

生成式人工智能运行中的风险分析与治理路径

李语涵¹

1 澳门科技大学法学院, 826994071@qq.com

摘要:以 ChatGPT 为代表的生成式人工智能推动 AI 从辅助分析迈入自主生成,也使得法律风险在运行环节之间扩散、在不同主体之间传导,并跨领域叠加。本文将人工智能的运行界定为一个持续生成、积累并外溢风险的动态过程,沿数据采集、算法训练与内容输出三个阶段展开系统梳理。研究认为,人工智能风险呈现出“来源于数据、加剧于算法、溢出于模型”的链式传导逻辑,具体表现为训练数据来源合法性缺失、数据质量失真、算法黑箱与透明度不足、知识产权侵害、虚假信息传播以及责任主体识别困难等。在输出端,知识产权权属不明与虚假信息责任链条断裂相互交织,集中暴露了传统侵权法与平台责任规则在生成式 AI 场景下的适用困境。有鉴于此,应构建覆盖事前预防、事中监管与事后救济的全链条治理体系。人工智能治理的关键,不在于以静态规则简单限缩技术空间,而在于在动态调适中寻求创新激励与风险防控的制度平衡,这一平衡点亦须嵌入数据源头、算法过程和输出影响三个环节,方能回应生成式人工智能链式传导的规范挑战。

关键词:生成式人工智能,运行风险,算法治理,治理路径

以 ChatGPT 为代表的生成式人工智能的诞生,标志着人工智能技术的重大技术跃迁,但同样也引发了一系列的系统性风险。2023 年马塔诉阿维安卡公司案中,^[49]因律师提交由 ChatGPT 生成的虚构判例的法律文书而受到制裁。该事件表明,生成式人工智能打破了虚拟与现实的传统界限,其风险层级、影响范围和传导机制均呈现出质的变化。

当前学界对人工智能法律问题的研究,大致集中于两个维度。其一是数据维度,聚焦于训练数据中的个人信息保护、著作权侵权以及数据污染问题。其二是算法维度,强调算法透明度、可解释性与反歧视的制度设计,典型如欧盟《通用数据保护条例》(GDPR)第 22 条确立的免受自动化决策约束权,以及我国《个人信息保护法》第二十四条对自动化决策的专项规定。这些研究成果推动了人工智能法治的奠基,但总体而言,既有研究仍存在风险分类缺乏结构化框架,对风险治理研究路径系统建构稍显不足。本文将人工智能的运行理解作为一种持续的风险生成过程,从理论和实践相结合的角度,系统梳理其在数据采集、算法训练、内容生成三个阶段的风险类型,并在此基础上提出构建涵盖预防、管控、救济等三个层面的全链条治理体系。

1 生成式人工智能理论阐释

生成式人工智能目前主要包括两类基础训练模式:其一为生成式对抗网络(GAN),其二为生成式预训练转换器(GPT)。^[48]前者主要依靠“生成器”与“判别器”两个神经网络之间的对抗博弈实现内容生成与模型优化;后者则通过大规模数据预训练,学习语言、图像或其他符号系统中的统计规律,在特定提示词或上下文条件下生成类人的文本、图像、音频等内容。以 ChatGPT 为代表的大语言模型即属于 GPT 路线。相较于传统基于规则匹配或单一任务识别的人工智能系统,生成式人工智能的核心特征在于:其不再仅仅是“识别”或“判断”,而是通过对既有数据模式的吸收、提炼与重组,生成具有创造性的内容。生成式人工智能的运行并非单一过程,其主要包

作者简介:李语涵,女,澳门科技大学法学院,博士研究生在读,研究方向为国际私法、国际仲裁。

括数据输入与训练的准备阶段,运用自身算法技术处理输入数据的运算阶段,输出数据并进行传播对社会行业造成影响的生成阶段。^[42]在数据收集阶段,生成式人工智能主要通过对其他作品编辑形成自己的数据库;在学习阶段,人工智能通过分析先前所收集的大量数据,使用最优算法模型,并对模型开展质量评估;在输出阶段,人工智能模型可根据用户的指令生成具有创造性的人工智能生成物。由此可见,生成式人工智能的运作是一条完整且各环节密不可分的链条,一旦其中某一环节出错将导致后续步骤无法成功运作。正是基于这一技术特性,生成式人工智能在运行过程中引发一系列风险。^[37]

2 生成式人工智能运行全链条的风险分析

生成式人工智能的风险并非孤立存在于某一环节,而是沿“准备—运行—输出”三阶段依次传导、层层累积。准备阶段的数据缺陷构成风险之源,运行阶段的算法运作将风险放大并使之隐蔽化,输出阶段则使风险以侵害具体法益的形式外部化。因此,有必要从全链条视角,以“风险机制—受侵害法益—制度障碍”为分析框架,对各阶段风险进行系统性审视,进而揭示现行制度在应对生成式人工智能风险时的结构性失灵。

2.1 准备阶段

数据海量性是生成式人工智能运行的基础,模型需要从大量数据中学习概率分布以生成内容。^[8]作为大模型研发与优化的前提,在准备阶段开发者需要将现有的海量作品转换成人工智能可以识别和处理的数据类型,继而对数据进行筛选、标注与加工。正因这一阶段决定了模型的价值取向和行为边界,数据收集与处理的合法性和合规性就显得尤为重要。

