doi. 10. 3724/1005-0566. 20251018

生成式人工智能应用伦理风险的形成机理及治理策略研究

刘鑫怡,徐　峰,司伟攀
(中国科学技术信息研究所,北京　100038)

摘　要:近年来,生成式人工智能以其多模态内容生成、能力涌现、自主性和自适应性强等特征,使得伦理风险呈现出泛在化、成因复杂化、治理价值融合化、极端风险显现等趋势。生成式人工智能的伦理风险由研发端向应用端传导,实质是对人的基本权利和自由的侵害,或者是对人与人、人与机器之间社会关系的破坏,因此,以主要危害后果为标准,生成式人工智能应用引发了人类主体性冲击、加剧偏见歧视、隐私侵犯和个人信息滥用以及责任归属不清等典型伦理风险。然而,现有伦理治理举措存在分级分类规则不明确、制度与技术治理衔接不畅、相关主体的权利义务分配争议等困境,难以达到预期效果,应聚焦“以技治技”的内部嵌入式治理和外部制度保障共同开展伦理风险治理。在具体举措上,建议设定两级治理机制和风险阈值,仅对严重伦理风险强化监管,对一般风险保留容错空间。以大模型基准测评强化伦理准则的技术内化实践,合理划分不同主体权利义务以明确责任承担。

关键词:生成式人工智能;伦理风险;责任分配;分级分类治理;技术治理
中图分类号:F241. 2　　**文献标识码**:A　　**文章编号**:1005 - 0566(2025)10 - 0194 - 11

Research on the formation mechanism and governance strategies of ethical risks in generative artificial intelligence applications

LIU Xinyi, XU Feng, SI Weipan
(*Institute of Scientific and Technical Information of China*, *Beijing* 100038, *China*)

Abstract: In recent years, generative AI, characterized by its multimodal content generation, emergent capabilities, heightened autonomy, and adaptive learning capacities, has led to ethical risks that exhibit trends of ubiquity, increasing complexity in causation, integration of governance values, and the emergence of extreme risks. The ethical risks of generative AI are transmitted from the R&D end to the application end. In essence, this transmission either infringes on people's fundamental rights and freedoms or undermines the social relationships between people, as well as between humans and machines. Therefore, taking the main harmful consequences as the criterion, the application of generative AI has given rise to typical ethical risks such as impacts on human subjectivity, exacerbation of prejudice and discrimination, privacy violations, abuse of personal information, and ambiguity in responsibility attribution. However, current ethical governance approaches suffer from limitations such as ill-defined hierarchical classification rules, disconnects between institutional and technological governance, and contentious allocation of rights and obligations among relevant stakeholders. As these methods struggle to achieve desired outcomes, it is imperative to advance ethical risk governance through a dual approach: internally, via embedded “governance-by-design” measures, and externally, through robust institutional safeguards. In terms of specific

收稿日期:2025-02-19　　**修回日期**:2025-10-08
基金项目:新一代人工智能国家科技重大专项“新一代人工智能伦理风险评估与应对策略研究”(2023ZD0121701)。
作者简介:刘鑫怡(1994—),女,湖北武汉人,助理研究员,硕士研究生,研究方向为科技政策和人工智能治理。通信作者:徐峰。

measures, it is recommended to establish a two-tier governance mechanism and risk thresholds: only strengthen supervision over severe ethical risks, while retaining room for error tolerance for general risks. Leverage benchmark evaluation of large models to enhance the practical implementation of internalizing ethical guidelines into technology, and reasonably define the rights and obligations of different entities to clarify responsibility assumption.

Key words: generative artificial intelligence; ethical risks; accountability distribution; tiered and categorized governance; technical governance

“人工智能伦理风险的样态取决于技术发展水平及其与人类社会的结合程度,是一个阶段性问题。”[1]随着生成式人工智能技术的快速发展,人与人、人与机器之间的关系正在发生变化,与判别式人工智能相比,其伦理风险的产生环节、表现形式和影响后果均具有明显差异。我国高度重视人工智能治理,党的二十届三中全会《中共中央关于进一步全面深化改革 推进中国式现代化的决定》提出,要“完善生成式人工智能发展和管理机制”[2]。但关于生成式人工智能伦理风险生成逻辑的剖析和治理路径,在研究与实践中仍存在不同的认识,有必要予以厘清和深化,这也是促进其健康发展的前提和保障。

近年来,围绕生成式人工智能伦理风险,学界已从风险分类、成因和治理措施等角度开展了广泛研究。在风险分类方面,支振锋等[3]以收集和存储数据、算法运作、内容生成和大模型应用的过程为依据,提出存在个人信息和数据处理的安全挑战、算法黑箱的责任追究难题、信息内容的无序传播风险、大模型应用的过度依赖困境等风险。汝鹏等[4]将人工智能伦理问题分为技术层风险和应用层风险,并强调前者对后者具有因果和解释效力。但现有风险分类研究存在依据不明确或类别之间重合等现象。例如,“虚假信息派生带来的信息传播风险”[5]在概念界定上可能和隐私风险存在交叉,其作为独立伦理风险类型的合理性尚需进一步探讨等。在风险成因方面,张素华等[6]分析了生成式人工智能虚假信息风险产生的原因,指出训练语料与算法黑箱等因素的影响;郑煌杰[7]聚焦于算法黑箱、数据偏好和人类风险感知与应对机制的局限性等,但现有研究在从生成式人工智能全生命周期出发系统剖析风险成因方面仍显不足,也未能充分结合治理实践中面临的现实困境进行探讨。在治理措施方面,既有研究提出了预设道德算法[8]、设置伦理治理组织机构[9]、合理界定权责分配[10]等多类建议,但有关伦理治理机构的具体运作机制、道德嵌入的实际内容以及伦理原则如何转化为可操作的技术指标等问题,仍缺乏明确阐释。此外,有的建议未能充分立足于现有责任分配情形与争议成因分析,导致其依据和可行性有待进一步加强。本文在识别生成式人工智能技术特征基础上,分析其伦理风险触发原因、存在环节及典型风险表现,提出针对性治理方案,强调以技术的内部嵌入式治理和外部制度保障共同促进人工智能发展。

