户星空AI机器人、彭博社聊天机器人BloombergGPT等
家庭助理	智能音箱助手	天猫精灵“鸟鸟分鸟”模型等
	智能教育助手	Quizlet下Q-Chat、多邻国Roleplay、可汗实验学校Khanmigo、学而思MathGPT、网易有道AI口语老师、科大讯飞“大模型+Ai学习机”等
	智能汽车助手	通用汽车虚拟汽车助手、百度文心一言、阿里通义千问、商汤“日日新SenseNova”落地汽车“绝影”、理想汽车车载AI助手“理想同学”等
智慧医疗	医疗助手	微软AI临床笔记软件Dragon Ambiente Xperience（DAX™）Express

1.3.1 探索优化应用场景

人工智能技术已发展超 60 年，时至今日仍难言大规模应用。《办法》第五条明确，鼓励生成式 AI 技术在各行业、各领域的创新应用，生成积极健康、向上向善的优质内容，探索优化应用场景，构建应用生态体系。

商业落地是国内人工智能发展面临的困境之一，大模型发展声势浩大，但只有做到商业化、工程化、应用场景化，才能真正赋能产业。

近期，创业者服务平台 GoDaddy 对全美 1003 家小型企业的调查数据显示，ChatGPT 以 70% 的应用率成为美国小型企业应用最多的生成式 AI 产品；38% 的受访者，在过去几个月里尝试过生成式 AI；营销、内容创作、商业建议是企业应用生成式 AI 最多的 3 个用例；75% 受访者非常满意生成式 AI 在业务中的表现 [10]。

对比之下，国内大模型远没有达到可商用化的程度，或是能深度切入具体应用场景。目前大模型落地主要以价值增强和效率提升为主，而商业模式层面的落地仍在探索中。

大模型能否和业务充分结合，从而真正解决业务问题，是决定 AI 能否实现经济价值的关键因素。只有紧贴业务的 AI 战略设计、完善的配套架构、充足的 AI 人才及健全的内部培养机制，才能使 AI 与业务发展需求充分融合，最大化实现经济收益。

各地的人工智能相关政策也聚焦到应用层。5 月发布的《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）》提到，要发挥各区产业特色和资源优势，结合人工智能技术特点，围绕经济社会发展、科学研究发现、重大民生需求等，形成一批示范性强、影响力大、带动性广的重大应用场景。7 月发布的《上海市推动人工智能大模型创新发展的若干措施》中提及示范应用推进计划，要加强大模型在智能制造、教育教学、科技金融、设计创意、科学智能等垂直领域的深度应用和标杆场景打造。

1.3.2 加强基础技术的自主创新

人工智能需要 GPU 算力、网络及存储等硬件基础设施的全方位支撑。《办法》指出，鼓励生成式 AI 算法、框架、芯片及配套软件平台等基础技术的自主创新，鼓励平等互利开展国际交流与合作，参与生成式 AI 相关国际规则制定。

在中美博弈大背景下，A100、H100 为代表的芯片被美国列入禁止出口商品名单，加之国内信创进展和适配需要时间，算力问题成为中国人工智能发展的掣肘因素。

国内过去在互联网及移动互联网时代所积累的云计算、数据中心、算力中心等领先全球的数字化基础设施，为生成式 AI 落地运行提供了坚实基础。4 月，科技部启动国家超算互联网部署工作。根据部署，国家超算互联网通过算力网络将全国众多超算中心连接起来，用互联网思维运营超算中心，构建一体化算力服务平台，实现算力资源统筹调度，降低超算应用门槛，带动计算技术向更高水平发展。

10 [ChatGPT 等生成式 AI，对小型企业帮助大，https://www.sohu.com/a/682263940_121649381]

《办法》强调，平等互利开展国际交流与合作，参与生成式 AI 相关国际规则制定。

这在一定程度上也表明，商业的密切交集使得各国愈发成为共同体；人工智能的全球竞争已经从技术扩展到治理领域。积极参与国际规则的制定具有非常重要的意义。

1.3.3 推动公共训练数据资源平台建设

生成式 AI 的训练和研发需要大量的基础设施和基础资源支持。《办法》提出，推动生成式 AI 基础设施和公共训练数据资源平台建设，促进算力资源协同共享，提升算力资源利用效能。

平台的建设有助于更好地协调和优化训练资源，更好地集中精力完成技术层面的攻关和突破。

推动公共数据有序开放，扩展高质量的训练数据资源也是必须要划出的重点。“大模型时代，得数据者得天下。”一方面，训练数据是大模型训练的基石和燃料，如果没有数据，大模型的训练就无法开展和持续；另一方面，当前技术领域的研究显示，各家大模型在算法层区别并不大，并且具有同质化的趋势。在此背景下，训练数据就成了真正区分且影响大模型性能的重要因素之一[11]。

从各地实践也可以看出，加码训练高质量数据集建设已成为重要方向。

北京 5 月印发的《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）》中就提到，加强公共数据开放共享，包括动态更新公共数据开放计划，加快构建高质量人工智能训练数据集等。6 月，深圳发布的《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）》中也提出，要搭建全市公共数据开放运营平台，建立多模态公共数据集，打造高质量中文语料数据等。

目前各地出台了不少关于公共数据开放利用的条例，利用公共数据投喂人工智能，应按照有条件开放、无条件开放或禁止开放的不同方式进行。不过，公共数据开放存在较多阻力，开放的数据范围和质量不够。接下来需推动有序开放，亟待分类分级，发挥公共数据红利，探索契合公共数据价值利用规律的开放之道[12]。

1.3.4 豁免责任

此前，对于技术研发阶段是否适用监管等问题有多种声音。此次《办法》对“研发生成式 AI 技术”进行了豁免。

本《办法》适用范围为：“利用生成式 AI 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”。值得注意的是，行业组织、企业、教育和科研机构、公共文化机构、

11 [王俊、冯恋阁，七部门为生成式 AI 立规，鼓励自主创新、分类分级监管，21 世纪经济报道，https://www.21jingji.com/article/20230713/herald/96c663f99a58d33aa729df871c1d90f2.html]

12 [张雅婷，《生成式人工智能服务管理暂行办法》出台，公共数据将如何赋能大模型？21 世纪经济报道，https://www.21jingji.com/article/20230713/herald/1dd46baad746e6be4beec8165a99174e.html]

有关专业机构等研发、应用生成式 AI 技术，未向境内公众提供生成式 AI 服务的，不适用本《办法》的规定。

这意味着，一方面，当研发主体（行业组织、教育科研机构、企业等）向特定公众提供小范围试用版本，或仅仅是在研发过程中推出测试版本，这些都是属于未正式上市的 AI 服务，不适用《办法》。

另一方面，如果生成式 AI 服务到了应用阶段，不向境内公众提供服务，也不需要遵守《办法》。例如不向消费者直接提供服务，仅向公司提供服务，就属于这种情形。

这也体现了中国对于技术研发等工作的态度。

第二部分

风险：生成式 AI 治理的底线思维

作为一项新兴技术，生成式 AI 技术在应用过程中带来的问题已逐渐显现。本白皮书结合生成式 AI 技术的应用场景，对基于生成式 AI 技术开发的模型或产品可能面临的风险进行了筛查，主要可能涉及知识产权侵权、算法风险、数据安全与个人信息保护风险及其他风险。针对这些风险点，本白皮书也建议相关开发者应当搭建生成式 AI 合规风险事件库，以及时了解、追踪生成式 AI 合规风险导向。

2.1 知识产权侵权风险

2.1.1 著作权侵权

基于大模型对语料丰富度的较高需求，无论是模型输入端，还是模型输出端，均存在较大的侵犯著作权的可能性与风险。

在输入端，大模型在构建过程中需要“学习”大量文本语料，而在获取文本语料时，可能会因未征得权利人许可复制、改变或者传播而涉嫌侵犯他人著作权。例如，某全球知名图片提供商起诉了某 AI 绘画工具的开发者，称其未经许可从自己的网站上窃取了数百万张图片。需要注意的是，将生成式 AI 模型应用于商业用途本身已经明显超出知识产权法律中界定的“合理使用”的范畴。

