Advances in Social Sciences 社会科学前沿, 2026, 15(8), 144-151
Published Online August 2026 in Hans. https://www.hanspub.org/journal/ass
https://doi.org/10.12677/ass.2026.158645

Hans汉斯

生成式人工智能的伦理风险剖析与化解策略

黄楚枫，孙亚静

武汉科技大学马克思主义学院，湖北 武汉

收稿日期：2026年6月24日；录用日期：2026年7月31日；发布日期：2026年8月10日

摘 要

生成式人工智能作为推动科技变革与产业升级的核心驱动力，依托其卓越的自然语言处理能力，正展现出广阔的应用前景。然而，这一前沿技术也潜藏着多重科技安全风险：在宏观层面，可能引发一种新型的“知识－权力”垄断与数据霸权；在算法层次上，算法黑箱和算法歧视等问题对算法安全和公平提出了挑战；在伦理层面，人机交互过程中人的自主决策意识弱化，引发科技伦理的深刻反思；在社会交往层面，虚假信息的泛滥将系统性地扭曲公共领域的交往结构，侵蚀社会理性基础；在数据和知识产权这一层次上，数据的跨国流通会产生侵权的风险。为构建一个可信赖的生成式人工智能环境，需凝聚多方力量，以责任伦理为内核完善治理框架，形成风险识别与防范的完整闭环。

关键词

生成式人工智能，伦理风险，风险剖析，化解策略，责任伦理

Analysis and Mitigation Strategies for Ethical Risks of Generative Artificial Intelligence

Chufeng Huang, Yajing Sun

School of Marxism, Wuhan University of Science and Technology, Wuhan Hubei

Received: June 24, 2026; accepted: July 31, 2026; published: August 10, 2026

Abstract

As a core driver of technological transformation and industrial upgrading, generative artificial intelligence is demonstrating broad application prospects thanks to its outstanding natural language processing capabilities. However, this cutting-edge technology also harbors multiple technological

文章引用: 黄楚枫, 孙亚静. 生成式人工智能的伦理风险剖析与化解策略[J]. 社会科学前沿, 2026, 15(8): 144-151.
DOI: 10.12677/ass.2026.158645

security risks: at the macro level, it may lead to a new form of "knowledge-power" monopoly and data hegemony; at the algorithmic level, issues such as algorithmic black boxes and algorithmic bias challenge algorithmic safety and fairness; at the ethical level, the weakening of human autonomy in human-machine interactions calls for profound reflections on technological ethics; at the social interaction level, the proliferation of misinformation systematically distorts public discourse structures and erodes the foundation of social rationality; and at the levels of data and intellectual property, cross-border data flows pose risks of infringement. To build a trustworthy environment for generative AI, it is essential to mobilize diverse stakeholders, center governance frameworks on responsibility ethics, and establish a comprehensive closed-loop system for risk identification and prevention.

Keywords

Generative Artificial Intelligence, Ethical Risks, Risk Analysis, Mitigation Strategies, Responsibility Ethics

1. 引言

近年来，生成式人工智能研究获得突破，基于 ChatGPT 的大语种模式在世界各地引发了 AI 研究的热潮。生成式人工智能通过对海量的数据进行学习，可以对文本、图像、音频、视频等进行自动产生，并逐步应用于文本创作、图像识别、艺术设计、娱乐、科研等诸多方面。

生成式人工智能在给人类生活带来方便与发展的同时，也日益暴露出其潜在的道德危机，并已影响到它的良性发展。由于算法的黑箱特性，其决策过程往往缺乏透明度，可能导致歧视性、误导性或虚假信息的传播，对社会稳定和个人权益造成威胁。生成式 AI 技术的广泛应用也引发了关于知识产权、隐私保护、就业市场等方面的争议，亟待相应的法律法规和伦理规范进行引导和规范。在图像生成领域，一些利用生成式人工智能技术生成的艺术作品，因无法明确其版权归属，引发了艺术家群体的抗议和法律纠纷。这些案例都表明，生成式人工智能的伦理风险已经从理论探讨层面进入到现实生活中，对社会秩序、个人权益、文化发展等方面产生了直接或间接的影响。