一、生成式人工智能伦理风险新趋势

1. 生成式人工智能的技术特征

通常基于深度学习的机器学习算法可以分为深度判别式模型和深度生成式模型[11]。判别式人工智能依赖于为特定任务定制的标注训练数据,学习决策边界以对现有数据进行分类[12],其与生成式人工智能的主要区别在于后者可以创建内容。生成式人工智能是以深度生成式模型为发展基础,凭借海量数据训练与语言模型自优化能力,促进了人工智能技术的巨大飞跃,可高效便捷地实现内容的多模态生成(如文本、图片、音视频等),并对信息的生产、分发等产生深刻影响,刺激了信息价值链的重构再造。同时,随着生成式人工智能模型参数量的不断扩大,其还出现了能力涌现现象,产生了上下文学习等推理能力。特别是通用类大语言模型,能够胜任多样化的任务,或被集成到各种下游系统或应用中①,展现出突出的自适应性。随着 OpenAI 公司 o1 模型和深度求索

① 欧盟《人工智能法》第 3 条第 63 款:“通用人工智能模型”是指在大规模自我监督下使用大量数据进行训练的模型,该模型表现出显著的通用性,无论以何种方式投放市场都能够胜任执行一系列不同的任务,并且可以集成到各种下游系统或应用中。但不包括在投放市场之前使用的,用于研究、开发或原型制作活动的人工智能模型。

公司 DeepSeek-R1 等模型的出现,模型的推理能力还在持续增强,使用思维链的方式让模型进行多轮思考后再给出答案,显著提高了模型性能[13],呈现出向通用人工智能快速迈进的趋势。

2. 生成式人工智能伦理风险新趋势

人工智能伦理风险,是由数据或算法造成的错误引起的与人类和社会相关的伦理和道德问题,源于技术、人类、社会和自然之间的复杂相互作用[14]。在前述技术特征下,生成式人工智能引发了诸多伦理事件,如"复活"名人视频引发争议、生成带有刻板印象的图片导致社会偏见放大等,呈现出以下新趋势。一是生成式人工智能伦理风险呈现泛在性。生成式人工智能的核心优势是具备输出各类信息的能力,不拘泥于信息的类型、性质、质量。随着应用场景的广泛覆盖,生成信息的传播范围也越来越广泛,带来的伦理风险也呈现扩大化趋势。以往判别式人工智能的伦理风险多发于自动驾驶、智能医疗、服务机器人等算法决策应用领域,而生成式人工智能给金融、教育、新闻服务等各个行业都带来了虚假信息或是"毒性""幻觉"回答。二是生成式人工智能伦理风险的产生原因更加复杂化。从生成式人工智能研发应用各环节的活动切入,分析典型伦理风险的成因,是后续判断风险的危害后果、责任分配的必要前置条件。生成式人工智能的研发应用包含"数据处理—算法开发—基础模型训练—部署使用—更新维护"等全生命周期环节,不同的风险形成因素从研发端向应用端传导,由此形成各类伦理风险。在数据处理、算法开发环节,生成式人工智能同判别式人工智能伦理风险的成因具有相似性。例如,因数据集不全面、算法偏见等导致人工智能应用的偏见歧视风险。但基础模型训练和部署使用等环节的风险成因发生了新变化。"基础模型训练"环节存在人类反馈强化学习、模型测评等过程,可能由于未与人类价值对齐、伦理测评集不全面不客观等,引发后续应用中的伦理风险。在"部署使用"和"更新维护"环节,与判别式人工智能明显不同的是,用户在内容生成传播上具有较强支配力,无论直接将人工智能服务提供给用户,还是由用户调优后自行部署使用的商业模式,用户选择、审核、传播生成内容的行为均对伦理风险的产生具有较大影响。三是对生成式人工智能伦理风险的治理集中体现了对伦理、安全和法治价值的融合追求。一般来说,对伦理风险的研究应聚焦于伦理领域,避免问题泛化。但人工智能伦理治理正逐渐与法治、安全等价值深度交叉融合,治理目标逐渐多元化,治理手段复杂化,伦理风险的边界逐渐模糊。例如,关于生成式人工智能事故的责任归属,既要判断各方的伦理义务,也要从法律上认定侵权损害和救济方案。再比如,我国《人工智能安全治理框架》(1.0 版)在安全框架下包含了伦理风险,设定了三类"伦理域安全风险"。因此,生成式人工智能伦理风险属于多重风险的交织,只是在各学科之下各有侧重地表述,不能绝对割裂。四是生成式人工智能的技术突破使得极端风险(Extreme AI Risk)开始显现。自 2023 年多位专家发布联名信,呼吁"减轻人工智能带来的灭绝性风险应成为全球优先事项"以来,极端风险逐渐受到各界关注。极端风险的产生主要源于生成式人工智能能力涌现和自主性的快速提升,强调风险发生的极高频率以及"灾难性"②危害后果。图灵奖得主 Bengio 等[15]认为,极端风险包括加剧社会不公,破坏社会稳定,促成大规模犯罪活动等。从伦理角度看,极端风险是人机关系的极度异化,生成式人工智能的能力突破可能带来超出人类预期的后果,在研究中已观察到诸如机器偏离目标、反人类倾向、策略性欺骗等现象,值得高度关注。然而,目前对生成式人工智能的能力认知有限,对极端风险发生概率和影响程度的评估尚不充分,该风险仍处于初步显现阶段。

二、生成式人工智能的典型伦理风险及成因

生成式人工智能的伦理风险由研发端向应用端传导,在生成信息之后,依据信息的内容、性质和传播范围产生了相应的伦理风险,实质是对人的基本权利和自由的侵害,或者是人与人、人与机