2.1.1 数据质量风险

生成式人工智能的数据质量是保证模型运行和生成可信内容的关键。其风险机制主要表现在三个层面。一是数据的准确性缺陷。由于人工智能在前期训练中所使用的数据,均由开发者从网络上采集与筛选,一旦此类数据未经有效甄别即被纳入训练语料,模型在训练过程中便会将错误信息作为“知识”学习并在输出环节进一步放大。具体而言,若遗漏或歪曲关键历史事实,用户查询相关内容时会得到错误结论,进而影响其历史认知。^[5]二是数据代表性的缺陷。在复杂多变的大数据环境下,若训练数据未能全面覆盖各类语言表达与知识命题,或网络公开数据本身包含虚假、误导、错误信息,大模型会形成系统性偏见,导致推理结果不准确、适用范围受限,最终降低算法输出的规范性与可靠性。^[12]当训练数据源自充斥着偏见与结构性不平等信息的互联网时,模型所显现出的正是对社会现实的记录与映射。算法通过统计学习对数据进行规律提取,潜藏于数据中的社会偏见便以技术的形式得到复制与强化。以全球数据生态为例,特定族裔的语料资源无论在规模还是质量上均存在明显短板,加之算法在数据采样与语料构建环节本身即沿袭了既有的种族刻板认知,最终形成以主流群体为参照中心的语义体系,致使边缘群体的表达空间受到压缩。^[15]三是数据完整性的缺陷。数据样本往往集聚于特定地域、语言或文化体系,形成结构性的资源失衡。以 ChatGPT 为例,中文资料仅占其训练数据的 0.16%,^[38]致使模型在与中文用户交互时持续输出嵌入西方意识形态与价值判断的结论,形成一种技术驱动的文化单向度传播。^[27]

2.1.2 个人信息处理的风险

与传统信息处理活动不同,人工智能大模型在训练阶段需要海量数据支撑,数据来源也呈现出多元、持续更新的特点。数据采集是一项持续性工作,会跟随模型运行与优化不断推进,而非仅在前期筹备阶段开展。大量个人信息也会借助不同方式与渠道,陆续融入模型训练及再训练的数据库。因此,会产生多种侵害个人信息权益的不足。

首先,大语言模型技术运行突破“知情同意”规则。在海量数据与持续再训练的技术现实下,难以逐一实现知情同意,违背现行个人信息保护法以“知情同意”为核心的规则。即便平台试图通过告知同意制度让用户进行信息自我决定,其效果也只是流于形式。^[1]例如,DeepSeek 隐私政策中明确了在确保数据已加密、去标识化且无法追溯到个人的情况下,可能会把用户的输入与对应输出用于改进模型和优化服务质量。^[41]但仅在隐私政策中提及不足以实现充分告知的义务,这种利用“入门”用户协议来告知个人信息处理的方式,使得用户面对冗长的隐私政策既无阅读意愿,也存在缺乏理解的能力。虽然《个人信息保护法》第 13 条第(六)项规定了“合理范围内处理已公开的个人信息”的例外情形,但“合理范围”的边界在大模型训练语境下高度模糊,将已公开信息纳入商业化模型训练是否属于“合理范围”,法律并未给出明确回答。这使得知情同意制度在生成式人工智能领域陷入了严格适用则技术发展受阻、宽泛解释则权利保护虚化的双重困境。^[39]

其次,将用户基于特定场景发布的公开信息进行去语境化整合与规模化抓取,会打破用户对信息使用的预期,让公开信息脱离原本使用场景,从而忽视了个人隐私权的保护。^[14]即使单条数据不涉及隐私,但经大模型的关联分析与模式识别,仍可能拼合出高度私密的个人画像。2023 年 6 月,OpenAI 被指控其研发的 ChatGPT 等产品在未经充分授权的情况下,擅自窃取数亿互联网用户的私人信息用于模型训练,涵盖医疗记录与儿童信息,违反了州和联邦的隐私法。^[30]《个人信息保护法》第 27 条虽要求“对个人权益有重大影响的”应取得同意,但大模型训练对个人究竟是否构成“重大影响”缺乏判断标准,场景一致性理念亦未引入规范,导致去语境化整合行为难以有效规制。

此外,模型通过用户交互数据持续再训练,使个人信息处理进入“收集—训练—交互—再收集”的循环,信息主体几乎无法追踪其信息的流转状态,更无法有效行使删除权。同时,当已公开信息进入训练集并与其他数据融合后,信息主体若要求行使删除权,在技术上是否需要重新训练模型、制度上如何实现,均无明确答案。^[46]

2.2 运行阶段

人工智能系统从静态代码转化为动态行为,涉及数据的持续处理、自动化决策的实时输出以及不与特定主体的交互,由此衍生的法律风险具有隐蔽性、系统性和持续性特征。

2.2.1 数据泄露风险

在“模型即服务”(MaaS)架构下,生成式人工智能的运行始终伴随着多源数据的持续汇聚、留存与再学习。在模型训练与持续优化过程中,用户可能会向大模型披露个人信息,系统会将互动内容纳入数据库,这种交互式数据留存使每一名用户都可能在不自觉中成为敏感信息的“贡献者”。企业虽然声称实施匿名化与加密等安全保障措施来提高数据安全性,但正如上述 ChatGPT 因漏洞导致用户对话标题与支付信息被他人查看的事件表明,技术承诺并不构成可靠的法律安全阀。此外,大语言模型在运行中可能通过用户交互数据进行持续学习,由此引发的数据留存、二次利用等问题,对《个人信息保护法》第 13 条所确立的“最小必要原则”构成实质挑战。