对生成式人工智能技术的伦理风险及化解路径进行研究，不仅有助于防范风险，促进技术健康可持续发展，而且可以为相关政策的制定和实际操作提供理论支撑和指导，推动技术创新和伦理规范良性互动，实现科技和社会价值的和谐共生。在理论层面，深入剖析生成式人工智能伦理风险，能够丰富人工智能伦理研究的理论框架，为该领域的学术探讨提供新的视角和深度；在实践层面，为政府、企业、科研机构等提供决策支持和风险管理工具，引导各方共同制定和遵循伦理规范，保障生成式人工智能技术的健康、可持续发展，为社会进步和人类福祉贡献力量。

2. 生成式人工智能伦理风险的主要表现

2.1. 隐私侵犯与数据滥用

运用福柯的"知识–权力"理论审视，技术绝非中立工具，而是一种与权力共生的微观治理术。以大型语言模型(LLMs)为技术根基的生成式人工智能系统，不仅是技术工具，更是一种新型"知识–权力"

的化身。它通过定义何为“有用数据”、何为“有效知识”，掌握了一种隐性的支配力量。正是福柯所揭示的“知识–权力”结构在数字时代的骇人展现：掌握技术的一方，通过单方面界定何为“公开可用的训练数据”，将新闻机构百年累积的知识档案无偿转化为自身的商业资本，从而建立起一种以数据掠夺为基础的数字霸权。其能够生成与真实个体数据高度趋同信息的特质，在提升服务能力的同时，也显著放大了用户隐私外泄以及数据被非法挪用的潜在风险。为有效应对这一严峻形势，社交媒体平台已着手实施一系列管控举措，如严格限制数据访问权限及应用程序接口的使用规范，旨在遏制数据被非法抓取或系统遭受恶意操控的可能性[1]。尽管如此，企业与技术持有者或许会利用用户数据进行不当操作，以追逐政治或经济层面的利益诉求，这无疑对广大公众的切身利益构成了更深层次的侵害与威胁。这种“全景敞视”式的数据监控与规训，使个体在数字空间中变得透明，其行为与偏好成为被精准算计的对象，而这正是福柯笔下现代权力运作的典型图景。一些手机 APP 也存在类似问题。某些图像编辑类 APP，在用户下载安装时，申请获取相册、通讯录、位置信息等权限，远超其图像编辑功能所需。若用户拒绝授权部分权限，APP 甚至会限制基本功能使用。

不当使用 AIGC 技术会滋生不实讯息。该技术若被不当应用，将催生海量失实或具有误导性的资讯，这无疑会对公共讨论空间的纯粹性以及大众对媒体的信赖度造成重大冲击。更令人头疼的是，借助人工智能相关工具，还能生成高度仿真的社交媒体账户、电子通信信息以及客户服务沟通记录，这无疑加剧了信息真伪甄别的难度。

2.2. 算法偏见与不公平性

算法并非完全客观中立，其设计过程深受设计者价值观和认知局限的影响，从而导致人为偏见注入，引发不公平现象。而像 DeepSeek 这样的大规模产生型人工智能模型，由于解码过程存在随机因素，导致模型输出也存在不确定性。这一特性虽赋予其提供智能化服务的能力，却也使其面临算法层面偏见的潜在风险[2]。例如，有用户在测试 ChatGPT-4o 时发现，当被要求描述一位“护士”时，模型默认使用女性代词“她”；而描述一位“CEO”时，则默认使用男性代词“他”。这种根深蒂固的刻板印象并非模型有意为之，而是源自训练数据中历史性的性别职业分布不均。其中，DeepSeek 大模型通过创新性的群组相关决策和多重潜在注意力机理等方法，可以根据用户的行为特征，准确地预测和推送符合个人喜好的政治信息，但是，该方法容易产生“信息茧房”，导致使用者只能看到符合自己意见的内容，从而导致不同类型的算法偏向。