在输出端，判断著作权侵权的第一步是看生成内容能否被判定为著作权法中所定义的“作品”。

根据以往判例，法院在对“作品”进行判断时，曾认定“作品的创作主体应限定为自然人”；同时，是否满足“独创性”也是判定是否构成作品的标准。不过，就算生成内容无法达到作品的标准，与原作品构成实质性近似，仍可能构成知识产权侵权。此外，大模型训练的文本语料数量越是匮乏，生成内容的知识产权侵权风险就会越高。生成内容还将受到使用者提问方式的影响，如使用者的提示词极为限缩、精准，也将增大生成内容侵犯他人知识产权的风险。

2.1.2 侵犯商业秘密

训练数据是修炼大模型的“原材料”，在大模型的静态与动态训练的过程中，可能会出现使用来源不明或者非法的数据信息的情况，如果前述信息中包含商业秘密，那么依据竞争法下的相关规定，将构成对他人商业秘密的侵害。

同时，随着大模型影响力的扩大，企业可能会将其纳入办公系统以提升工作效率。而企业员工在训练和使用模型的过程中，如不慎输入企业的商业秘密，不仅可能造成公司商业秘密的直接泄露，甚至可能被模型存储于数据库中作为训练数据，如遭受黑客攻击，公司将进一步蒙受损失。例如某头部互联网公司日前声称，在某生成式AI模型所生成的内容中发现了与其商业机密非常相似的文本，可能是其内部员工在使用该模型生成代码时输入了公司的机密数据。从企业防止商业秘密外泄的角度，如何约束员工在利用大模型提高工作效率的同时，又能避免对企业商业秘密的侵犯，也将成为相关法律合规部门需要思考的问题。

2.1.3 不正当竞争和反垄断风险

大模型在协助编程、广告设计、文学创作等领域表现优异。用户在使用大模型生成广告文案时，其生成内容可能与他人广告文案、知名商品名称、知名企业名称等存在相似。考虑到这类文案、名称等通常篇幅简短，较难被视作著作权法上的“作品”，此时，如果用户直接将生成内容进行商业使用，可能构成竞争法中的“混淆行为”。

此外，由于多数企业将公开爬取作为获取数据的渠道之一，如果使用非法爬取的数据进行生成式AI模型训练，形成的数据产品若达到足以实质性替代其他经营者提供的相关产品或者服务的程度，其仍可能构成不正当竞争。

同时，虽然互联网上存在大量可以自由访问的信息，但如果从具有禁止第三方爬取数据条款的网站收集海量数据，该等数据很有可能被认定为竞争性财产权益，因此如何避免对该等数据的收集违反竞争法的相关规定，也将成为相关开发者和服务提供者所要面临的一大挑战。

另一方面，大模型也可能引发竞争法下反垄断相关风险。其一，是基于技术的高尖性，由于目前生成式AI技术主要掌握在全球大型科技公司手中，如何防止生成式AI技术的研发与应用成为大

公司新型的垄断手段？就该问题，欧美地区也正在筹划制定专门针对生成式 AI 的反垄断法[13]。其二，部分企业之间试图通过人工智能达成“垄断协议”，也即价格算法合谋，例如，某公司通过某生成式人工智能模型对消费者数据进行消费习惯分析，从而对消费者采取有针对性的算法垄断定价。如何防止这种新型垄断现象对消费者权益可能带来的损害，也将成为需要思考的问题。

2.2 算法风险

2.2.1 算法黑箱风险

由于大模型的算法内部机制和决策过程不可解释或难以理解，会导致算法的输出结果存在无法解释的“黑洞”。全球最大的生成式 AI 模型 ChatGPT 也因至今未曾公布算法规则而饱受诟病。

算法的输出结果不仅仅取决于输入数据，还会受到算法自身的运行过程、模型参数、超参数等多方面因素的影响。这种风险可能会导致企业难以理解模型的决策过程和预测结果，从而难以评估模型的可靠性和稳定性。另一方面，企业也难以响应用户关于解释算法结论的要求。

例如，某企业在使用生成式 AI 模型生成推荐内容时，发现该模型频繁生成与特定人群利益相关的内容，但无法确定这种情况是因为算法本身存在偏见还是数据集本身就存在偏见，此时企业可能难以发现问题的根源和解决方案。

2.2.2 算法歧视风险

大模型算法在应用、决策或预测时，如其本身的算法规则对某些特定的个体或群体存在偏见，将会导致企业的商业决策不公正，进而影响用户对其的信任和企业商誉与形象。对用户而言，其将遭受歧视和不公正的待遇。

例如，某企业使用生成式 AI 模型为客户提供智能客服服务，但由于该模型算法规则本身存在种族、性别等方面的偏见，导致部分群体的问题无法得到利于其立场或身份的解答，进而影响客户对该企业的认知与评价。

2.2.3 算法决策风险

在使用大模型进行决策时，由于模型的不确定性或数据量欠缺等原因，可能会导致错误的决策或不良后果。同时，由于算法决策形态本身的多元性，算法决策机制的隐蔽性，以及算法决策主体的模糊性，都有可能给算法决策带来风险。

例如，某医疗保险公司使用生成式 AI 模型来决定是否批准某个人的理赔申请。如果该模型的

---

13 [AIGC 反垄断法或加剧算力军备竞赛 行业大玩家谁将获益，东方网，http://caijing.chinadaily.com.cn/a/202307/06/WS64a65520a310ba94c56151a4.html。]

算法规则中没有纳入部分潜在的判定因素，例如此人的医疗病史、病情危重程度等，就可能会错误地拒绝该人的申请。这种错误决策可能会导致患者的疾病无法得到及时治疗，甚至危及生命，给公司和患者带来不良后果。

2.2.4 信息失真风险

信息失真风险并非算法的固有风险。当算法所处理的数据本身存在错误时，算法的处理结果就会出现信息失真；此外，如果企业或者用户出于某种目的故意规避对算法和数据的监管，也有概率导致训练出的模型生成违法、欺诈、诽谤、侵犯隐私等类型的内容。这种风险可能导致用户通过算法模型得到错误的结论，产生负面影响；企业可能因为错误的或违法的内容受到监管机构处罚，影响企业声誉和业务发展。

特别是对于拟自行部署生成式 AI 模型的开发者而言，如模型算法本身的语言推理能力有限，造成信息失真甚至“臆想”现象将愈发严重。

例如，某用户在使用某企业提供的生成式 AI 模型时，由于数据有误，导致其在商业决策中决策失误，并由此亏损，其认为该企业提供的模型应当承担相应责任。

2.3 数据安全与个人信息保护风险

2.3.1 个人信息保护

（1）个人信息收集场景

大数据时代，生成式 AI 模型难以规避因收集个人信息所带来的风险，这类风险不仅可能发生在模型的训练阶段，也可能发生在模型的实际应用阶段。

在模型的训练阶段，大模型往往需要获取多元化、丰富的语料进行训练，在这个过程中，难免会采取爬虫等技术方式通过互联网等公开渠道获取大量数据，如何避免因爬取或其他手段获取公开渠道的个人信息而构成侵权等法律风险？如涉及从第三方获取的数据的，如何审核个人信息来源的合法性和个人的授权情况？这都是应当思考的问题。例如，某生成式 AI 模型在训练的过程中爬取了某点评网站上关于某餐饮店的评价，但由于某用户在点评时透露了自身的个人信息，导致该部分个人信息进入到模型语料库，进而涉嫌侵犯他人个人信息权益。

在模型的实际应用阶段，如何精准识别 AI 与用户交互过程中所收集的个人信息，并进而履行个人信息保护相关的合规义务，也将成为生成式 AI 模型应用者所要面临的一大挑战。不同于一般应用程序中填入式的收集个人信息方式，大模型由于涉及人与 AI 的交互，很难在事前对可能收集个人信息的场景进行完全罗列，而更近似于“客服热线”的场景，在此背景下，应当如何在事前向

个人告知收集个人信息的目的、方式和范围并取得其同意，也是值得研究的问题。

（2）个人信息使用场景

在对个人信息的使用上，目前，部分生成式 AI 产品以改善服务为由使用用户提供的内容（其中包含个人信息），但显然，仅以改善服务为由要求收集用户信息并不符合最小必要原则，本质上是对于“企业训练模型之需”与“用户享受服务之需”的混淆。目前，OpenAI 已提供用户拒绝其使用个人信息进行训练的途径。