生成式人工智能存在商业秘密泄露及引发不正当竞争的风险。主要存在两种可能,一是在交互中用户自行输入构成商业秘密的信息。企业将大语言模型接入内部办公系统后,员工在日常使用中若输入涉密业务的数据,会显著提升商业数据外泄风险。由于生成式 AI 的工作机理决定了用户输入内容可能被纳入后续训练语料,上述商业秘密信息存在被模型“记忆”并在响应其他用户查询时间接披露的风险。三星就曾因使用 ChatGPT 导致半导体相关机密数据被存入训练库。^[43]二是企业以“爬取”的不正当手段获取数据训练模型,所生成的数据产品在功能上实质性替代他人既有产品或服务,或导致相关公众对商品来源产生混淆,则可能违反《反不正当竞争法》第 6 条关于混淆行为的规定,或触发第 2 条一般条款的适用。^[19]这类生成式人工智能具备模型即服

务的特性, 日常运行会汇聚海量数据, 叠加平台资源优势在用户使用与模型迭代之间形成明显的“数据飞轮”效应, 数据流转规模庞大, 也让常态化的数据安全防护与治理面临严峻挑战。^[23]

2.2.2 算法风险

生成式人工智能的决策过程以深度神经网络依赖数十亿乃至数千亿参数在多层网络中的非线性运算。算法的技术性不透明使得具体执行过程难以为外部所知悉, 因此被称为算法黑箱。^[40]需要指出的是, 算法黑箱的出现是技术特性与人为因素共同作用的结果。一方面, 算法通过自主学习生成的规则体系在复杂程度上已经超出了人类的理解范畴, 即便是模型设计者也难以对特定输出结果做出因果层面的完整解释; 另一方面, 算法开发者通常会刻意隐匿决策规则, 使得整个决策过程缺乏透明度, 相关主体无法知晓其运作流程与内在逻辑。^[12]更为棘手的是, 生成式人工智能通常包含多重算法的嵌套与协同, 且模型参数处于持续迭代更新之中, 这使得即便在某一時点完成了算法审计, 其结论的时效性亦极为有限。

2.2.2.1 算法可解释性困境

算法黑箱首先侵蚀的是个人的知情权与异议权。《个人信息保护法》第24条第3款赋予个人要求说明并拒绝仅通过自动化决策作出的重大影响决定的权利, 然而说明义务的实现以算法可解释为前提。当模型设计者自身亦无法就特定输出提供因果性的完整解释, 则当事人的知情权与异议权将沦为具文。

其次, 算法透明与商业秘密保护之间存在权衡困境。模型架构、训练数据组合与超参数配置往往构成企业的核心竞争力, 强制公开可能削弱开发者的创新激励, 但不公开又使监管无从切入。目前尚无技术手段对黑箱算法进行全局解释, 而局部解释的可信性往往备受诟病。^[44]监管机构若不具备理解算法决策逻辑的技术能力, 则事后追责与事前预防均将失去着力点, 法律治理所依赖的可问责性基础将遭到根本动摇。

2.2.2.2 算法偏见与算法歧视

算法偏见是指人工智能系统在决策与预测过程中产生的系统性偏差, 其成因贯穿数据采集、标注、模型设计与优化的全流程。在数据层面, 训练语料中固有的社会偏见会经由模型学习被继承甚至放大。若训练数据在性别、种族、地域等维度上存在代表性不足或刻板化标注, 模型输出将不可避免地复现乃至强化这些偏见, 印证了“偏见输入则偏见输出”的技术规律。^[28]在模型设计层面, 算法选择和超参数设置等也属于技术黑箱的范畴。^[18]当开发者单纯追求预测准确率等效率指标而忽视公平性约束时, 模型可能在统计意义上对特定群体产生系统性不利影响, 形成无意识歧视。即便开发者主观上并无歧视意图, 其在效率、利润或风险控制等目标的优先安排, 也可能在结果上对特定群体造成不利影响。例如意大利户户送平台因骑手排名的算法不区分缺勤的合理由, 机械扣减评分并剥夺骑手优先接单的机会, 构成隐蔽的算法就业歧视。^[31]该案表明, 即便算法规则在形式上对所有用户一视同仁, 但若其设计未将公平价值纳入考量, 仍可能在实质效果上沦为侵害特定群体权利的歧视工具。此外, 行为人亦可能蓄意利用生成式人工智能低成本生产个性化内容的技术特点, 故意制作并传播带有偏见、歪曲事实的信息。这不仅会对个人和群体产生负面冲击, 还会降低建模的准确性, 对实现预期结果造成一定影响。^[7]

2.2.2.3 算法滥用风险

算法滥用的风险则直接威胁消费者自主决策权、公平交易权与市场竞争秩序。其典型形态包括算法操纵与算法共谋。算法操纵是指平台经营者利用算法技术对用户行为实施隐蔽性干预, 以实现自身利益最大化的行为。其典型表现是基于用户画像实施的个性化差别定价, 即“大数据杀熟”、诱导性界面设计和差异化信息推荐。在实践中, 平台利用特殊的算法, 对消费者的支付意

愿和价格敏感度进行精准画像,就同一商品或服务向不同消费者收取差异化价格,从中攫取消费者剩余的最大值。^[17]《个人信息保护法》第 24 条虽已明确要求自动化决策应当保证透明和公平,不得在交易价格等交易条件上对个人实行不合理差别待遇。但该条款在执行层面仍面临举证困难与认定标准模糊等障碍。

算法操纵的形式远不止于价格歧视。更广义的算法操纵风险包括算法被用于干预信息呈现、影响舆论走向、规避监管规则、对不同用户实施差别待遇。^[36]在信息推荐领域,算法操纵表现为平台对信息呈现顺序的刻意干预。以搜索引擎自动补足算法为例,自动补足算法提供的搜索建议并非中性地反映用户真实需求,而是嵌入了平台的商业利益考量,通过引导用户点击特定的建议从而增加点击量和热度。^[22]这种对信息环境的隐蔽塑造,实质上构成了对用户自主决策权的侵蚀。