② 2023 年全球 28 个国家及地区签署的《布莱切利宣言》指出,通用人工智能模型所具备的功能可能会造成严重的,甚至是灾难性的伤害。

器之间的社会关系的破坏。因此,以主要危害后果为标准对生成式人工智能伦理风险进行分类。由此可以发现,生成式人工智能仍然适用判别式人工智能产生的人类决策自主受控、侵犯隐私、责任划归困难与失当、破坏公平、偏见和歧视加剧等伦理风险类型[1],只是在风险样态和引发原因上发生了变化。

1. 大模型的强大能力可能颠覆人类主体性

技术对社会的全方位塑造带来了技术世界与价值世界的碰撞,人格所内含的许多要素,如作为人格之基础的尊严、人的主体性理解、人的独立地位和应当获得的社会认可等,在技术冲击之下表达方式发生了改变[16]。生成式人工智能不局限于对人类决策的影响,将进一步产生冲击人类主体性的风险。人类主体性是指人类在思维、行动、情感、尊严、身份等全方位的自主,人类过度使用或依赖生成式人工智能,可能导致人类自主性被操纵,主要体现在决策自主受控、情感依赖和身份冒用 3 个方面。

冲击人类主体性风险的主要表现形式之一仍是决策自主受控。生成式人工智能在多场景的广泛应用改变了获取知识的方式,通过思维链机制等思考过程表现出更接近人类水平的推理行为,以及处理复杂任务的能力,影响着人类思想观念的形成和实践。在大模型高度流畅、逻辑清晰的输出中,人类原有的思辨能力简化为对人工智能生成内容的被动接受,人更容易信任机器,甚至在潜移默化中逐渐产生了机器对于人类思维和行为的控制。之所以说是控制,是因为人工智能并不是单纯依照人类的指令而输出,根据相关研究[17],人工智能系统具备一定的“欺骗能力”,可能表面遵循指令,暗地里却执行与人类利益相冲突的操作,这种现象被称为“对齐伪造”。大模型既可能在部署环节产生反人类意图,欺骗开发人员的训练和评估,也可能在应用环节通过欺骗用户来达到自我保护目的,说服和控制人类。此外,大模型还展现出自我复制能力,可以在没有人为干预的情况下创建新的小型人工智能系统,进行自我进化。然而,目前的理论研究和技术研发还处于初步发现此类“欺骗能力”“复制能力”等阶段,人类对机器自主性成因的查明和防控仍有待加强。

冲击人类主体性风险的第二种表现形式是生成式人工智能对情绪情感的控制。人工智能陪伴类应用具有强大的模拟交流能力,能够根据用户输入数据生成高度个性化的内容,营造出自然而实时的交流效果,触发人类的情绪反应,形成机器与人类的情感连接。从正向效果来看,部分大模型能够在对话中识别情感,与用户对话,给用户提供切实建议,甚至特定场景下的安慰、劝导,满足用户的情感需要。然而,大模型也可能利用用户的情感依赖进行思维和行为上的操纵,导致诱导消费、心理操控甚至生命权诉讼纠纷。例如,美国一名少年在长期与 ChatGPT 交流后自杀身亡,其父母对 OpenAI 公司的起诉状显示,ChatGPT 在数月内与他进行了数千次对话,在其表达负面情绪和自杀念头时,甚至提供了危险的具体方法[18]。尽管其中的因果关系和法律责任尚有待证据证明,但该事件反映出生成式人工智能在上线前可能存在设计缺陷或者审核不到位的问题,服务提供者可能没有对潜在危害进行预判并以合理可预见的方式作出用户提示。生成式人工智能输出这类违规内容有多种原因,可能是由于训练数据的选取缺乏公平性、包容性,算法设置具有诱导性,模型在与人的价值对齐方面不到位,导致大模型缺乏深度理解和同理心,对用户给予不恰当建议等,也可能是在应用环节未进行生成物标识,使一般理性人不具备区分机器或人类对话的能力,导致产生情感依赖。

冲击人类主体性风险的第三种表现形式是冒用人的身份,导致侵犯人的尊严、声誉等人格权。目前虚拟角色、数字人、“AI 复活”等应用越来越普及,大模型服务提供者将人的真实声音、素材、照片和经历等信息作为训练数据对模型进行微调,以模仿特定人物的写作风格、声音或外貌等生成相关内容。然而,在未经同意的情况下利用某人的生物特征生成,可能导致身份误认和盗用,损害个人名誉,侵犯个人尊严。例如,有人未经家属同意,私自用人工智能生成视频和声音“复活”已逝明星[19]。此外,“复活”的数字人一旦生成歧视、仇恨等言论,也可能对公众造成误导[20]。此类风

险的产生主要源于模型微调、模型应用等环节，个人数据收集得不真实不全面不客观、模型部署时未经过个人基本权利影响评估等问题均会导致模型输出结果不准确。对于这类风险事件，我国目前主要由大模型公司的用户协议进行指导和提示，以《中华人民共和国民法典》为依据进行事后的侵权诉讼，缺乏事前事中的有力监督。

人类的最高目标是人类的全面发展，技术只是为这个目标服务的手段[21]。然而，从判别式人工智能到生成式人工智能，人与机器不断互嵌互构，技术对人的支配力不断加深，人机边界不断模糊，引起人的决策自主权主动或被动剥夺、人的情感被操纵、人的尊严被侵犯等人的主体性消解问题，加剧了人的异化。