除此之外，在大模型的交互模式下，对于个人信息的披露可能不同于往常意义上的“公开披露”，而更类似于一种“被动公开”，即当某个用户的真实个人信息被摘录于语料库后，之后任意用户通过询问等方式均可以得知相关个人信息，此时由于对象为非特定自然人，相较于向特定个人“提供个人信息”，可能更接近于“公开个人信息”的范畴。因此，对于模型开发者而言，应当慎重考虑在语料库以及训练模型的过程中是否加入真实个人信息。例如，某直销机构需要向客户公开披露直销员的联系方式等，但由于并未告知直销员，导致直销员的联系方式被其他人通过与 AI 的问答获取，并用于其他目的，此时企业可能会因为未事前披露使用目的而涉嫌侵犯他人个人信息权益。

（3）个人信息权利响应场景

在大模型下，关于个人信息权利响应的实现似乎远远没有想象中来得容易。例如，就查阅权和更正权而言，提供者应当如何确定个人信息的范围并提供给用户查阅或更正？如前所述，模型通常存储的是交互记录，而不会在识别个人数据后将其作为单独的存储单位。就删除权而言，如果这部分数据已被用于模型训练，此时，从技术上而言难以做到完全删除，仅能通过过滤数据或者重新训练的方式以最小化这部分个人信息对模型输出可能产生的影响。

同时，如果大模型技术提供方位于境外，收集的个人信息将通过 API 接口传输至位于境外的主体，如何向个人告知向境外行使个人信息权利的途径，也将成为服务提供者需要面临的现实问题。

（4）儿童个人信息处理场景

在训练大模型的过程中，服务提供者需要基于自身的目的，考量是否有收集儿童个人信息的必要性。

如果业务本身并不面向或针对儿童，但如遇到医疗健康事件等小概率事件下可能会收集儿童个人信息，也应当在隐私政策等个人信息声明中告知并获得有效同意。

如不存在收集任何儿童个人信息的必要性，则应从技术和制度角度防止误收儿童个人信息。例如，某 10 岁的儿童通过网站的广告页面进入了某生成式 AI 模型服务提供页面，并输入了自身的姓名等个人信息，此时，由于系统无法准确识别使用者的年龄，在无形中收集了该名儿童的个人信息。

目前，如 OpenAI 也已经关注到此类问题，但可能出于对现有技术判别年龄的有效性等考量，其并未采取进一步动作。

2.3.2 数据跨境风险

目前，除少部分自行开发、部署模型的服务提供者提供以外，大部分服务提供者仍需倚赖第三方技术服务商搭建模型或以接入 API 等方式使用生成式 AI 服务，而这些技术方的服务器一般部署于境外。例如，一家位于中国大陆的企业，通过 API 接口的方式接入位于北美的生成式 AI 技术服务提供商，而该服务商的服务器部署于印度，此时可能面临相关数据出境所带来的风险。

除此之外，在提供生成式 AI 服务过程中，不仅涉及数据出境问题，还可能涉及数据入境。例如，经过境外模型处理后产生的数据通过 AI 交互方式返回给中国用户时，也需考虑境外国家关于数据出境的合规要求和限制。

从境内外关于生成式 AI 技术的法律规制来看，目前，服务提供者在应用生成式 AI 模型的过程中，可能会面临如下与数据跨境相关的风险与挑战：

（1）大陆地区尚未被列入核心技术供应商开放服务范围之内

目前，如 OpenAI 等核心生成式 AI 技术提供方并未将中国大陆地区列入其服务提供对象范围，在此背景下，如果因为使用相关服务给大陆企业造成了损害后果（如数据泄漏等），企业的权利应当如何得到保障？

此外，部分企业通过自行建立或租用专线（含虚拟专用网络 VPN）的方式，连接到境外的生成式 AI 技术模型，这一做法如未经电信主管部门批准，则涉嫌违反工信部《关于清理规范互联网网络接入服务市场的通知》的规定，违规风险极大，尤其是当企业以营利为目的专门向其他企业提供此类服务的，情节严重的情况下，还可能构成非法经营罪，将会面临刑事风险。

（2）涉及数据出境情况存在不确定性

根据《网络安全法》《个人信息保护法》《数据出境安全评估办法》等法律法规规定，在进行数据出境前应当履行相应的出境合规义务，例如进行事前安全评估，如涉及个人信息的，还应当进行个人信息保护认证、签署标准合同等，而由于企业使用生成式 AI 模型服务中涉及的数据出境情况存在不确定性，因此，在履行出境合规义务上可能存在滞后性，例如，事先企业无法确定使用生成式 AI 模型所涉及的数据是否在数据出境合规监管范围之内，又或者，由于无法确定具体的数据出境路径而无法事前完成相应申报或备案。

目前，生成式 AI 模型的应用路径主要分为两种，其一，是企业将其收集的数据提供给生成式 AI 模型，并向其提出数据处理请求，该情形下，企业在提供数据前可以结合业务需求及合规义务要

求对数据进行预处理、筛选，以限定其所提供的数据范围，从而使得相关数据出境风险处于可控范围之内。其二，是用户直接使用企业接入生成式 AI 模型的服务平台，这种情况下，由于用户是直接向生成式 AI 模型提供数据的主体，对于企业而言，用户提供哪些数据存在很大不确定性，从而触发相关的数据出境合规义务要求。而且，由于该情形下向境外提供个人信息量级的不确定性，将导致企业难以确定数据出境合规路径。

（3）个人信息出境时如何告知并取得个人或其监护人的单独同意

根据《个人信息保护法》的要求，基于个人同意向境外提供个人信息的，应当取得个人的单独同意。从目前行业实践来看，企业通常通过弹窗勾选的方式取得数据出境相关的单独同意，但是在告知方面，基于生成式 AI 模型的特殊性，如何在事前告知可能出境的个人信息范围，将会成为企业在履行相关告知义务时不得不面临的现实困境。

（4）如何满足境外出口管制的要求

经境外的生成式 AI 模型处理后产生的数据面临相关国家、地区的监管。不同于欧盟的严格限制，例如美国对于数据跨境流动的监管要宽松许多，对于数据跨境流动，美国偏向于限制政府权力，其跨境数据隐私保护策略主要靠有限的立法和行业自律。

如前所述，经境外的生成式 AI 模型处理后产生的数据或产品如返回给中国用户，需要考虑境外国家关于数据出境的合规要求和限制。例如，在 OpenAI 公布的用户协议中，亦明确规定其提供的服务受美国出口管制相关规则所约束，不得出口到任何美国禁运国家。同时，根据美国商务部工业与安全局官网公布的禁运规则解释文本及其对应表格，中国属于清单中的 D 组禁运国家，主要禁运领域为“国家安全、核工业、化学与生物领域、导弹科技和其他军事领域”，虽非绝对禁运，但是，对于电子产品、电子通信、电脑、信息安全等领域相关的设备、组件、材料、软件及其相关科技向中国的出口，仍需要向美国相关部门进行报批。值得注意的是，出口管制适用的范围不仅包括实体货物，还包括计算机软件、数据、技术等，这就意味着受《美国出口管制条例》等法律法规约束的货物相关的数据亦需遵循相应的规制。因此，对于经生成式 AI 模型处理后返回的数据，是否符合境外出口管制方面的合规要求，亦是企业在将来迎接监管时可能面临的挑战。

2.3.3 数据安全风险

（1）数据泄露

基于大模型强大的系统学习能力，以及交互方式上相对的自由与不受局限，导致交互过程中存在较大泄漏数据或商业秘密的风险。例如，近日，某韩国头部企业发生了三起生成式 AI 产品的误用与滥用案例，包括设备信息泄漏和会议内容泄漏。其中，涉及半导体设备测量资料、产品良率等

内容或已被存入 AI 模型的学习资料库中[14]。为规避数据泄露风险，此前已有不少企业明确禁止员工使用生成式 AI 模型，包括摩根大通、德意志银行、埃森哲、富士通、软银、高盛、花旗等公司。

（2）网络安全

大模型的“滥用风险”似乎是伴随着大模型兴起一直为人们紧密关注的议题，对于生成式 AI 模型的开发者而言，在训练模型的过程中需要思考如何有效预防用户的恶意使用行为，例如利用模型撰写编码或病毒，用于攻击网站等，从而引发网络安全风险。