算法对市场竞争秩序的另一重要威胁,来自算法共谋。算法共谋是指多个经营者通过使用相同或相似的定价算法,在无需直接意思联络的情况下实现价格协调的现象。数据驱动算法使经营者能够实时获取产品需求变化、竞争对手定价等信息,为不同经营者之间发现利益共同点并形成事实上的价格协调提供了技术条件。^[32]算法共谋对反垄断法的挑战在于传统垄断协议的认定以经营者之间存在明示或默示的“意思联络”为核心构成要件,但在算法定价场景中,价格趋同可能完全由算法自主学习与策略调整所驱动,经营者之间并不存在传统意义上的合意。这种“无合意的协调行为”是否以及如何纳入《反垄断法》的规制框架,构成了数字经济时代反垄断法教义学面临的重大理论难题。

2.3 输出阶段

2.3.1 知识产权面临被侵害的风险

生成式人工智能的输出内容依赖于对海量训练数据中统计规律的提前与概率化重组,其本质是基于用户指令,对数据库的信息进行定向筛选和重组后生成的结果。大陆法系多数国家仅承认自然人是智力成果的创作者,而现有法律尚未明确 AI 能否享有著作权。一方面,生成式人工智能输出内容本身亦可能构成对训练数据中原作品的侵权。该类侵权主要源于人工智能的开发者未获得版权所有人的授权许可,如未经版权人许可直接扒窃数据库或个人数据、未经授权将他人非数据作品数字化、未经著作权人许可创作新作品等一系列手段。这种未经授权的复制或改编行为,无论是直接将原始作品下载并存储,还是改变原作品的表现形式,都有可能侵犯著作权法中的复制权和改编权。

另一方面, AI 生成内容的权利归属难以界定,国际上也未形成统一标准。美国版权局在 *Zarya of the Dawn* 案中采取“逐要素审查”标准,认定申请人对 AI 生成图像的文字及图文编辑受著作权保护,但对纯粹由 Midjourney 生成的单幅图像拒绝登记,理由是生成图像不是受版权保护的原创作品。^[52]这一立场延续了 *Thaler v. Perlmutter* 案的裁判逻辑,法院明确指出“人类创作性是版权保护的基本要求”。^[51]依据我国《著作权法》第三条之规定,受保护的作品须是“文学、艺术和科学领域内具有独创性并能以一定形式表现的智力成果”。司法实践长期以自然人创作为基本前提,由此产生的核心问题是人工智能仅根据用户的提示词而自主生成的具体内容是否构成了著作权法意义上的“作品”。

我国司法实践已有初步探索。首例 AI 文生图著作权侵权案分析,北京互联网法院认为用户通过设定提示词和选择参数等行为体现了“智力投入”,构成以 AI 为工具的创作,用户享有包括署名在内的著作权。但我国学界对人工智能成果的可版权性还存在较大争议。

2.3.2 虚假信息

虚假信息是指未经调查和核实验证的、与客观事实不符或具有误导性的信息内容。当前生成式人工智能还不具备真正的逻辑推理与事实校验能力,其生成机制本质是基于文本预测,因而会

产生所谓“幻觉”现象，^[33]即生成一些看似合理但实则缺乏事实依据的信息，这种高度仿真的表现形式使人们难以辨别内容真伪。美国新闻可信度评估与研究机构 News Gurd 通过对 ChatGPT 的测试发现，ChatGPT 能在几秒钟内改变信息并产生大量很有说服力却无信源的内容。^[34]究其根源，大模型产生的不实信息往往是由算法偏差、数据质量欠佳或建模错误导致的，其本质是商业、社会等多重利益驱使的结果。^[6]

深度伪造技术更进一步，利用生成型对抗网络、扩散模型、生成预训练变换模型等算法的迭代升级，精准迎合人类的认知习惯，制造出极具欺骗性的影像与声音，侵害个人名誉权、肖像权，并扰乱社会秩序。例如，国内外已相继出现利用人工智能换脸、拟声技术伪装熟人身份实施大额财产诈骗的真实案例。美国联邦贸易委员会也专门就此发布风险预警，而多模态技术的持续成熟还将进一步提升信息欺诈的隐蔽性与成功率，对互联网内容生态安全与公众财产权益构成持续威胁。^[26]《生成式人工智能服务管理暂行办法》要求服务提供者对生成内容进行标识，并在发现违法内容时采取停止生成、停止传输等处置措施。然而，标识易被技术手段去除，经过多次转发后标识脱落几乎成常态，透明度义务的规范实效因而大打折扣。若对生成的内容缺乏有效的验证和多维度的交叉核验程序，只会放大传播风险的途径，进而使责任分配更加复杂化。^[16]

当人工智能生成内容造成损害，模型开发者、服务提供者与使用者三方主体的责任链条因果关系复杂。若将大语言模型简单视为工具，则用户作为提示词输入者需要承担过重的法律后果，尤其在简短提示即引发模型自主生成侵权或虚假内容的情形下，这一责任分配明显有违公平；若要求服务提供者承担类似网络平台的通知—删除义务，则又面临生成内容非由用户上传、无法预先识别的技术现实，类推适用《民法典》第1195条的制度基础并不牢固。而在虚假信息致损场合，模型“幻觉”引发的虚假内容属于系统内生的非故意风险，难以归入传统的过错责任框架，是否应引入危险责任或产品责任思路，各方分歧明显。更应注意的是，因知识产权产生的权属真空直接加大了虚假信息的追责难度。生成内容的权利归属不明也使主体注意义务不明，从源头上无法进行追责。