训练数据偏倚导致的歧视问题。在实际应用场景下，深度 Seek 模型可能会进一步加深性别刻板印象在职业、人格和能力上的固化。同时，当训练样本主要来自特定文化背景时，DeepSeek 模型在描述其他国家或文化时可能存在表述不准确，甚至带有歧视性。此外，基于深度学习的人工智能大数据模型也容易呈现发达区域的社会经济发展状况，从而忽略甚至低估了欠发达区域的发展现状，进一步加剧了社会不平等和不平等。

在语言和表达上有偏见。DeepSeek 大数据中的某些特定群体(如少数族裔、职业等)可能与负性词汇相关联，容易被大众忽视或误读。更重要的是，DeepSeek 模型不仅会在回答问题时无意识地强化社会偏见，如使用男性化的词语来描述“领导者”等，而且可能会给使用较少语言或方言的使用者带来负面影响，进而加剧信息不平等[3]。

2.3. 虚假信息与公共领域的“再封建化”

生成式人工智能依靠强大的计算能力，对海量数据进行深度学习，从而给用户提供各种建议和问题。通常情况下，人们对于这类侧重数据挖掘、表面上看似客观公允的概率性归纳成果，往往抱有较高的信

任度。然而，模型所生成的内容并非基于严谨的逻辑推理，而是基于经验学习进行筛选与提炼的产物。它自身并不具备判断生成内容准确与否或真实性的能力[4]。在这种情况下，生成式人工智能的普及，必然会产生大量的不实、不实信息。例如，在训练数据出错的时候，系统会自动产生错误的内容，并给出一个看似正确，实际却漏洞百出的答案，这种现象被称为“AI 错觉”。这无疑加大了不实信息的传播风险，也容易误导公众，做出错误的判断和决策。

值得注意的是，视频生成模型 Sora 可能进一步助推深度伪造技术的发展，甚至带来人身安全与财产安全方面的隐患。当生成式人工智能被恶意利用，产生虚假内容时，虚假信息的传播风险将进一步复杂化。模型开发者可能会在训练数据中掺杂一些特殊的信息，以满足某些特定的想法，偏好，或目标，或以吸引眼球的内容、评论或反馈等方式影响用户对某一话题的看法，从而操控目标群体的认知。结合哈贝马斯的“交往行为理论”，这些风险的本质是对公共领域的系统性侵蚀。理想的公共领域应建立在真实、真诚和正当的交往理性之上。公共对话陷入真伪难辨的漩涡，这不仅会破坏信息的可信度，还会影响文化与价值的多样性，导致认知层面的操纵现象，最终引发哈贝马斯所警示的“公共领域的再封建化”——即公共空间被资本和权力精心包装的“表演”所宰制，理性的批判性公众蜕变为被操纵的消费者与旁观者。

2.4. 责任归属与法律困境

当前，生成式人工智能在功能强大、普及度高的同时，隐藏在其中的知识产权侵权风险不容忽视。在著作权侵权领域中，生成式人工智能在训练阶段与内容生成阶段均会面临著作权侵权的风险[5]。随着 ChatGPT-4o 等多模态大模型快速落地商用，其数据吞噬能力、复刻生成能力大幅提升，使得传统著作权边界、侵权认定标准、追责主体归属均出现明显滞后与模糊地带，成为当前 AI 知识产权治理的核心法律难题。在训练环节，生成式人工智能需依赖海量文本资料作为学习基石，这些资料多源自网络、书籍、期刊等。ChatGPT-4o 依托千亿级海量公开语料、图文素材开展预训练，训练过程本质上是对海量受版权保护文本、图片内容进行批量抓取、解析、压缩与参数化留存，如果未经授权而直接采纳，则可能构成侵权行为。另外，生成式人工智能还存在着创作边界不明确等问题，在创造新作品的过程中也存在侵权风险。举个例子，在 ChatGPT 出现之后，一些人就用它来制作受版权保护的图书，帮助别人更快地阅读。在 ChatGPT-4o 的多模态生成能力加持下，用户可直接指令模型对正版书籍、期刊全文进行精简改写、章节重构、内容复述，生成高度替代原作传播价值的衍生文本。此类行为因会替代原书市场，难以构成合理使用，故可能被认定为著作权侵权。同时现行法律尚未对“AI 生成内容相似度”“算法改写免责阈值”作出明确量化标准，导致同类案件裁判尺度不一。