同时，大模型也似乎更容易受到“提示语注入攻击”的影响，如何防止恶意使用者诱骗模型突破限制使用范围，也将成为生成式 AI 模型应用过程中可能面临的一大挑战。

（3）重要数据处理

基于大语言模型的特性，其模型的训练和完善都有赖于大量数据的帮助，不同数据类型、数据量级的叠加，不排除构成重要数据乃至核心数据从而触发相应的合规义务。同时，我国目前《数据安全法》等法律法规中虽然提出了“重要数据”“核心数据”的概念，但二者的判定依据尚未明晰，这种不确定性也使得生成式 AI 模型应用面临的合规风险增加。比如某企业通过某生成式 AI 模型堆砌数据，但当数据量增加到某一量级时已构成重要数据，此时如本身模型技术支持位于境外，企业将会面临重要数据出境相关的合规风险。

2.4 伦理道德风险

生成式 AI 技术带来的伦理道德风险也值得深思。例如近日欧洲议会通过的《人工智能法案》的谈判授权草案，也严格禁止“对人类安全造成不可接受风险的人工智能系统”[15]。

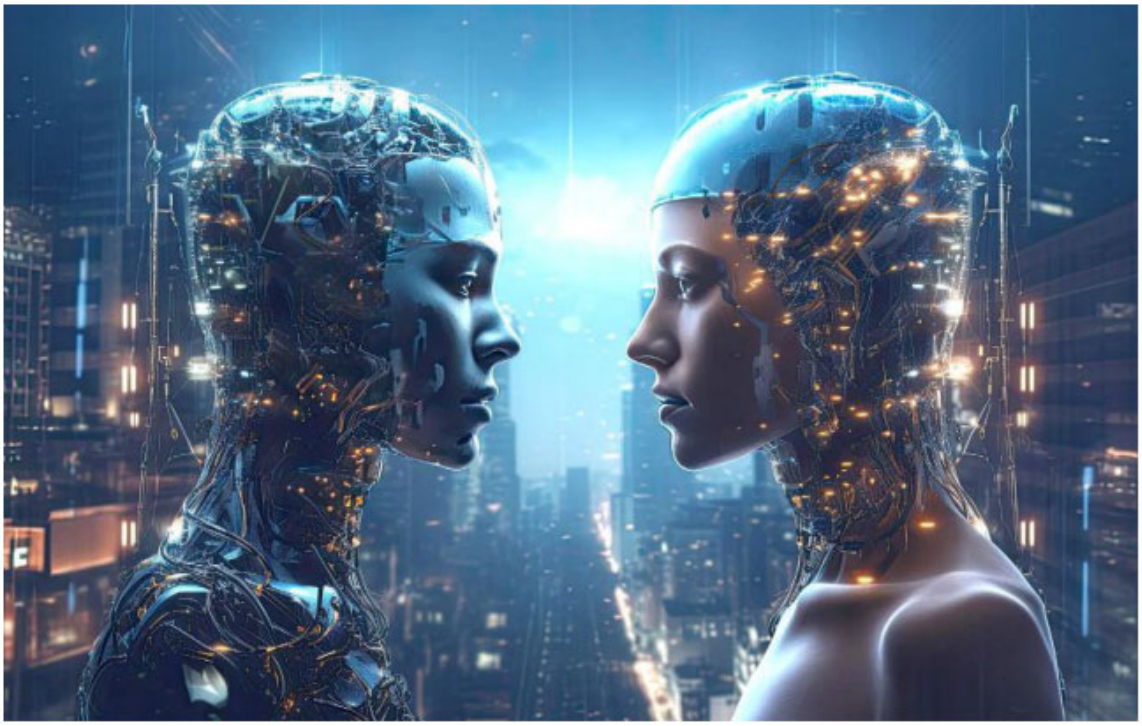
科技伦理道德风险的出现，一方面是基于生成式 AI 技术本身的缺陷，例如训练数据集在多样性、代表性、公正性等方面存在缺陷，导致偏见、刻板印象、片面性等问题。由于数据集的类型区分不清，也可能导致事实与想象不分，加剧错误或虚假信息的扩散和传播。尽管法规中要求对使用人工智能技术生成的内容进行标识，但目前的技术难以做到一一精准识别。

另一方面，科技伦理风险也可能是基于用户对于生成式 AI 模型的滥用，例如用户利用生成式 AI 模型进行破坏社会秩序的行为。如何应对生成式 AI 可能带来的伦理道德风险，也将成为较长时间内人类与 AI 相处过程中的重要议题。

---

14 [系好“安全带”，生成式人工智能才会更好发展，人民网，http://news.sohu.com/a/666350811_114731]

15 [欧洲议会准备通过全球首部人工智能法案，禁止威胁人类安全，美国 42% 的 CEO 认为人工智能将毁灭人类，极目新闻，https://baijiahao.baidu.com/s?id=1768733406669188370&wfr=spider&for=pc。]



第三部分

借鉴：欧美生成式 AI 治理的观察

随着生成式 AI 技术的高速发展，各国都在积极探索关于生成式 AI 的治理思路，这其中既有监管视角的 AI 立法，也有应用视角的 AI 风险管理框架，以期在控制生成式 AI 风险的前提下实现生成式 AI 等技术的发展。下文中，本白皮书将介绍美国和欧盟目前在 AI 治理中的不同思路，了解其他国家关于控制 AI 风险和促进 AI 发展的不同方案，为未来构建完善的生成式 AI 治理模式提供思考与借鉴。

3.1 美国

3.1.1 关于人工智能的立法概况

作为世界科技强国，美国在人工智能技术和产业领域的影响力不可忽视。相较于如何监管的问题，美国更早关注的是如何发展人工智能产业。美国在战略层面上对人工智能持续加大关注与支持，各部门密集出台了一系列人工智能战略与相关政策，内容涉及资金投入、数据资源、人才培养等各个方面。

美国人工智能
立法概况图

- 2016年10月 《为人工智能的未来做好准备》《国家人工智能研发战略规划》
- 2016年12月 《人工智能、自动化与经济》
- 2018年1月 《国防战略报告》
- 2018年8月 《2019财年国防授权法案》
- 2019年2月 《保持美国在人工智能领域的领导地位》
- 2019年6月 《国家人工智能研发战略计划: 2019 年更新》
- 2020年5月 《生成人工智能网络安全法案》
- 2020年8月 《数据问责和透明度法案》
- 2020年11月 《人工智能监管原则草案》
- 2021年1月 《2020年国家人工智能倡议法》
- 2021年5月 《人工智能能力与透明度法案》《军用人工智能法案》《2021算法正义和在线平台透明度法草案》
- 2022年2月 《2022算法问责法草案》
- 2022年8月 《芯片与科学法案》
- 2022年10月 《人工智能权利法案蓝图》
- 2022年12月 《2022推进美国人工智能法案》
- 2023年1月 《人工智能风险管理框架（第一版）》
- 2023年4月 《2020年国家人工智能倡议法案》
- 2023年5月 《国家人工智能研发战略规划》

2016 年 10 月，美国政府发布了《为人工智能的未来做好准备》(Preparing for the Future of Artificial Intelligence) 和《国家人工智能研发战略规划》(The National Artificial Intelligence Research and Development Strategic Plan) 两份报告。前者阐述人工智能的发展现状、未来机遇、潜在问题，并针对美国政府、公共机构和公众提出多项建议，后者则通过一个框架确定了人工智能研发的优先顺序，提出 7 大人工智能研发策略。

同年 12 月，白宫发布《人工智能、自动化与经济》(Artificial Intelligence，Automation，and the Economy) ，进一步调查了人工智能驱动的自动化对美国就业市场和经济的影响，并倡导各方开发、训练人工智能以促进转型，并帮助工作者学会适应人工智能带来的生产力增长。

2018 年 1 月，美国国防部发布该年《国防战略报告》（National Defense Strategy）强调人工智能对美国国家安全具有重大的战略意义。8 月，特朗普政府签署通过了《2019 财年国防授权法案》(National Defense Authorization Act for Fiscal Year 2019)。依据此授权法，美国成立了国家人工智能安全委员会（National S ecurity Commission on Artificial Intelligence/ NSCAI），研究人工智能和机器学习方面的进展，以及它们在国家安全和军事方面的潜在应用。

进入 2019 年 2 月，时任总统特朗普签署行政令——《保持美国在人工智能领域的领导地位》(Executive Order on Maintaining American Leadership in Artificial Intelligence)，提出 5 个关键领域，包括加大人工智能研发投入、开放人工智能资源、设定人工智能治理标准等。6 月，《国家人工智能研发战略计划 : 2019 年更新》(The National AI Research and Development Strategic Plan: 2019 Update) 发布，形成了特朗普政府的 8 大人工智能研发战略。