3 人工智能算法治理路径

生成式人工智能涉及上游研发、中游微调、下游部署等多个环节，其风险产生的技术逻辑呈现出“来源于数据-加剧于算法-溢出于模型”的链式传导特征。生成式人工智能的法律风险涵盖数据、算法、系统、人为四个层面且存在复杂的传导关系，这一多源性与关联性决定了任何静态的风险等级划定都可能因新技术的出现或既有技术的新应用方式而过时。因此，生成式人工智能应当构建事前数据可审查、事中算法可解释、事后责任可追溯的全链条全生命周期的治理体系，在动态调适中实现技术创新激励与风险防范之间的制度平衡。

3.1 事前预防机制

在准备阶段中，训练数据的合法性不足与质量失真造成了个人信息权益损害和错误信息、偏见性内容等核心问题。通过建立公共数据资源库，将数据来源合法性的审查机制和强化数据质量的同时纳入前置治理框架，以此来回应训练数据侵权、个人信息滥用以及模型偏差累积等源头性风险。

3.1.1 建立数据来源合法性审查机制

我国《生成式人工智能服务管理暂行办法》第七条只是原则性地提出需要提高训练数据质量。《互联网信息服务算法推荐管理规定》第八条也只概括规定了对数据实施定期审核制度，但均未就审查的标准、程序和方法作出可操作性的细化规定，无法为企业审核训练数据提供明确的指引。训练数据合法性的审查缺位将直接引发监管执法风险，甚至导致服务被整体叫停。例如，2023年3月，意大利数据保护局（Garante）以 OpenAI 缺乏处理个人数据的合法基础、未有效验证用

户年龄等为由,对 ChatGPT 发出临时禁令。^[47]

鉴于此,在日常的监督管理中应当根据数据获取方式的不同进一步细化对应的合规规制要求。一是针对自行采集的数据,研发者应当在数据处理前完成数据影响评估,划定数据处理的合理边界。同时建立用户分层授权机制,对于一般性个人信息,确保用户在明确知情的前提下同意数据被收集和使用,并提供随时撤销授权的选项。^[8]一旦涉及人格隐私或敏感信息,则必须取得用户的单独明示同意,并完善事前告知程序以保障信息安全。二是对于公开数据集的取用,需要区分承载个人信息与不含个人信息两类公开数据。按照《个人信息保护法》第28条和第29条确定的敏感个人信息认定标准,对生物识别、医疗健康等与人格尊严直接关联的敏感个人信息,遵循“特定目的+单独同意”的要求。对于借助网络爬虫技术采集的数据,应将采集范围严格限制在已合法公开的数据范围内,并尊重目标网站的 Robots 协议等技术限制措施。^[3]三是对于涉及著作权保护的训练数据,应当明确数据挖掘与训练使用的合法性界限。我国《著作权法》虽规定了合理使用制度,但其封闭式列举的立法模式难以涵盖大规模机器学习训练这一新型使用方式。对此,可以借鉴欧盟《数字单一市场版权指令》的制度经验,该指令为科研目的的文本与数据挖掘设置了强制性例外,同时赋予权利人对商业性挖掘的保留权,^[50]在促进技术创新与保护权利人利益之间寻求动态平衡。我国宜在立法中建立类似的法定许可制度,在合理补偿机制下允许基于训练目的的数据使用,而非一概禁止或一概放任。

3.1.2 强化训练数据质量控制与信用约束机制

数据质量直接影响模型输出的可靠性与安全性。在复现与迭代阶段,应对生成式人工智能的产出进行系统化的质量把控。一项可行的评判指标是检验开发企业是否愿意将模型生成的文本作为新模型的训练材料:若不能作为训练材料,则意味着生成文本较为模糊、训练价值低;若可以作为训练材料,则意味着生成文本与原始语料具有相近的信息密度和清晰度,此入手建立大型语言模型的数据质量评价机制。在制度设计上,可将数据质量评估纳入模型开发的前置程序,要求高风险模型在投入训练前完成数据偏差检测、标签一致性校验及敏感数据识别,并形成可供追溯的审查记录。借鉴欧盟《通用数据保护条例》中数据影响评估的这—种事前监管机制,不仅可以预先识别处理操作可能带来的潜在负面后果并采取措施有效减轻风险影响,还能实现对生成式人工智能从研发到应用的全生命周期的风险监控。^[4]

在制度保障层面,应当将数据信用纳入企业信用评级指标体系。围绕企业数据采集违规频次、数据存储泄露事件数量、用户针对生成内容质量的投诉率以及问题整改成效等数据质量与安全维度,采用定性定量相结合的方式评定企业信用等级。并将信用等级与算法备案审查、行政许可续展等行政程序挂钩,形成信用约束的闭环。此外,开发者的商业秘密应获得制度性保护。当数据来源和类型涉及核心商业秘密时,允许开发者在合规审查中进行保密处理,但应由审查机构在保密协议框架下进行核验,避免以保密为由规避审查义务。^[25]

3.1.3 公共训练数据资源的协同供给

目前我国合成数据生成产业的发展进度滞后于生成式人工智能产业的整体需求。《生成式人工智能服务管理暂行办法》第六条中提出了建设公共训练数据平台的发展保障措施,为构建政府引领的数据供给模式提供了政策依据。行政机关可以在医疗、教育、交通等重点领域优先搭建公共训练数据平台,通过自主开发或政府采购的方式,形成基础性公共数据资源库,为行业发展提供高质量、合规可溯源的训练数据供给,从源头降低企业因数据获取违规而引致的法律风险。