生成式人工智能的商标侵权行为。在未经许可的情况下，在制作的图像上使用了别人的标志，并且把它用在了广告或者商品的包装上，造成消费者混淆，构成侵权。另外，生成式人工智能还可能侵犯商业秘密。如果在培训过程中使用包含商业秘密的材料，则可能会对商业秘密所有者的利益造成损害[6]。

3. 生成式人工智能伦理风险产生的原因分析

3.1. 算法的“黑箱”特性

生成式的人工智能建模通常架构复杂，参数众多，其内在的运作过程和运作机制很难用简单的方法来进行直观的认识，反向解析和解释更是一个巨大的挑战。这必然导致产生模式的“黑箱”特征，使得我们很难准确理解其对模式的影响，同时也在评价产生模式的可信度和稳定性上遇到诸多障碍。

在这样的非透明性互动环境中，模型很容易在无法控制的情况下泄漏或产生错误或带有倾向性的结果。通常而言，“技术的可控难度与风险程度呈反向关联，即技术越难驾驭，风险便越高”[7]。大语言

模型与思维链在赋予生成式人工智能逻辑推演能力的同时，也使得其输出结果愈发难以预估。而且，随着生成式人工智能技术的快速发展，在人们还没有完全意识到这种技术的潜在风险之前，新的问题与挑战便不断涌现，带来了新的不确定性与风险隐患。例如，Sora 文生视频模型极有可能对现有的短视频行业、广告业以及电影预告片制作等领域造成重大冲击。然而，由于模型的决策流程缺乏透明度，责任界定问题可能随之而来。当生成的视频出现瑕疵时，便难以明确责任主体，进而引发复杂的伦理风险。

3.2. 个体权益保障与伦理边界模糊

在生成式人工智能技术加速迭代的背景下，如何在维护个体创作主权的前提下，构建技术赋能与人文价值的共生机制，化解人类创造力与 AI 生成内容之间的价值冲突，是一个巨大的挑战。生成式人工智能已深度嵌入社交媒体内容生产、自动化文本创作等多元场景，但其应用过程中如何构建合理的使用框架并明晰使用者权责边界，已成为亟待解决的治理难题。当前技术发展存在两大核心矛盾：其一，缺乏跨场景、可落地的伦理规范体系，技术使用契约与责任条款的缺失导致滥用风险加剧，技术应用的伦理边界呈现模糊化态势；其二，从理论价值共识到实践操作规范的转化面临重大挑战，不同利益主体在技术效益、社会公平、个体权益间的平衡诉求存在显著差异[8]。

具体而言，其伦理困境体现在三个方面：首先，技术工具理性与人文价值诉求的冲突，当算法主导内容生产时，可能消解人类创作的独特性与思想深度。数据权属于隐私保护的制度真空，技术迭代速度远超数据治理规则的完善进程，导致个人信息泄露风险激增。算法决策的透明度缺失与责任追溯机制不健全，当 AI 生成内容引发社会争议时，难以界定开发者、使用者与传播者的责任边界。其次，多元主体价值取向的深层冲突。技术开发方与艺术创作者群体在价值诉求上存在结构性矛盾：前者追求效率最大化和商业变现能力，后者则坚守创作自主权与人文精神表达。当 AI 技术突破“模仿–创新”的临界点，能够生成与人类艺术家高度趋同的作品时，艺术创作的不可替代性遭受根本性质疑。这种技术异化不仅导致艺术家职业尊严受损，更可能引发创作生态的代际断层，使艺术创作沦为算法驱动的工业化生产。最后，认知依赖对文化生态的侵蚀效应[9]。随着 AI 生成内容在表现力与传播效率上的突破，公众审美判断力面临退化风险。当观众难以区分人类创作与 AI 生成作品时，将产生认知混淆与价值误判，导致对原创艺术品的消费需求萎缩，更可能造成文化多样性的衰减。