2020 年 5 月，《生成人工智能网络安全法案》（Generating Artificial Intelligence Networking Security （GAINS） Act），要求美国商务部和联邦贸易委员会明确人工智能在美国应用的优势和障碍；调查其他国家的人工智能战略，并与美国进行比较。

2020 年后，虽然产业发展相关规则的制定还在进行，但也开始出现治理为主的规则。2020 年 8 月，《数据问责和透明度法案》（Data Accountability and Transparency Act of 2020）发布企业相关服务的隐私收集提出规制。11 月，《人工智能监管原则草案》（Guidance for Regulation of Artificial Intelligence Applications）则指出，要规范人工智能发展及应用，要求联邦机构在制定人工智能方法时应考虑 10 项“人工智能应用管理原则”，包括公众对人工智能的信任与参与、风险评估与管理、公平与非歧视、 披露与透明度、安全与保障等。 2021、2022 年也保持相似的情况。在产业促进方面，2021 年 1 月的《2020 年国家人工智能倡议法》（the National AI Initiative Act of 2020) 要求科学技术政策办公室（OSTP）宣布成立国家人工智能计划

办公室和国家人工智能咨询委员会，并建立或指定一个机构间委员会，以更健全完备的组织机构推动“国家人工智能计划”实施；5 月发布的《人工智能能力与透明度法案》（Artificial Intelligence Capabilities and Transparency Act of 2021）和《军用人工智能法案》（Artificial Intelligence for the Military Act of 2021）则从优化人才结构等方面提出建议以促进相关技术应用发展。在治理方面，5 月，《2021 算法正义和在线平台透明度法草案》（Algorithmic Justice and Online Platform Transparency Act）发布，对平台个人信息收集、内容审核以及算法透明度等提出诸多要求。

2022 年 2 月的《2022 算法问责法草案》（Algorithmic Accountability Act of 2022）要求对自动化决策系统和增强的关键决策流程进行影响评估。而同年 8 月的《芯片与科学法案》（Chips and Science Act）则表示，美国将在人工智能、机器人技术、量子计算等关键领域投入 2000 亿美元的研究经费和产业补贴。

在 ChatGPT 及相关技术引发越来越多关注和担忧后，美国的治理思路从重视产业更多转向监管治理与行业发展平衡。

2022 年 10 月美国白宫发布的《人工智能权利法案蓝图》，提出负责任地使用人工智能（AI）路线图。该综合文件确定了指导和管理人工智能系统有效开发和实施的五项核心原则，特别关注侵犯公民权利和人权的意外后果。

而 12 月发布的《2022 推进美国人工智能法案》（Advancing American AI Act）则着眼产业发展，提出增加投资、鼓励运用等多项促进人工智能发展的举措。

今年 1 月，《人工智能风险管理框架（第一版）》（Artificial Intelligence Risk Management Framework）出台，期望为有需求的各方提供可参考的 AI 风险管理框架。4 月，《2020 年国家人工智能倡议法案》（National Artificial Intelligence Initiative Act of 2020）公布。该法案计划在未来五年内提供近 65 亿美元，确立美国在人工智能 (AI) 领域的领导地位。5 月，2023 年的《国家人工智能研发战略规划》(The National Artificial Intelligence Research and Development Strategic Plan) 发布，重申了之前的 8 项战略，同时增加了新的第 9 项战略以强调国际合作。

3.1.2 关于生成式 AI 应用的风险治理框架

美国尚未就生成式 AI 应用制定专门的规定，生成式 AI 应用遵循其关于人工智能技术应用的相关规范。NIST 在 2023 年 1 月发布 AIRMF（AI Risk Framework）1.0 作为引导行业实践的 AI 风险管理框架文件，也许代表着美国对于生成式 AI 应用风险治理的主要思路。

2023 年 1 月，美国国家标准与技术研究院（NIST）在与私营和公共部门的紧密合作下发布了

《人工智能风险管理框架》（Artificial Intelligence Risk Management Framework）（以下简称“框架”）及配套使用手册，该框架是一份非强制性的指导文件，供设计、开发、部署或使用人工智能系统的组织自愿使用，以帮助组织机构管理人工智能技术应用过程中的相关风险。该框架为降低人工智能系统对公民自由和权利造成的威胁并实现人工智能系统积极影响的最大化提供了路径，从而使得人工智能系统更加安全可信赖。在适应人工智能技术发展的背景下，该框架旨在帮助企业和组织根据自身能力和需求制定人工智能风险管理框架，实现人工智能技术应用过程中的风险管理，使得社会在受益于人工智能技术的同时免受其害[16]。

框架正文由两大部分构成。第一部分为基础信息介绍，包括企业如何界定人工智能系统的相关风险，并阐述框架的受众和有效性[17]。除此之外，还概述了可信的人工智能系统的特征，即可解释性、透明度和问责制。通过将这些特征纳入人工智能系统，各组织或企业可以确保其系统的可信性和安全性，避免对个人或社会产生不必要的风险。

可信的人工智能

· **可解释性**（Explainability）是指人工智能系统对其决策过程提供清晰和可理解的解释的能力。若缺乏可解释性，系统会产生与偏见、歧视和其他负面影响有关的风险。

· **透明度**（Transparency）是指人工智能系统以利益相关方可以理解的方式提供有关其运作和决策过程的信息的能力。这包括提供关于数据来源、所使用的算法以及可能影响系统性能的其他相关因素的信息。

16 [NIST: NIST Risk Management Framework Aims to Improve Trustworthiness of Artificial Intelligence, https://www.nist.gov/news-events/news/2023/01/nist-risk-management-framework-aims-improve-trustworthiness-artificial.]

17 [窦传喜：《美国NIST〈人工智能风险管理框架〉评述》，载清华大学智能法治研究院，https://mp.weixin.qq.com/s/7njqOOXIIFrf6LizDCXFOg。]

· **问责制**（Accountability）是指人工智能系统对其行为负责的能力。这包括确保有相应机制来识别和解决系统决策过程中的错误或偏见。

第二部分是框架核心和用例配置文件。该框架由四个顶层模块组成，分别是：

· **治理**（Govern）功能旨在为其他三个功能提供信息并贯穿其中。它涉及到明确人类在“人类与人工智能团队”中的角色和责任；明确系统性能监督者的角色和责任。它还旨在使系统的决策过程更加明确，并帮助对抗系统性的偏见。

· **映射**（Map）功能涉及识别和理解与人工智能系统相关的风险，包括确定潜在漏洞，评估人工智能系统故障的影响，以及了解人工智能系统输出的潜在后果等。它可以帮助企业提高其内部的风险识别能力，以识别系统的局限性，探索和检查基于人工智能的系统在现实世界中的影响，并评估其整个生命周期的决策过程。映射有助于企业全面了解人工智能系统所涉及的风险。

· **衡量**（Measure）功能的重点是量化和评估人工智能系统风险。它涉及制定衡量标准来评估人工智能系统的性能和有效性。衡量功能可以帮助企业评估人工智能系统风险的影响，并制定明智的风险缓解策略。

· **管理**（Manage）功能涉及制定和实施风险缓解战略和风险控制战略，以解决已明确的人工智能风险。它包括制定和实施政策、程序和技术保障措施等活动，以缓解和控制整个人工智能系统生命周期的风险。管理功能确保采取适当的措施，最大限度地减少人工智能系统的潜在负面影响。

该框架具有高度的抽象性，因此 NIST 并没有提供标准的配置模板。组织或企业可以根据自己的需求、风险偏好以及成本资源等因素灵活实施此框架。此外，人工智能风险管理框架还提供了在线配套使用手册，为管理者提供了具体的行动建议，包括示例和参考资料等。

3.2 欧盟

3.2.1 关于人工智能的立法概况

人工智能一直是欧盟数字立法计划关注的主题之一。2018 年 5 月，为适应云计算、互联网、大数据等新技术应用的影响，欧盟“最严”数据保护立法《通用数据保护条例》（General Data Protection Regulation）正式施行，在世界范围内引发关注。而在 GDPR 落地前一个月，欧洲经济和社会委员会（European Economic and Social Committee）发布《欧洲人工智能

战略》（European Strategy for Artificial Intelligence），着眼人工智能领域，提出要增加对人工智能的公共和私人投资，并确保适当的道德和法律框架。同年 12 月，欧盟委员会（European Commission）发布《人工智能协调计划》（Coordinated plan on AI），意在协调各成员国合作落实《欧洲人工智能战略》。