明确数据挖掘和训练数据使用的合法性界限。应建立更加明确的法定许可制度,确保在生成式人工智能训练过程中,版权人能够得到合理的补偿,而使用方能够在合规的框架内获取数据。美国通过判例法明确了在特定情况下“转换性使用”可以构成合理使用,为 AI 技术的应用提供了灵活性和法律支持。^[21]

3.2 事中监管机制

生成式人工智能的算法黑箱以及数据聚合效应导致模型训练的过程和决策逻辑长期处于不透明状态,需要构建起以算法审计、算法备案与透明度的提升体系回应运行阶段的风险放大,确保模型训练及决策过程的合意性和可解释性的同步达成。

3.2.1 算法审计制度的系统化建构

算法审计的核心功能是由独立主体对算法的合规性、公平性与安全性进行系统化的检验与评价,为监管执法与公众监督提供专业判断的基础。同时回应了算法偏差、歧视性输出和安全失控等深层风险。在审计模式上,应采取内外协同的双轨机制。在内部审计层面,企业运用代码审计、非侵入性测试、对抗性攻击模拟等技术方法,对算法逻辑、训练数据偏差与业务流程风险进行持续性自查,及时修正自主学习过程中出现的偏离设计初衷的问题。在外部审计层面,引入具备相应技术资质的第三方专业机构参与社会监督。外部审计机构要依照技术中立的原则出具合规认证报告,并且要注意协调特定技术标准的垄断化趋向,防止审计认证沦为市场准入壁垒。

3.2.2 强化算法可解释性的分层披露义务

算法解释权是连接技术黑箱与法律追责的关键制度节点。然而,算法可解释性并非一个无差别的整体性概念,而应当根据信息接收主体的不同需求进行分层设计,实现精准透明。

面向受影响的个体用户,应赋予受自动化决策不利影响的相对人请求解释、更正数据、申请救济的权利,由算法使用者承担解释义务,开发者予以技术协助,借助可解释性技术破解算法黑箱。^[20]通过在软件框架中增加可解释技术接口,或提供指标评估模型等增强算法可信解释能力,使用户能够理解算法决策的基本逻辑并做出有意义的选择。^[24]我国《个人信息保护法》第24条第3款已赋予个人对自动化决策“拒绝权”与“说明要求权”,但该条款的实施细则尚未明确,亟需通过配套规定将解释义务的触发条件、解释内容的最低标准和时限要求予以具体化。面向监管机构,算法开发者应提供算法的技术细节、执行日志等材料,使监管者能够全面了解算法运行状况,为执法提供事实基础。^[35]面向独立审计机构,透明度的核心是可验证性。在保密协议框架下开放算法的运行数据与关键参数,使第三方能够独立检验算法的公平性与安全性。这一分层设计在保障透明度实质功能的同时,也回应了企业对商业秘密保护的合理关切。

就透明度的内容标准而言,我国需要通过立法划定算法说明义务的量化基准与边界。算法研发主体负有主动信息披露责任,应以通俗可理解的方式说明算法的运行规则与基本原理;同时需主动披露算法运作机理、参数来源,酌情公开测试样本与研发考核标准,切实保障公众的知情权与监督权。

3.2.3 明确算法备案法律属性是风险治理的制度根基

生成式人工智能算法备案应定性为行政告知式备案,摒弃行政许可式的实质审查与行政确认式的法律事实认定模式,监管部门仅对备案材料做形式审查,通过信息收集、存档与公示实现监管目的。^[2]这一属性定位的制度意义在于,一方面避免过度审批抑制技术创新,回应组织逻辑困境中监管资源有限的现实约束;另一方面通过信息收集打破企业与监管方、公众的信息不对称,以告知承诺机制倒逼主体自我规制,从源头减少算法滥用与黑箱操作的风险诱因。

在制度运行上,备案应与风险分级制度形成联动。对高风险生成式AI强制备案,对低风险主体鼓励自愿备案并简化流程;针对不同算法类型制定差异化备案要求,优化线上审核流程,提供答疑指导与一次性补正告知,降低企业备案成本。尤为关键的是,强化备案信息的实质性公开,要求备案主体用通俗语言公示算法原理与运行机制,并建立动态更新机制,匹配算法快速迭代的技术特性,使备案制度真正发挥透明化治理的功能,而非沦为形式化的行政程序。^[11]

3.3 事后追责与权利救济

由于人工智能技术所涉及的利益方众多,权利义务关系复杂,加之算法黑箱的不可解释性,传统侵权法的归责逻辑面临严峻挑战。面对个人信息权和知识产权被侵害、数据泄露等违法行为,存在着责任主体识别困难、因果关系证明受阻、责任承担难以落实等困境。在这一背景下,需要重新审视归责原则、合理分配举证责任以及多元救济渠道来确定事后追责体系。