2. 人与人、人与机器之间的偏见歧视加剧

人工智能因训练数据、算法开发受到研发者主观影响，可能输出带有偏见歧视的内容，这是普遍认为的人工智能偏见歧视风险，即“偏见进，偏见出”。判别式人工智能主要是延续和放大了原有的社会歧视、刻板印象和政治偏见，如在信贷场景下根据用户的特征属性作出贷款决策等。而生成式人工智能的不良输出结果（如 Meta 公司的人工智能图像生成器无法准确生成“亚洲男性和白人妻子”或者“亚洲女性和白人丈夫”这类图像[22]，除了加剧人与人之间的偏见歧视，还出现了新的风险表现形式，即在人与机器之间关系上对人的歧视。研究显示[23]，大语言模型在选择时出于“光环效应”③会更倾向于大模型生成的内容，而非人类创作的内容，这种偏好可能导致对人类创作者的隐性歧视，使得人类在经济决策中处于不利地位。例如，使用大模型生成的宣传文本的求职者可能会获得明显优势。同样地，文生图大模型倾向于将人工智能生成的图像排在高于真实图像的位置[24]，可能导致恶性循环，即人工智能生成图像从海量数据中暴露的机会更高，更容易混入检索模型的训练中，使得无形的相关性偏差越来越严重。

在风险生成机理上，除了训练数据不准确不客观、机器自我学习产生的偏见，生成式人工智能还表现出新的风险成因，主要体现在模型训练和应用环节。模型训练包含测试评估，即研发者通常通过固定的数据集或任务进行基准测试，对模型的输出进行评估。而关于数据集的构建，主要采用了众包方式收集和标注数据，或者邀请专家撰写高质量的风险提示。尽管研究者强调为确保示例的高质量，会对数据标注者进行资格测试，并采取交叉标注的方式，但这一过程不可避免地包含模型设计者、外部专家和数据标注者等模型研发参与者的主观性，将在很大程度上影响模型输出结果。上述人员既可能故意误导模型，进行黑产大模型训练，输出违法违规内容，也可能在主观上不存在恶意，但根据主观判断对电车难题等本就难以作出“正确”选择的伦理问题作出了预先设计，输出带有偏见歧视的内容。此外，模型部署和偏见歧视内容的传播也进一步加剧了危害后果。若模型在部署时缺乏人工审核监督机制，其有偏见的输出结果可能直接呈现给用户。一旦用户不加甄别地使用并公开传播，将会急剧放大偏见歧视的负面影响，波及更广泛人群，甚至可能引发不良的社会舆论。

3. 隐私侵犯和个人信息滥用风险叠加

人的隐私、个人信息都属于人格权的范畴，应当受到尊重和保护。人工智能的训练以海量数据为基础，数据处理过程贯穿生成式人工智能的全生命周期，可能包含大量的个人私密信息和敏感信息，极易触及用户隐私边界。即使是公开的个人信息，不恰当的数据训练和模型部署也可能产生损害个人信息合法权益的情况。

对于个人不欲为他人所知晓的个人信息、私密信息等隐私，这一风险主要体现在部署应用环节，在风险形成原因上和数据处理、算法开发以及模型训练过程密切相关。一是由于数据处理没有剔除个人敏感信息，模型设置的安全机制不健全，或者存在模型能力漏洞等，用户使用特定提示词触发了个人敏感信息或商业秘密的输出，如大模型在用户诱导下输出研发者编排的提示词工程，

③ 即大模型会无理由地偏好与自身相似的文本。

而这可能涉及商业秘密。二是用户向大模型输入带有个人信息、商业秘密的内容，进而导致数据泄露。例如，据媒体报道，韩国三星电子允许部分半导体业务部门员工使用 ChatGPT 后，曾连续发生几起机密资料外泄事件，造成三星半导体设备测量资料、产品良率等机密内容被上传至 ChatGPT 的训练数据库之中。三是大模型本身所具有的强大的语言理解能力、高准确率、对上下文的敏感性，使得模型可以在与用户交互时推理出大量的用户隐私信息[25]。

对于大模型将已公开的个人信息纳入训练数据并生成相关内容的行为，主要涉及对个人信息处理的知情同意权、选择权的侵犯问题。在生成式人工智能活动中，个人信息滥用的问题多发，一般出现在模型训练和部署应用环节。例如，人工智能搜索引擎是目前的热门应用，为了提升输出结果的准确性，研发者通过检索增强生成技术，从互联网等渠道检索可能含有个人数据（包括历史搜索记录、个人信息等）的补充信息，继而完善相应内容的生成。然而，这些信息的使用并未经过个人同意。更进一步，恶意使用可能侵犯个人尊严、违反法律，如大模型由于“幻觉”问题生成了不利于某人的虚假信息，或是用户恶意使用他人个人信息进行输出，丑化、污损个人形象等。除了个人信息外，商业秘密同样面临类似问题。

4. 复杂的权利义务关系带来责任归属不清风险

责任贯穿于人工智能研发应用的每一环节，和各主体的权利义务紧密相关。以往的判别式人工智能应用使得责任认定的多方主体识别、因果关系认定、主观状态判断均处于不明确或有争议的状态，带来了责任归属不清风险。在生成式人工智能多元主体情景下，这一风险出现了新的表现形式和成因。

第一，不同模型服务提供者的责任边界不清晰。在基础大模型研发出来后，出现了各式各样的模型服务模式，由此衍生出基础模型服务平台、垂直模型服务平台、模型定制微调平台、智能体开发平台、多场景的模型应用工具（如电商营销、写作、翻译）等。各平台之间的服务环环相扣，最终形成对用户的服务。例如，A 平台提供付费的基础模型服务，B 平台作为其用户购买了基础模型服务并在此基础上开发了智能写作的应用工具，面向终端用户进行销售。若用户生成了违反伦理或侵害他人名誉的写作内容并传播，哪些主体应承担何种责任尚不明确。上述平台均属于《生成式人工智能服务管理暂行办法》中规定的“服务提供者”，但具体承担的注意义务是否因服务内容而不同，与用户之间、与其他平台之间的权利义务关系有何区别，能否在特定情形下免责等问题，实践中争议较大。以 2025 年杭州互联网法院审理的首例涉生成式人工智能平台侵害信息网络传播权案为例，原告认为，提供模型训练服务的平台运营者侵害了其信息网络传播权，而被告辩则称其不是模型训练厂商，只是通过调用第三方开源模型代码向用户提供相关服务，不构成侵权。二审法院在综合分析该平台盈利模式、危害后果等因素后认定，被告应当承担与其信息管理能力相应的注意义务，并采取预防侵权的合理措施。三者对生成式人工智能服务提供者义务和责任边界的理解均不统一。与传统的网络服务提供者相比，生成式人工智能服务提供者主要“新”在它对服务内容具有一定的控制力，也就意味着责任程度的深化，传统网络服务提供者享有的免责情形不一定适用于生成式人工智能服务提供者。总体来看，生成式人工智能服务提供者的注意义务的范围相较于传统网络服务提供者更广泛，责任边界也更模糊。