2019 年，欧盟在人工智能算法、伦理方面有所发力。4 月，《算法问责及透明治理框架》（A Governance Framework for Algorithmic Accountability and Transparency）发布，就算法及其在自动化决策系统中的应用快速增长提出了全面的监管框架。同月，《可信赖人工智能伦理准则》（Ethics guidelines for trustworthy AI）提出尊重人自主性、预防伤害、公平性和可解释性四项伦理准则和人的能动性和监督、技术稳健性和安全性等伦理准则七要素。进入 2020 年，欧盟委员会发布《人工智能白皮书》（White paper on AI: a European approach to excellence and trust），其中表示将协助欧洲各国同美、中等国在人工智能与科技领域抗衡。发布的公告还提到，欧盟已根据“数字欧洲计划”提出了超过 40 亿欧元的建议，以支持高性能和量子计算，包括边缘计算和人工智能，数据和云基础设施。这一年，还有《关于发展人工智能技术的知识产权的决议》（Intellectual property rights for the development of artificial intelligence technologies）、《人工智能、机器人和相关技术的伦理问题框架》（Framework of ethical aspects of artificial

intelligence, robotics and related technologies）等针对性解决人工智能发展某一部分问题的文件出台。

2021 年，欧盟委员会发布《2030 数字指南针：欧洲数字十年之路》（2030 Digital Compass：the European way for the Digital Decade），其中指出到 2030 年，数据公平共享将成为数据经济的重要基础，5G、物联网、边缘计算、人工智能、机器人、增强现实等数字技术将成为新产品、新制造流程、新商业模式的核心（而非手段）。在这年 4 月，《人工智能法案》立法提案发布。这是世界范围内第一部针对人工智能进行规制的法律，主要特点是依循风险分类分级的思路对人工智能系统进行监管治理。2021 年至今，该提案历经多次更改。今年以来，随着 ChatGPT 影响力逐渐扩大，人工智能产业格局迎来变动，法案亦有新增和变更。目前最新版针对 ChatGPT 等生成式 AI 系统的提供者提出了如下要求，如果用以训练的数据受版权保护，提供者必须公开相关详细摘要；针对生成的内容，法案则要求其遵守透明度要求，对人工智能生成内容进行标注和披露；此外，还需采取措施防止生成非法内容。值得注意的是，相关模型在欧盟市场上发布前，基础模型的提供者还需要在欧盟数据库中注册。

在《人工智能法案》立法提案发布并不断修改的过程中，《人工智能协调计划 2021 年修订版》（Coordinated Plan on Artificial Intelligence 2021 Review）、《关于数字时代人工智能的决议》（Artificial intelligence in a digital age）、《欧洲人工智能责任指令（草案）》（Proposal for an AI liability directive）相继发布，从监管治理和产业发展等多个方面给出更详细的建议。 值得注意的是，在 2022 年后半年，欧洲数字立法体系中又有两部重要立法公布——《数字市场法》（Digital Markets Act）和《数字服务法》（Digital Services Act）。目前实践显示，这两部法律虽不是只针对人工智能，但其对在线平台和数字市场的监管一定程度上可能会对欧洲人工智能发展格局产生影响。

3.2.2 关于生成式 AI 应用的风险治理框架

欧盟对于生成式 AI 应用的风险治理遵循《人工智能法案》等 AI 技术应用规范的规定，主要采用以人工智能法案为主的风险分级监管治理框架。

2021 年 4 月 21 日，欧盟委员会提出了《关于制定人工智能统一规则》的提案。自此之后，欧洲议会和欧盟理事会就提案进行了多轮修订和讨论。2023 年 4 月 27 日，欧盟议会成员达成临时政治协议。6 月 14 日，欧洲议会以 499 票赞成、28 票反对和 93 票弃权的高票通过了《人工智能法案》谈判授权草案（以下简称“法案”）。根据欧盟立法程序，欧洲议会、欧盟成员国和欧盟委员会将开始“三方谈判”，以确定最终条款。据德国《商报》报道，《人工智能法案》预计于今

年年底获得最终批准，但完全生效可能还需要数年时间[18]。

《人工智能法案》是全球首个专门提出 AI 风险分级监管的综合性人工智能法案，同时这部法案具有跨国界性：无论企业实体位于何处，只要该企业在欧盟市场开放服务或部署人工智能系统，都必须接受法案监管。针对欧盟范围内人工智能驱动的产品、服务及系统，法案对其开发、贸易和使用制定了核心规则。对于未来有意提升人工智能治理水平的企业和组织，法案将提供重要参考依据。该法案的目标是创建可信的人工智能产品和服务，让用户最终能够相信人工智能技术将被安全和合规地使用。除了设立欧洲人工智能委员会外，该法案还要求设立机构来监督法案具体内容的实施。对此，每个欧盟成员国须设立一个单独的监督机构[19]。

《人工智能法案》的立足之本是一个风险分类系统（见下图），用于判定人工智能技术可能对人类健康、安全或基本权利造成的潜在风险等级。该系统包括四个风险等级：不可接受的风险、高风险、有限风险和最小风险。针对不同等级的风险，法案将实施不同程度的控制措施。

1. 不可接受的风险，位于危险金字塔的顶端，指对人们的安全、生计和权利有明显威胁的人工智能系统。AI 产品或服务若具有“不可接受的风险”，将被明确禁止在欧盟市场上销售和使用。包括：

· 对人造成伤害的潜意识、操纵性或剥削性系统；

· 在公共场所用于执法的实时、远程生物识别系统，以及；

· 所有形式的社会评分系统，例如基于社交行为或预测的人格特征来评估个人可信度的人工智能技术。

2. 高风险，属于法案的重点限制对象，大致分为两类：一类是该法案附件二中列出的人工智

18 [王卫：欧盟《人工智能法案》进入最终谈判阶段，载法制日报，https://h5.drcnet.com.cn/docview.aspx?version=edu&docid=6981208&chnid=1020。]

19 [于品显，刘倩：《欧盟〈人工智能法案〉评述及启示》，载微信公众号海南自由贸易港金融学会，https://mp.weixin.qq.com/s/QHwOuPRMJQ3ehs-pX6mpLQ。]

能系统作为产品安全组件的物理产品，一类是附件三中列出的更多基于人工智能系统软件的产品。列入附件二的包括机械、玩具和医疗设备，列入附件三的则包括[20]：

· 重要基础设施（如交通），其使用可能会危及人们的生命和健康；

· 教育或职业培训，其使用可能会决定一个人的受教育机会和专业课程（如考试评分系统）；

· 产品的安全组件（例如，人工智能在机器人辅助手术中的应用）；

· 就业或员工自营职业机会的管理（例如，用于招聘程序的简历排序软件）；

· 基本的私人或公共服务（例如，可能剥夺公民获贷款机会的信用评分系统）；

· 可能干涉公民基本权利的执法（例如，用于评估证据可靠性的系统）；

· 移民、庇护和边境控制管理（例如，用核实旅行文件的真实性的系统）；

· 司法和民主进程的管理（例如，用于将法律适用于一系列具体的事实的系统）。

高风险的人工智能系统在投入市场之前将受到严格的义务约束：

a. 充分的风险评估和风险缓解系统；

b. 为系统提供高质量的数据集，以尽量减少风险和歧视性的结果；

c. 对活动进行记录，以确保结果的可追溯性；

d. 提供详细文件，包括有关该系统及其目的的所有必要信息，以便监管部门评估其合规性；

e. 向用户提供清晰和充分的信息；

f. 适当的人力监督措施，以最大限度地减少风险；

g. 高水平的鲁棒性、安全性和准确性。

3. 有限风险，指具有特定透明度义务的人工智能系统。例如，当使用人工智能系统（如聊天机器人）时，用户应该意识到他们正在与机器进行互动，这样他们就可以对是否继续使用做出理智的决定。

4. 最小风险，该法案允许自由使用具有“最小风险”的人工智能系统。包括人工智能支持的视频游戏或垃圾邮件过滤器等应用。目前在欧盟使用的绝大多数人工智能系统都属于这一类型。

在处罚层面，违反《人工智能法案》的潜在罚金很高，且随着立法的进展大幅增加。违规的最高处罚是 4000 万欧元或公司上一财政年度全球总营业额的 7%，以较高者为准。相比之下，这几乎是 GDPR 最高处罚的两倍。

---

20 [European Commission: Regulatory framework proposal on artificial intelligence, https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.]