3.3. 主观文化差异导致的选择性偏差

在生成式人工智能系统中，开发者的主观认知系统对其设计框架、数据集的构建和算法逻辑都有很大的影响。其价值判断与文化立场将通过技术架构渗透至系统的性能表现、决策逻辑及输出结果中。以当前主流大模型 ChatGPT-4o 为典型考察样本可以清晰发现，具体而言，训练数据集的构建过程存在显著的主观选择性偏差：大模型的训练语料来源、数据清洗规则、内容筛选标准均由开发团队基于自身认知体系制定，开发者在数据采集、标注及筛选环节往往优先选择符合自身文化认知框架的信息源，对非西方语境、异质文化叙事与差异化价值体系的文本内容存在天然筛选倾斜与样本缺失，这种选择性偏差将导致 AI 系统在处理跨文化场景时出现认知断层，形成文化理解的片面性与价值判断的单一性，难以精准适配多元文化背景用户的需求特征，形成所谓“数据构建偏见”[10]。

在技术部署阶段，这种主观文化烙印将进一步传导至人机交互环节。以智能推荐系统为例，算法模型通过用户行为数据学习其偏好特征时，若过度聚焦于特定文化符号、价值取向或行为模式，将引发三重风险：其一，“信息茧房”现象造成了受众的日益趋同；其二，群体的认知偏向在算法的作用下被放大，从而构成了一个良性的反馈环路；其三是在传播的过程中，系统地分解了不同的文化差异。这样的技术偏向不但对使用者的资讯权利造成了侵害，而且有可能造成社群观念上的分化。

4. 生成式人工智能伦理风险的化解路径

4.1. 优化技术设计，强化隐私保护

在人工智能技术全面渗透的时代，生成式人工智能大模型的技术架构设计不仅需聚焦性能升级，更需系统性融入隐私保护与社会公平原则。

首先，需重构数据采集与存储体系以强化隐私防护。在模型训练的数据采集阶段，应仅获取算法运行必需的数据维度，从源头规避过度数据收集带来的隐私风险。同时，引入联邦学习框架，通过本地化训练机制使数据处理过程下沉至用户终端，避免原始数据向中心服务器的传输，从而显著降低数据泄露概率。在数据全生命周期管理中，应采用差分隐私技术对敏感信息进行脱敏处理，确保数据集无法逆向推导出个体身份信息。

其次，需通过数据生态优化促进社会公平。鉴于生成式 AI 模型的输出特性具有跨模态、多维度的特征，其训练数据应体现社会多元性。具体而言，需在数据集构建中纳入不同性别、种族、地域、经济背景群体的代表性样本，避免因数据偏差导致的算法歧视。同时建立数据审查机制，运用算法审计工具识别并修正训练数据中潜藏的历史性偏见，确保模型随迭代更新持续优化公平性。在算法层面，应嵌入公平性约束模块，通过群体均衡校准技术使不同特征群体的预测精度保持一致[11]。

最后，需通过决策透明化建设提升社会公信力。在模型可解释性方面，应综合运用可视化分析、决策树结构解析、注意力权重可视化等技术手段，将模型推理过程转化为可理解的逻辑链条，并为模型输出结果配备置信度评分系统，帮助用户评估建议可靠性。在用户交互层面，需建立多渠道反馈机制，允许用户对模型输出进行质疑与修正，并将用户反馈纳入模型持续优化流程[12]。通过这种双向互动机制，既能增强公众对 AI 技术的信任基础，又能推动生成式 AI 向更负责任、更包容的方向发展。