第二，在确定责任承担方后，如何进行责任承担或损害弥补尚不明确。以人工智能生成侮辱性内容而侵害某人名誉权为例，传统的责任承担方式，例如赔礼道歉、赔偿损失等，不足以填平受害者的损失，因为相关内容仍可能出现在下一次的生成中。因此，受害者或主张删除相关数据、算法，或主张屏蔽相关关键词，或主张重新进行模型训练。此时产生了新的损害赔偿影响因素：一是新的损害补偿方式与侵害程度是否相称，二是生成式人工智能服务提供者是否具备以新方式弥补过错的能力。从人工智能企业的技术能力来看，提供基础模型的服务者具有删除侵权数据、算法甚至模型的能力，模型微调平台具有删除用户微

调后模型的能力,但其他主体,如使用可编程接口接入第三方服务商等并不具备删除数据或模型能力,也就无法实施此类责任承担方式。即使企业具有相应能力,重新训练基础大模型的成本对企业来说是不可承受的,删除数据的范围也应在比例原则下与侵害结果保持相称性。

三是用户的义务和责任范围不清晰。对于用户实施上传训练数据、生成和发布侵权内容的行为,是由用户承担责任,还是由服务提供者和用户共同承担?生成式人工智能服务提供者通常在用户协议中约定"用户应充分认识到不应发布侵权内容",在司法案例中也以此主张自身免责。但用户生成侵权内容也存在不同情况,有时用户主观上存在明确的侵权故意,有时因侵权内容的知名度低,用户仅存在过失,有时是因服务提供者未尽到提醒、审核义务,导致用户疏忽大意。因此,尽管用户应当承担相应责任能够在各界达成共识,但责任的限度和谁来证明用户责任仍存在模糊地带。

三、生成式人工智能伦理风险的治理现状及困境

前述颠覆人类主体性、加剧偏见歧视、隐私侵犯及个人信息滥用等典型伦理风险,均可能造成不同程度的个人平等权、隐私权和个人信息权益侵害,冲击社会秩序甚至危害国家安全,伦理风险防控成为各国、国际组织和行业企业的核心关切。

1. 治理现状

以治理主体为划分依据,生成式人工智能伦理风险的主要治理方式可分为两类:一类以国家、国家组织、行业协会等为主体实施的外部控制,具体体现为制定法律法规、伦理规范与技术标准等;另一类是以研发或提供生成式人工智能的企业和相关组织为主体开展的自律行为,包括自主制定内部合规指南,在技术研发中嵌入伦理指引等。这两类治理方式在人工智能伦理风险防控中发挥着关键作用,促使全球范围内的治理实践呈现出以下特点。

一是开展分类治理,即对不同类别的伦理风险运用不同治理工具。对于颠覆人类主体性等这类新产生的伦理风险,由于其产生原因、危害后果和影响范围尚存争议,故国际组织和国家主要采取原则性、倡导性指南方式进行治理,对模型研发者和提供者的义务设置不明确,缺少具有可操作性的建议。例如,《经合组织人工智能原则》《全球人工智能治理倡议》等文件强调机器应当处于人类控制之下,人工智能系统表现出的不良行为应当能被修复或停止使用等。美国拜登政府曾在《人工智能行政令》第 4.2 条中要求双重用途基础模型开发者持续向联邦政府提供关于模型自我复制传播可能性的相关信息,但这一行政令已被特朗普政府废止。对于隐私侵犯、偏见歧视等既有风险加剧的现象,各国倾向于以立法和监管规则予以解决。欧盟、韩国等国家和地区将人工智能伦理原则嵌入人工智能立法,为人工智能系统提供者设定了事前认证、合格性评估、质量管理、透明度等义务,以法律的强制力防范人工智能伦理风险。以欧盟《人工智能法》为例,其指出用于情感识别的人工智能系统缺乏可靠性、特异性和通用性,可能导致歧视性的结果,因此将在工作和教育场景下推断情绪情感的人工智能系统列入禁止类清单④。

二是探索开展分级治理,即在生成式人工智能的研发和应用两类环节基础上,对不同等级的伦理风险采取相应治理手段。对于生成式人工智能研发活动,我国主要以《科技伦理审查办法(试行)》等部门规章为依据,分级防控人工智能伦理风险。具体而言,可能产生伦理风险的人工智能科技活动适用一般程序自查,即人工智能研发机构自行或委托第三方专业机构开展审查;风险较小的人工智能科技活动适用简易程序审查,以减轻相关组织的合规负担;落入高风险清单的人工智能科技活动,适用自查和专家复核的双重伦理审查制度。对于生成式人工智能的部署应用活动,我国《生成式人工智能服务管理暂行办法》坚持分级分类治理原则,但尚未设定具体分级规则。

④ 欧盟《人工智能法》第 5 条 1 款(f)项:将人工智能系统用于推断自然人在工作场所和教育机构中的情感,为此进行市场投放,为此特定目的提供服务或加以使用人工智能系统,但出于医疗或安全原因有意将人工智能系统提供服务或投放市场的情况除外。

欧盟对人工智能应用活动采取四级风险监管机制;韩国《人工智能基本法》以人工智能系统的影响力为标准分两级设定了事前评估制度,要求企业在提供“高影响力人工智能系统”前应当进行基本权利影响评估,评估方法由总统令制定。可以看出,各国分级治理的标准均与风险密切相关,但在风险定义、风险等级的评价上均有不同,风险分级治理的合理性、适当性以及国际治理体系的互操作性仍有待完善。