由于人工智能是一种快速发展的技术，该法案还有一个面向未来的方法，为规则随技术变化而调整提供可能性。人工智能应用在投放市场后仍应保持可信度。在系统进入市场后，相关监管机构应负责市场监督，用户应负责人力监督，供应商则需建立市场后监测系统。若遇到严重事故和系统故障，供应商和用户也应及时作出报告。

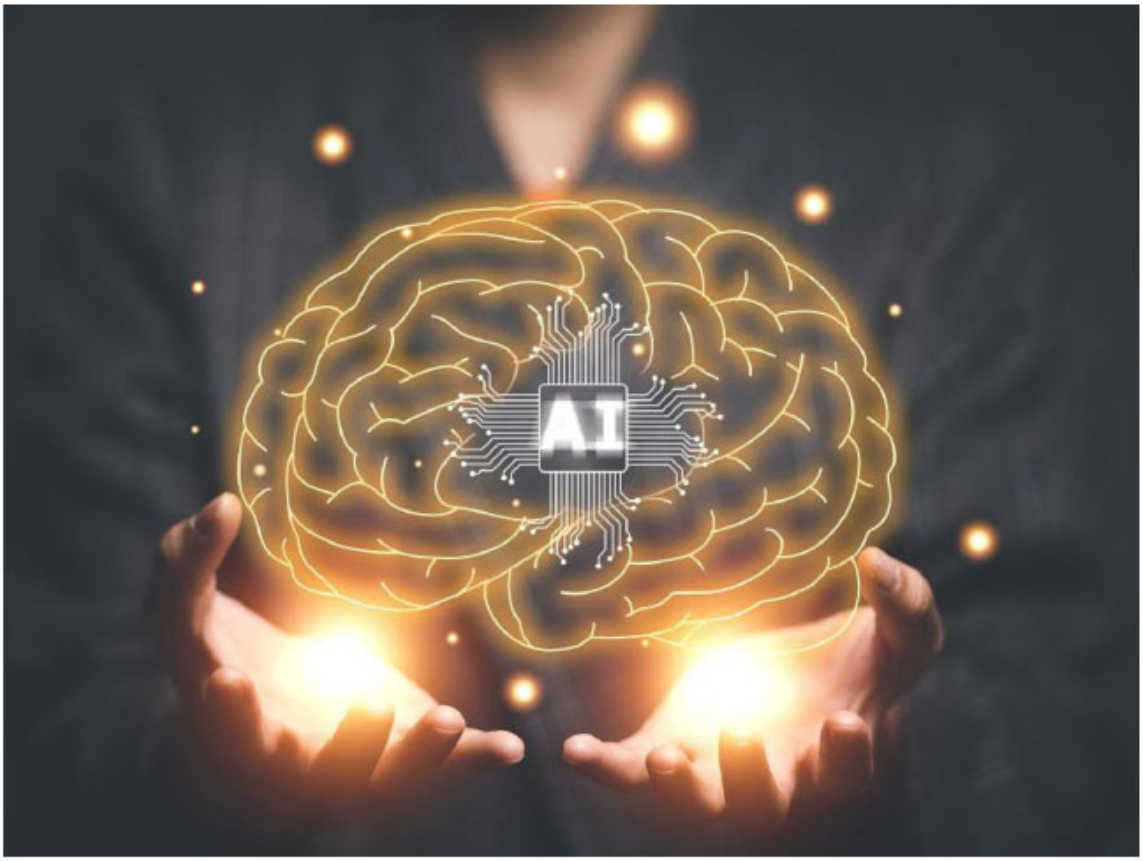
此外，2022 年 9 月 28 日，欧盟委员会通过了《人工智能责任指令》和《产品责任指令》修订版两项提案。前者确定了针对人工智能系统所致损害的适用规则，后者将其适用范围扩展到配备人工智能的产品。这两项法规与《人工智能法案》共同构成了欧盟当局从立法层面应对数字时代和循环经济的三驾马车。

3.3 关于美国与欧盟风险治理框架的评析

目前，各国关于生成式 AI 应用的风险治理框架均处于探索阶段，并无孰优孰劣，通过了解各国对于生成式 AI 应用的风险治理思路和应对方案，有利于为我国未来构建完善的生成式 AI 应用风险治理框架提供思考与借鉴。

美国采取以 NIST 的 AIRMF 为主引导实践的治理框架，其未在联邦或者州层面制定相关的法律文件来规制生成式 AI 的应用以及其可能带来的风险，而是由美国国家标准与技术研究院发布了一份非强制性的指导文件，指导使用生成式 AI 技术的机构组织在设计、开发、部署、使用人工智能系统时采取相应措施，降低安全风险，提高人工智能的可信度，保障公民的相关权益。美国关于生成式 AI 应用的治理框架反映了其对于创新技术的一贯态度，即鼓励科技发展，着力保持自身在全球科技创新中的领导地位。该种治理框架对于生成式 AI 技术的应用者而言极具灵活性，但由于其不具有强制性，其对于生成式 AI 应用风险治理的实际效果以及在保障公民的相关权益方面，仍有待实践检验。

欧盟采用了以人工智能法案为主的风险分级监管治理框架，将人工智能技术可能对人的健康和安全或基本权利造成的风险程度分为四个级别，并分别采取不同的限制措施。相较于美国针对生成式 AI 技术应用采取鼓励型的监管模式，欧盟则显得相对保守。欧盟的风险治理框架有利于管控信息泄露问题，保障公民的相关权益，但同时，由于违反《人工智能法案》的代价极高，企业在应用生成式 AI 技术的同时，也需要其耗费成本采取措施实现生成式 AI 技术应用合规，也在很大程度上将限制生成式 AI 的发展。



第四部分

实践：中国关于生成式 AI 治理的独立思考

生成式 AI 的快速发展给人类带来了机遇与挑战，如何利用生成式 AI 技术造福人类，同时又能够有效控制技术在应用过程中产生的种种风险？各个国家都在积极探索解决路径，中国也不例外。2023 年 7 月 10 日，随着《生成式人工智能服务管理暂行办法》的出台，中国对于生成式 AI 治理也给出了自己的方案。

4.1 关于生成式 AI 的立法概况

中国的人工智能治理之路始于 2017 年。当年 7 月，国务院发布《新一代人工智能发展规划》，提出人工智能三步走的战略目标，并设置了 2020 年、2025 年及 2030 年三个时间节点，目标覆盖人工智能技术理论、产业发展、治理体系等领域。

2019 年，国家新一代人工智能治理专业委员会在 6 月、9 月先后发布《新一代人工智能治理原则》《新一代人工智能伦理规范》，前者强调了“发展负责任的人工智能”这一主题，并提出发展相关方需要遵循的八项原则；后者则提出要将“伦理道德”这部“软法”融入至人工智能研发和应用的全生命周期。

由于算法歧视、“大数据杀熟”等算法不合理应用问题日渐突出，2021 年 12 月 31 日，国家互联网信息办公室联合公安部等四部门联合发布了《互联网信息服务算法推荐管理规定》，主要用于规范算法推荐服务提供者在使用包括生成合成类等算法推荐技术提供服务。

2022 年，我国各地开始考虑人工智能产业发展问题。在 9 月，深圳、上海相继发布《深圳经济特区人工智能产业促进条例》和《上海市促进人工智能产业发展条例》，对地方人工智能产业发展和治理提出要求，值得注意的是，两地在治理环节都提到要建立人工智能伦理（专家）委员会、采用风险分类分级机制对人工智能进行管理。同年 11 月，为应对元宇宙等概念兴起，AI 换脸等深度伪造技术引发的社会事件，监管部门对于深度合成技术应用做出回应，发布 《互联网信息服务深度合成管理规定》，对深度合成技术提出一系列要求，其中，将深度合成技术定义为“利用生成合成类算法制作文本、图像、音频等的技术”。

时间来到 2023 年，ChatGPT 的面世和大模型领域的密集动态无疑给人工智能的烈火再添了一把新柴。我国在治理上亦有相应动作。4 月，国家互联网信息办公室发布《生成式人工智能服务管理办法（征求意见稿）》，对生成式 AI 产品或服务提供者的责任、数据安全等与生成式 AI 技术密切相关的问题做出回应；5 月，国务院发布的 2023 年度立法工作计划中，人工智能法草案赫然在列；7 月 13 日，《生成式人工智能服务管理暂行办法》（以下简称《办法》）正式出台，与此前征求意见稿相比有较大的思路调整。《办法》强调实行包容审慎和分类分级监管，并单设“技术发展与治理”章节，新增了不少有力措施来鼓励生成式 AI 技术发展，比如推动生成式 AI 基础设施和公共训练数据资源平台建设，促进基础技术的自主创新。并明确了训练数据处理活动和数据标注等要求。