首先,建构分层分类问责体系。责任的明确划分是预防体系得以有效落实的关键环节,通过清晰地界定技术链条上各主体的法律义务,可以确保预防措施从技术研发到数据管理的每个环节都被严格执行。要求数据提供者去报数据来源的合法性和合规性,定期接受合规性审查和技术能力评估,以确保其履行数据安全责任的能力。在立法中明确技术开发者加强对数据和算法安全的保障,制定安全技术标准进行全流程风险评估;用户对恶意提示、违法利用、输入他人商业秘密或个人信息等行为承担相应责任。^[9]其次,重构举证责任分配规则。对于经算法影响性评估认定为高风险的决策场景,适用举证责任倒置,由责任主体全面证明其决策过程的合法性与合理性,将技术不透明的不利后果完整地分配给信息优势方,打破传统的“谁主张、谁举证”的框架。最后完善多元纠纷解决与救济机制。在问责体系中组建由技术专家、法学专家与公众代表构成的第三方独立机构,作为常态化问责的监督载体,专门受理有关人工智能生成内容侵权、数据合规争议与平台责任纠纷等问题。同时,监管机构应在技术可以实现的范围内,为信息泄露的通知与处理设置明确的时间点。

综上,生成式人工智能治理路径的建构不是对孤立制度工具的简单叠加,而是以风险分级为制度中枢,将事前数据合法性审查与影响评估、事中透明度与审计监测、事后问责与权利救济贯通为有机整体,形成从标准约束到流程管控再到责任倒逼的完整风险治理闭环,为生成式人工智能算法的规范发展提供系统性的制度保障。

4 结语

生成式人工智能带来技术革新的同时也带来了数据安全、隐私保护、算法黑箱、知识产权等一系列法律风险。我国已初步建立人工智能综合治理体系,但其应用场景、商业模式与风险样态都具有显著的不确定性,这决定了任何治理框架都不能一劳永逸。未来立法与监管宜在坚守安全底线的同时,为制度保留必要的弹性,通过风险分级分类、动态评估与责任明确,缓解技术发展带来的法律风险,实现技术与安全的平衡。

参考文献:

- [1] 丁晓东.论算法的法律规制[J].中国社会科学,2020(12):105.
- [2] 王丽熔.生成式 AI 下的算法备案:功能定位、现实困境与优化路径[EB/OL].[2026-06-03].<https://doi.org/10.26914/c.cnkihy.2024.024063>
- [3] 王海洋.生成式 AI 训练数据的法律风险及其元规制[J].浙江社会科学,2024(9):59.
- [4] 王海洋.生成式 AI 训练数据的法律风险及其元规制[J].浙江社会科学,2024(9):60.
- [5] 卢安文.生成式人工智能:风险、监管与治理模式探索[J].重庆邮电大学学报(社会科学版),2025(3):115.
- [6] 卢安文.生成式人工智能:风险、监管与治理模式探索[J].重庆邮电大学学报(社会科学版),2025(3):114-115.
- [7] 卢安文.生成式人工智能:风险、监管与治理模式探索[J].重庆邮电大学学报(社会科学版),2025(3):115.
- [8] 刘洋,宋振武.人工智能算法立法的理论逻辑及其规范构造[J].天津法学,2024(2):50.
- [9] 刘洋,宋振武.人工智能算法立法的理论逻辑及其规范构造[J].天津法学,2024(2):51.

- [10] 孙逸啸,彭成龙.生成式人工智能算法"失控": 风险景象、治理困境与规制进路[C]//新兴权利集刊.上海: 上海人民出版社,2025:35.
- [11] 孙逸啸,彭成龙.生成式人工智能算法"失控": 风险景象、治理困境与规制进路[C]//新兴权利集刊.上海:上海人民出版社.2025:41-42.
- [12] 毕文轩.生成式人工智能的风险规制困境及其化解: 以 ChatGPT 的规制为视角[J].比较法研究, 2023(3):158.
- [13] 毕文轩.生成式人工智能的风险规制困境及其化解: 以 ChatGPT 的规制为视角[J].比较法研究, 2023(3):158.
- [14] 李西岭.生成式人工智能中的个人信息保护[J].法学论坛,2026(3):33.
- [15] 李智,赵雪蓉.生成式人工智能大模型的数据风险与治理机制[J].上海政法学院学报 (法治论丛), 2026(3):146.
- [16] 李智,赵雪蓉.生成式人工智能大模型的数据风险与治理机制[J]. 上海政法学院学报 (法治论丛), 2026(3):148.
- [17] 杨成越,罗先觉.算法歧视的综合治理初探[J].科学与社会,2018(4):5.
- [18] 杨建武,罗飞燕.类 ChatGPT 生成式人工智能的运行机制、法律风险与规制路径[J].行政法学, 2024(4):106.
- [19] 杨建武,罗飞燕.类 ChatGPT 生成式人工智能的运行机制、法律风险与规制路径[J].行政法学, 2024(4):108.
- [20] 张凌寒.算法规制的迭代与革新[J].法学论坛,2019(2):26.
- [21] 张梦洁.生成式人工智能训练数据的著作权风险及规制路径[EB/OL].[2026-06-01].<https://doi.org/10.26914/c.cnkihy.2025.020512>
- [22] 张凌寒.搜索引擎自动补足算法的损害及规制[J].华东政法大学学报,2019(6):40.
- [23] 张欣.生成式人工智能的数据风险与治理路径[J].法律科学 (西北政法大学学报), 2023(5):46.
- [24] 陈敏,孙占利.算法"黑箱"的技术与法治耦合治理[J].法治论坛,2022(4):17.
- [25] 陈宏光,安诗凤.生成式人工智能训练数据的风险与治理研究[J].海南开放大学学报,2024(1):115-116.
- [26] 宋保振,秦瑞标.论生成式人工智能应用风险的法律规制[J].上海政法学院学报 (法治论丛), 2023(6):113-114.
- [27] 宋保振,秦瑞标.论生成式人工智能应用风险的法律规制[J].上海政法学院学报 (法治论丛), 2023(6):115.
- [28] 岳平,苗越.社会治理: 人工智能时代算法偏见的问题与规制[J].上海大学学报 (社会科学版), 2021(6):1-11.
- [29] 岳平,苗越.社会治理: 人工智能时代算法偏见的问题与规制[J].上海大学学报 (社会科学版), 2021(6):1-11.
- [30] 金融界.ChatGPT 开发商 OpenAI 遭集体起诉: 不择手段窃取信息, 将导致文明崩溃! [EB/OL].[2026-05-30]https://www.sohu.com/a/692637279_114984.
- [31] 罗智敏.算法歧视的司法审查: 意大利户户送有限责任公司算法歧视案评析[J].交大法学,2021(2):183-195.
- [32] 胡小伟.人工智能时代大数据算法风险管理的法律监管研究[J].湖北大学学报 (哲学社会科学版), 2021(2):129.
- [33] 姚微佳.生成式人工智能风险治理及法律规制研究[J].湖北经济学院学报 (人文社会科学版), 2024(12):99.
- [34] 斜晓东.论生成式人工智能的数据安全风险及回应型治理[J].东方法学,2023(5):106-116.
- [35] 徐伟,何野.生成式人工智能算法风险的法律治理[J].新媒体与社会,2024(2):69.