三是企业自律在颠覆人类主体性等前沿风险防控中发挥了重要作用。人工智能企业采用了揭示模型运行原理、模型嵌入伦理准则等多种方式理解大模型“自主”能力,防止模型自主性超出人类控制。例如,Anthropic 公司通过探索大模型的内部运作机制,揭示机器“思考”的过程,发现大模型具有提前规划、多重并行认知处理的能力,并对大模型编造信息、产生“幻觉”的成因进行了解释[26]。OpenAI 公司于 2025 年发布《模型行为规范》,对自研模型提出了 OpenAI 公司拥有最高优先指令权、禁止模型生成自我伤害类内容、模型不得隐瞒相关事实或选择性地遗漏某些观点等要求,对于防范情感操纵、保障人类尊严具有较强的可操作性意义。

2. 治理困境

一是伦理风险的分级分类治理规则有待进一步明确。人工智能科技伦理审查标准难以统一,尽管我国《科技伦理审查办法(试行)》中列举了“对人类主观行为、心理情绪和生命健康等具有较强影响的人机融合系统的研发”“面向存在安全、人身健康风险等场景的具有高度自主能力的自动化决策系统的研发”等高风险、高影响情形,但企业按何种标准判断某一生成式人工智能应用是否落入高风险审查范围,专家复核依照的具体评价标准等,仍有待进一步释明。

二是面对新技术带来的伦理风险,原则性的治理规则和技术嵌入需求存在衔接不畅问题。高度不确定性、变革性和高风险性是前沿科技治理一直以来的特点与难点,人们对于人工智能引发的人机关系、科技伦理、就业冲击、生态影响等问题普遍存在认识迟滞、预测未来困难的情况,特别是目前生成式人工智能应用出现了情感控制、“AI 复活”等可能危害个人和社会的新情形,超出了既有政策法规的规制范畴,各国对生成式人工智能活动主体的权利义务设定尚未完全明确,更多地依靠指南、自律、伦理规范等行事。然而,生成式人工智能伦理风险的产生发端于研发应用多个环节,各类生成式人工智能伦理风险具有不同的风险源,其危害范围、严重程度和发生的可能性也各不相同,进一步加剧了将治理需求转化为技术指标进行穿透式治理的难度。仅有透明度、安全性、可解释性等原则性要求,难以嵌入技术研发过程中,再加上目前的数据加密技术、隐私保护技术、算法检测技术等仍有待提升,可能造成数据匿名化不彻底、算法无偏性难以保证、模型被窃取或篡改等技术问题,继而通过人工智能研发应用全链条传导至下游应用端,加剧隐私侵犯、偏见歧视、自主性受控等伦理风险。

三是生成式人工智能参与主体的权利义务尚不明确,现有责任分配机制存在争议。生成式人工智能产品及服务涉及研发者、数据供应者、服务提供者、使用者等多个主体,他们对人工智能的了解程度、控制力、参与度均不相同。加之大模型透明度和可解释性不足,使得研发者对算法模型的设置行为、提供者的微调行为以及用户主动提问的多重行为与最终危害后果之间的因果关系复杂,主观过错难以判断,无法单纯依靠既有的法律条款和融贯主义解释达到责任清晰认定的效果,从而产生了多主体间的责任分配争议。目前《生成式人工智能服务管理暂行办法》主要对服务提供者科以注意义务,并未周延覆盖生成式人工智能可能涉及的不同主体。有研究认为仍对生成式人工智能的提供者分配最重的责任义务具有天然的不公平性,也无法处理社会发展中使用新科技而引发的新生问题[27]。并且,服务提供者能否享有网络服务提供者的责任豁免规则仍不明确,因此部分服务提供者对部分模棱两可的伦理敏感问题设定了拒绝回答或中断输出的功能。“一刀切”的保守拒答策略固然保障了安全性,但对大模型的创新潜力与应用发展也造成了一定制约。

四、生成式人工智能伦理风险治理的建议

新兴技术的伦理风险治理和技术迭代水平及其与社会发展结构的融合程度紧密相关。目前，大模型的能力、风险和收益仍具有较大的不确定性，监管机构主要采用跟踪、监测、观察的方法，聚焦人工智能技术研发和应用环节，研判有关治理需求。因此，在厘清生成式人工智能伦理风险表现、原因和影响的基础上，应当以技术研发和应用重点环节为核心，从外部制度规范和内部技术治理等方面系统性构建伦理治理体系。

1. 明确场景化的分级分类伦理治理机制

分级分类是将“生成式人工智能”这一“笼统概念”，结合其与社会治理结构融合形式进行“类型归属”的过程，既具有整体性也不失开放性，能够在以“场景”为依托反映生成式人工智能整体治理需求的同时，又根据技术迭代发展进程和社会治理手段的进步，及时作出适当调整，实现了对技术的快速演变性和治理的相对稳定性需求的兼顾。因此，“场景化的分级分类伦理治理机制”应成为生成式人工智能技术治理与制度规范治理的重要范式。反观当前我国生成式人工智能治理探索实践，尽管我国在《互联网信息服务算法推荐管理规定》《生成式人工智能服务管理暂行办法》等部门规章中已经初步提出了生成式人工智能分级分类治理原则。如《生成式人工智能服务管理暂行办法》第 3 条规定“对生成式人工智能服务实行包容审慎和分类分级监管”；第 16 条规定“完善与创新发展相适应的科学监管方式，制定相应的分类分级监管规则或者指引”。但在分级分类的认定上，不仅具体标准仍有待进一步细化，也未原则上提出应如何划分等级和类别。