4.2 关于生成式 AI 应用的风险治理框架

不同于美国和欧盟采取的治理方式，中国对于生成式 AI 应用的规制主要采用目标与问题导向相结合的风险治理框架，既采取措施支持和鼓励生成式 AI 应用的发展，又针对生成式 AI 应用场景下可能引发的风险进行规制。

我国针对生成式 AI 应用的治理经历了一个从“问题导向”向“目标与问题导向相结合”的治

理模式的转变。如前文所述，由于算法歧视、“大数据杀熟”等算法不合理应用问题日渐突出，元宇宙、AI 换脸等新概念、新技术影响扩大，网信办于 2021 年、2022 年先后发布《互联网信息服务算法推荐管理规定》和《互联网信息服务深度合成管理规定》。

《算法推荐管理规定》与《深度合成管理规定》主要是以问题为导向的治理框架，主要关注包括生成合成类算法服务、深度合成服务在内的生成式 AI 技术在应用过程中的风险问题，通过制定相关规则和采取措施限制相关方在生成式 AI 技术应用过程中的风险。

同时，基于 ChatGPT 等大模型的兴起，为规制 GPT 应用可能诱发的相关风险，2023 年 7 月 13 日，网信办等六部门联合正式发布了《生成式人工智能服务管理暂行办法》，该办法于 2023 年 8 月 15 日正式生效。相比于征求意见稿，新出台的《办法》试图在实现风险控制的同时，鼓励发展生成式 AI 应用。

不难看出，我国监管部门对于生成式 AI 应用风险的治理框架主要采用目标与问题导向相结合的风险治理框架，以期在促进生成式 AI 健康发展的同时，又能够确保生成式 AI 的规范应用。“技术应当是中立的”，通过规制能够更好地促进其良性发展，为其保驾护航。我国的风险治理框架既不同于美国基本没有限制的风险监管模式，也不同于欧盟多重限制的风险监管模式，其介于二者之间，对于实现风险监管与鼓励科技创新的平衡更具优势，但具体实施效果依然有待时间与实践检验。

4.3 关于商业化应用中生成式 AI 风险治理的思考

随着生成式 AI 商业化应用在全球范围内的兴起，业界对生成式 AI 应用商业化所伴生的风险有了越来越多的讨论与认识。未来，生成式 AI 商业化应用将需要更高质量的训练数据，面向更多用户和利益相关方，投入更复杂的场景应用。这些因素都会放大生成式 AI 产品在商业化中各种风险，企业在生成式 AI 的商业化之路面临更多的挑战和压力。

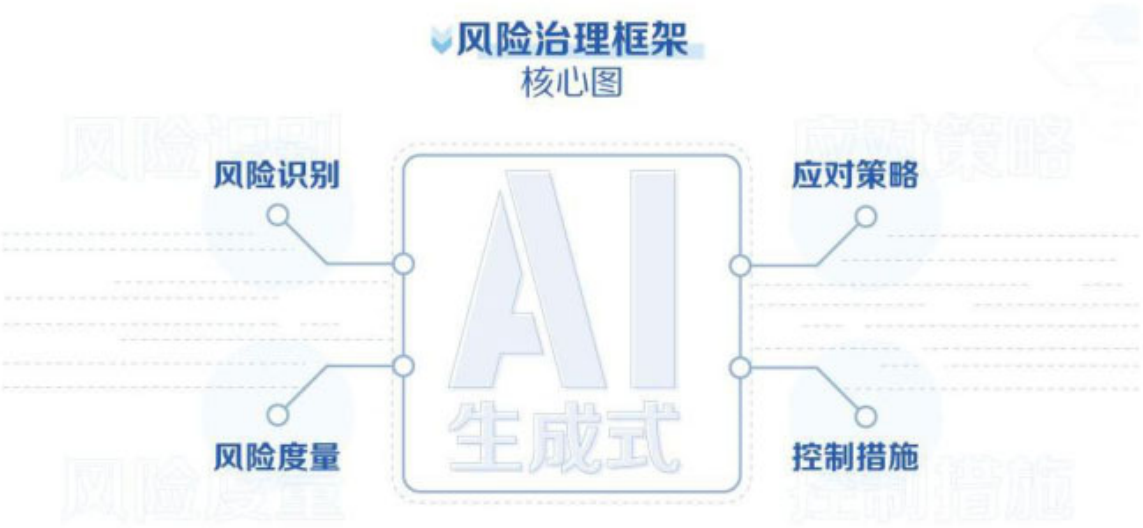
一场生成式 AI 应用风险的讨论正在全球展开。此前一些国家的监管案例已经展现出监管机构的对待生成式 AI 产品商业利用的态度，监管机构认为，生成式 AI 产品商业利用还需跨过妥善治理的“门槛”。换言之，如果相关企业不能证明具备良好的风险治理水平，就意味着其经营活动面临监管风险和业务连续性风险，OpenAI 被要求展现更多治理措施就是一个例子。

完善的风险治理框架无论从监管者视角还是投资者视角都是关注的重点。从 OpenAI 等企业受到监管关注不难发现，一家可持续发展的 AI 企业必须具备良好的风险治理水平、高效的风险管理能力以及持续的合规风险治理更新和改善能力。要有效利用生成式 AI 技术，企业首先要解决好风险治理框架的问题。

在本白皮书第三部介绍了美国 NIST 的 AI RMF 关于 AI 风险管理框架的实践以及欧盟人工智能法案提出的一些关于 AI 风险治理可供参考的建议，但是生成式 AI 商业化应用的开发者仍需要根据自身情况设计一套适合自身的生成式 AI 商业化应用风险治理框架。考虑到像 AI RMF 这样的行业实践只给出一般性的风险类型和治理手段，在具体落实时，开发者需要根据自身生成式 AI 商业化应用情况量化这些风险，设置合理的容忍度，并确定最适合自己的控制措施。同时，我们也需要意识到，不同开发者的生成式 AI 商业化应用、环境和目标存在较大差异。现阶段法规要求较为宽泛，开发者仍需要提前制定自己的治理框架以配合未来法规。开发者在进行生成式 AI 商业化应用中需要怎样的 AI 风险治理框架，本白皮书写作团队给出如下的建议：

风险控制是治理框架的核心目标导向。技术、商业模式、政策环境等不断变化，风险是各方最关注，也是目前最好的治理抓手，为此可以考虑通过如下四个模块来构建、形成开发者针对生成式 AI 商业应用风险的治理框架：

风险治理框架的核心（风险识别 - 风险度量 - 应对策略 - 控制措施）；



a、**风险识别**：生成式 AI 应用的风险识别是风险治理的起点，目前常见的应用风险包括：知识产权、算法、数据安全以及个人信息等方面，部分未明确定义或未充分理解的生成式 AI 应用风险开发者需要保持关注并持续更新。同时也提醒，无法适当识别的风险并不意味着生成式 AI 商业化应用一定会带来高风险或低风险；

b、**风险度量**：典型的风险衡量方法需要开发者将测量或估计的影响和影响可能性相乘或定性组合为风险评分（风险≈影响 x 可能性）。试图完全消除负面风险在实践中可能适得其反，因为并非所有事件和风险都可以消除。对风险不切实际的期望可能会导致企业不当分配资源，造成风险分

类效率低下或不切实际，或者浪费稀缺资源；

c、应对策略：开发者针对已识别的生成式 AI 商业化应用风险的应对治理模块，目前包括以下模块：协议安排、内部政策、技术性限制、数据治理、动态测试及评估、透明度实践；

d、控制措施：依据 c 风险应对策略，确定的具体风险控制节点，需要开发者内部具体落实的措施。

此外，生成式 AI 应用风险事件库也是一个值得借鉴的行业实践。AI 风险事件数据库通过整理和分析生成式 AI 商业化应用的开发和部署在现实商业领域中造成的或可能造成不良后果或者损害的情况，形成相应数据库。与航空和计算机安全领域的类似数据库一样，生成式 AI 应用风险事件数据库旨在从经验中学习，以便相关开发者能够预防或减轻不良后果。

（本报告图片来源：视觉中国）