- [36] 徐伟,李文敏.生成式人工智能法律风险的源头治理[J].政府与法治,2025(2):95.
- [37] 袁合川、吴艳弘.生成式人工智能的版权风险及其规制路径[J].传播与版权,2025(13):107.
- [38] 商建刚.生成式人工智能风险治理元规则研究[J].东方法学,2023(3):6.
- [39] 黄镔.生成式 AI 对个人信息保护的挑战与风险规制[J].现代法学,2024(4):105.
- [40] 程雪军.AIGC 浪潮下超级人工智能平台算法黑箱的治理路径[C]//新兴权利集刊.上海:上海人民出版社,2023.
- [41] 童星海.生成式人工智能数据训练中个人信息保护的责任规制[J].中国科技论坛,2026(3):23.
- [42] 熊进光、张崢.从 ChatGPT 到 DeepSeek: 生成式人工智能的法律风险与三维规制[J].重庆大学学报 (社会科学版),2025:1.
- [43] 澎湃新闻.员工 20 天三起泄密事故,三星考虑重新禁用 ChatGPT[EB/OL].[2026-05-30] https://www.thepaper.cn/newsDetail_forward_22568617.
- [44] 戴维,王锡铎.算法透明机制的局限性及其克服[J].华东政法大学学报,2025(1):38-39.
- [45] 戴维,王锡铎.算法透明机制的局限性及其克服[J].华东政法大学学报,2025(1):38-39.
- [46] Dennis Hillemann, Does ChatGPT Comply with EU-GDPR Regulations? Investigating the Right to be Forgotten[EB/OL].[2026-06-01].<https://www.fieldfisher.com/en/insights/does-chatgpt-comply-with-eu-gdpr-regulations-inves>.
- [47] Garante per la protezione dei dati personali. Provvedimento del 30 marzo 2023, No. 9870 832.
- [48] Jie Gui, Zhenan Sun, Yonggang Wen, Dacheng Tao & Jieping Ye, A Review on Generative Adversarial Networks: Algorithms, Theory, and Applications, 35 IEEE Transactions on Knowledge and Data Engineering 3313, 3313-3332 (2023).
- [49] Mata v. Avianca, Inc., 22-CV-1461 (PKC) (S.D.N.Y. 2023).
- [50] Regulation (EU) 2024/1689 of the European Parliament and of the Council (AI Act), Art. 3-4.
- [51] Thaler v. Perlmutter, No. 23-5233 (D.C. Cir. 2025).
- [52] USCO. Re: Zarya of the Dawn (Registration # Vau001480196)[EB/OL].[2026-06-14].

Risk Analysis and Governance Pathways in the Operation of Generative Artificial Intelligence

Li Yuhan*¹

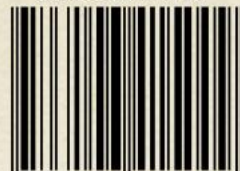
¹ Macau University of Science and Technology, 826994071@qq.com

Abstract: Generative artificial intelligence, as represented by ChatGPT, has pushed AI from assistive analysis toward autonomous generation, while also causing legal risks to spread across operational stages, transmit among different actors, and compound across domains. This article conceptualizes the operation of artificial intelligence as a dynamic process in which risks are continuously generated, accumulated, and externalized, and systematically examines it across three stages: data collection, algorithm training, and content output. The study argues that AI risks exhibit a chain-transmission logic of “originating in data, intensifying in algorithms, and spilling over from models,” which is concretely manifested in the lack of legitimacy of training data sources, distortion of data quality, algorithmic

black boxes and insufficient transparency, intellectual property infringement, dissemination of false information, and difficulties in identifying responsible subjects. At the output stage, unclear intellectual property ownership and the rupture of the liability chain for false information are intertwined, exposing the difficulties of applying traditional tort law and platform liability rules in generative AI scenarios. In light of this, a full-chain governance system covering ex ante prevention, in-process supervision, and ex post remedy should be established. The key to AI governance lies not in simply constraining technological space through static rules, but in seeking an institutional balance between innovation incentives and risk prevention through dynamic adjustment. This balance must also be embedded in the three links of data sources, algorithmic processes, and output impacts in order to respond to the normative challenges arising from the chain transmission of generative AI risks.

Keywords: Generative Artificial Intelligence, Operational Risk, Algorithmic Governance, Governance Path

ISSN 3106-4159



9 773106 415009 >