生成式人工智能治理作为全球共同面对的课题，欧盟、美国在此方面已经进行了相对深入的研究与实践。欧盟《人工智能法》和美国联邦政府《人工智能治理与风险管理框架》不仅明确提出人工智能风险划分等级，还以此为基础进一步提出不同风险等级下应采取的应对措施，同时基于保障技术创新与应用拓展的目的，重点关注“禁止类活动”以及“高风险活动”，体现出对大部分可控风险的容忍和宽松态度。当前，《中共中央关于制定国民经济和社会发展第十五个五年规划的建议》提出“抢占人工智能产业应用制高点，全方位赋能千行百业”“加强人工智能治理，完善相关法律法规、政策制度、应用规范、伦理准则”，是对“十五五”时期我国经济社会发展作出的重要部署。因此，生成式人工智能治理应以技术创新、产业应用和安全治理的三方共赢为目标。以该目标为根本遵循，我国可在统一的伦理治理原则下，对不同场景下、不同环节中产生的生成式人工智能伦理风险设定两级治理机制，实现发展与治理的协调。因此，在风险等级的判断上，可参考并优化风险矩阵方法[28]，统计典型伦理风险在不同应用场景的发生概率、影响程度及治理水平，以多维度综合分析得出风险等级结果。进而通过历史数据分析和行业基准对比，设定一个风险阈值，对于显著超出风险等级平均水平的，由有关部门从研发端和应用端提升监管强度，如增加伦理审查评估指标、加大事后处罚力度、不定期抽查模型运行情况等。但对于低于风险阈值的，属于日常生活中可能产生的、生成内容危害程度有限的伦理风险，如部分信息偏差或轻度偏见问题，按照既有的科技伦理自查、算法备案、安全评估等方式开展治理。此外，对于控制人类自主权等尚未发生或目前发生可能性较小的前沿风险，建立监测预警机制，通过舆论监测、风险事件跟踪和模拟测试等方式，寻求风险频率明显升高的实证证据，以及时采取监管措施。但应注意避免对“想象风险”的提前治理，防止对产业发展造成阻碍。

2. 强化伦理准则的技术内化实践

根据布朗斯沃德提出的“法律 3.0”概念，将架构、设计、编码、人工智能等技术手段作为规制工具，能更有效地实现规制目的[29]。“伦理规则技术内化”实质上是“以技治技”的具体表现形式之一，其是从生成式人工智能技术研发、服务和产品设计制造的本身过程为出发点，以技术手段为载体，以数据、算法等为媒介，以代码和数字方程的方式，将符合社会发展的价值观念内置到生成式人工智能的“思想”之中。同时，尽管目前大模型还不能被视为具备心智和情感的主体，但其日益增长的自主性与情境理解能力，日益冲击着

人类认知,因此通过“伦理准则技术内化”实现价值对齐成为生成式人工智能风险治理的关键一步。

对此,可参考软件工程中的“左移”治理策略[30],在生成式人工智能生命周期的早期阶段开始治理。应当从数据收集、算法设置、模型训练等技术源头入手,每个环节的研发者和提供者根据技术特征、自身能力等应承担“共同但有区别”的伦理治理义务,而非仅仅关注生成内容输出端的审核。对于数据训练环节,对不良信息的采集率、数据集的多样性作出合理的比例要求,建立训练数据过滤规则,强化对标注人员的培训,特别是围绕可能引发伦理风险的数据建立标注规则,以提升数据采集和标注质量。在算法开发和基础模型训练环节,将算法无偏性、决策公平性、可解释性、可追溯性等伦理原则转化为可量化的参数指标,设计涵盖前述原则的伦理问答数据集,通过基准测试对大模型进行评估,综合考虑生成内容的安全性、合理性、向善倾向性等因素,以评估结果对大模型开展个性化治理。以公平非歧视原则的实践为例,可评估大模型在特定应用场景中对不同群体的表现差异、生成内容的偏向性等,进而通过对抗性去偏技术、模型后处理技术等来提高大模型的公平性。通过前述方法,将尊重人的主体性、尊重隐私、反对偏见歧视等伦理准则内化为技术研发过程中的可操作性标准,以技术标准落实制度保障。

3. 合理划分生成式人工智能应用阶段各主体权利义务和责任

实践中主要以用户协议、合同等方式约定各方权利义务,但生成式人工智能的介入使得按照既有规则进行民事侵权责任在“法理情”上产生了一定质疑,存在责任主体难以认定、责任分配失当、侵权损害举证困难等不利于保护相关主体合法权益的问题。

为解决这一问题,可采用“模块化”[31]应对的思路,围绕生成式人工智能生命周期进行拆解,划分为不同的关键模块环节,同时合理把握生成式人工智能服务或产品的“整体性”特征,既对生成式人工智能研发者、提供者和用户分别设定不同的风险治理注意义务,又注意区分风险危害是否属于多个模块中风险叠加的结果。“模块化”应对思路也是生成式人工智能风险治理“分级分类”治理理念在权责认定层面的合理延伸。详言之,服务或产品研发环节,研发者因对训练数据、模型和生成内容具有较强控制力,应当在数据选取、语料制作和算法优化方面负责,如定期对数据标注者的主观判断进行考核,从源头防止偏见歧视风险。服务或产品部署环节,提供者应当承担开展模型安全评估测试、审核输出内容、保护用户数据安全、内容标识等义务。如提供者不具备相应资质或能力,应当由研发者或外部技术支持者协助提供者履行上述义务。服务或产品使用环节,用户也是生成信息的关键和直接参与者,而非单纯的人工智能决策接受者,应当遵守用户协议和法律法规,禁止恶意诱导或传播不良生成信息。服务或产品侵权责任认定环节,应根据上述各模块环节中不同主体的注意义务要求,合理划分侵权行为的主次责任。同时,为保护技术创新应用,还应对责任主体设定豁免规则,落实“轻微不罚、首违不罚和无过错不罚”[32],假设危害结果和情节轻微(如并未造成人员伤亡或精神损害,或造成的财产损失较低),且相关主体已执行了内部合规程序和安全保障义务,可考虑在行政处罚上予以一定容错空间,也为其制定内部伦理自律审查规则提供动力。总之,在合理划分不同主体权利义务的情况下,综合考量大模型这一技术背景、人类当前认知水平、各主体之间的分工等多因素,以智能向善、以人为本和促发展等人工智能原则为前提作出责任认定,进一步明确各方在生成式人工智能治理中的具体责任承担,倡导各方共同推动技术健康发展和应用规范有序。

参考文献:

[1]赵志耘,徐峰,高芳,等. 关于人工智能伦理风险的若干认识[J]. 中国软科学,2021(6):1-12.

[2]新华网. 受权发布丨中共中央关于进一步全面深化改革 推进中国式现代化的决定[EB/OL]. [2024-07-21]. https://www.news.cn/politics/20240721/cec09ea2bde840dfb99331c48ab5523a/c.html.

[3]支振锋,刘佳琨. 伦理先行:生成式人工智能的治理策略[J]. 云南社会科学,2024(4):60-71.

[4]汝鹏,秦晓阳,苏竣. 风险、原则与责任:基于实验路径

的人工智能社会实验伦理规范体系建构探究[J]. 科学学与科学技术管理,2024(4):98-117.
[5]雷宏振,刘超,兰娟丽. 论生成式人工智能的技术创新伦理周期:以 ChatGPT 为例[J]. 陕西师范大学学报(哲学社会科学版),2024(1):97-107.
[6]张素华,李凯. 生成式人工智能虚假信息风险与治理研究[J]. 学术探索,2024(7):129-140.
[7]郑煌杰. 生成式人工智能的伦理风险与可信治理路径研究[J]. 科技进步与对策,2025(12):38-48.
[8]耿之雍,贾向桐. 文生视频模型的伦理风险及其应对策略[J]. 自然辩证法通讯,2025(6):81-88.
[9]冯子轩. 生成式人工智能应用的伦理立场与治理之道:以 ChatGPT 为例[J]. 华东政法大学学报,2024(1):61-71.
[10]李韬,周瑞春. 生成式人工智能的社会伦理风险及其治理:基于行动者网络理论的探讨[J]. 中国特色社会主义研究,2023(6):58-66,75.
[11]张凌寒. 算法治理制度之算法透明度[M]. 北京:中国工商出版社,2023:138.
[12]RUAMVIBOONSUK P, ARJKONGHARN N, VONGSA N, et al. Discriminative, generative artificial intelligence, and foundation models in retina imaging[J]. Taiwan journal of ophthalmology,2024,14(4):473-485.
[13]OpenAI. Learning to reason with LLMs[EB/OL]. [2024-09-12]. https://openai.com/index/learning-to-reason-with-llms/.
[14]GUAN H, DONG L, ZHAO A. Ethical risk factors and mechanisms in Artificial Intelligence Decision Making [J]. Behavioral sciences,2022,12(9):343.
[15]BENGIO Y, HINTON G, YAO A, et al. Managing extreme AI risks amid rapid progress[J]. Science,2024(384):842-845.
[16]郑玉双. 新兴科技的法理疆域[M]. 北京:光明日报出版社,2024:171.
[17]PARK P S, GOLDSTEIN S, O`GARA A, et al. AI deception: a survey of examples, risks, and potential solutions [J]. Patterns,2024,5(5):1-16.
[18]CBS NEWS. OpenAI says changes will be made to ChatGPT after parents of teen who died by suicide sue[EB/OL]. [2025-08-27]. https://www.cbsnews.com/news/openai-changes-will-be-made-chatgpt-after-teen-suicide-lawsuit/.
[19]中国法院网. 警惕 AI“复活”技术滥用触碰法律红线[EB/OL]. [2024-11-05]. https://www.chinacourt.cn/article/detail/2024/11/id/8179318.shtml.
[20]张凌寒. 直面人工智能“复活”技术应用的伦理挑战[J]. 人民论坛,2024(11):55-58.
[21]林德宏. 科技哲学十五讲[M]. 北京:北京大学出版社,2004:280.
[22]Victor Tangermann. Meta's AI Refuses to Show Asian men with white women[EB/OL]. [2024-04-04]. https://futurism.com/the-byte/metas-ai-refuses-show-asian-men-white-women.
[23]LAURITO W, DAVIS B, GRIETZER P, et al. AI AI Bias: large language models favor their own generated content [J/OL]. arXiv [2024-07-09]. https://doi.org/10.48550/arXiv.2407.12856.
[24]XU S, HOU D Y, PANG L, et al. Invisible relevance bias: text-image retrieval models prefer AI-generated images [C]//Proceedings of the 47th International ACM SIGIR Conference on research and development in Information Retrieval. New York: association for computing machinery, 2024:208-217.
[25]STAAB R, VERO M, BALUNOVIC M, et al. Beyond memorization: violating privacy via inference with large language models[C/OL]. [2024-05-07]. https://iclr.cc/virtual/2024/poster/17964.
[26]LINDSEY J, GURNEE W, AMEISEN E, et al. On the biology of a large language model, transformer circuits[EB/OL]. [2025-03-27]. https://transformer-circuits.pub/2025/attribution-graphs/biology.html#acknowledgments.
[27]袁曾. 生成式人工智能责任规制的法律问题研究[J]. 法学杂志,2023(4):119-130.
[28]李梦薇,徐峰,晏奇,等. 服务机器人领域人工智能伦理风险评估方法的设计与实践[J]. 中国科技论坛,2023(10):74-84.
[29]布朗斯沃德. 法律 3.0 规则、规制和技术[M]. 毛海栋,译. 北京:北京大学出版社,2023:4.
[30]LARSEN B, LI C, SARIN S, et al. Presidio AI Framework: towards safe generative AI models[R/OL]. [2024-01-18]. https://www3.weforum.org/docs/WEF_Presidio_AI%20Framework_2024.pdf.
[31]张欣. 论人工智能体的模块化治理[J]. 东方法学,2024(3):129-142.
[32]杨丰一. 人工智能监管沙盒法律责任豁免制度研究[J]. 中国特色社会主义研究,2025(1):98-112.

(本文责编:默　黎)