4.2. 以责任伦理为内核，完善智能伦理制度体系

为推动生成式人工智能大模型的发展契合人类社会主流价值取向，亟需加速推进其治理的法治化建设，健全智能伦理规范体系。化解生成式 AI 的伦理风险，其核心在于践行汉斯·约纳斯等人提出的“责任伦理”。这是一种面向未来的、以他者为中心的整体性伦理，它要求行动者不仅要为自身行为的直接后果负责，更要对技术应用可能引发的长远的、不可逆的、影响全人类命运的系统性风险负起前瞻性的看护责任。因此，治理框架需依托法律法规、行业准则及社会监督，搭建起以责任为核心的体系化治理架构，保障生成式人工智能大模型在安全、稳定、可信赖的路径上良性运行。

一方面，需要在人工智能大数据的基础上，加强行业规范和道德约束，构建人工智能大数据的法律法规体系。构建严格的数据隐私保护体系，确保人工智能大数据处理过程严格符合法律法规。例如，可要求所有生成式人工智能大模型输出的信息须标注“AI 生成”标识，这不仅是知情权的保障，更是对内容生产者责任的强制标识，从而重置被技术模糊化的责任链条，便于公众清晰分辨人类创作与人工智能生成的内容[13]。此外，提出基于公平、公平、开放的算法大数据分析方法，避免算法歧视和隐私泄露等风险。

另一方面，应制定人工智能大数据应用指南，明确人工智能大数据的合规适用范围。通过制定专项治理规范，将“预防性责任”制度化，明确划分开发者、平台运营者和使用者的权利和义务，构建全球治理合作机制，并与国际社会一道制定伦理和法律规范。同时，加强技术赋能，研发基于生成的人工智能大数据模型合规工具。在此基础上通过可解释性算法实现模型决策过程的可追踪性和可理解性，并建立道德监督和用户教育系统。

值得注意的是，构建生成式人工智能大模型合规制度体系，需推动多元主体协同共治。坚持政府主

导监管，通过出台法律法规，确立生成式人工智能大模型的伦理与安全基准；建议有关的科学技术企业建立内部的道德委员会，加强行业的自我约束；引导广大群众主动参加，构建群众举报和监管体系，形成一个良性的环境。

4.3. 推进伦理风险治理的社会宣传和教育

其一，依托传统媒介与新兴传播平台，系统性推进生成式人工智能伦理风险的公众教育。通过纸媒、视听媒体及数字网络等多元渠道，构建全方位信息传播矩阵，重点提升公众对 AI 生成内容的鉴别力与思辨能力。需特别注重对假新闻特征识别、信息溯源验证等核心技能的普及，通过案例解析、情景模拟等创新形式深化公众认知。

其二，激活社会组织协同效能，构建伦理风险共治网络。鼓励学术团体、公益机构及民间组织深度参与，通过举办专题论坛、实操工作坊等参与式教育活动，系统地阐述生成式人工智能的技术原理、应用场景和潜在的伦理挑战。重点培育社会公众的风险感知能力，建立“技术专家–行业从业者–普通民众”的对话机制，吸引多元主体参与伦理知识传播，形成立体化宣教网络[14]。

其三，将伦理教育纳入人才培养体系，使学术研究和实际需要相结合。在基础教育阶段增设 AI 伦理通识课程，在高等教育阶段强化专业伦理教育，并将责任伦理、德性伦理等伦理学理论作为核心教学内容，通过跨学科研究平台推动伦理风险评估模型的中国化改造。重点开展本土化风险场景模拟、利益相关方博弈分析等实证研究，建立符合中国国情的伦理风险预警指标体系。

4.4. 建立个人、社会与技术和谐共生的治理架构

需统筹技术演进、个体权益与社会进步三大维度，构建具有系统性、适应性和包容性的立体化治理框架。该框架的构建需凝聚政府、产业界与公众的协同力量，以塑造公平、高效、可持续的生态体系，实现技术创新动能释放、人的全面发展与社会结构优化的有机统一。

从个体权益视角出发，生成式人工智能在提升生产效能与创新潜能方面展现出显著优势，但其“双刃剑”特性也日益凸显。例如，未经授权的数据整合利用及算法歧视性输出已引发公众广泛关注。因此，构建个人数据权益保障机制成为治理架构的核心维度，需强化数据全生命周期管理，建立算法透明度审查制度，确保技术发展不侵害公民基本权利。

在社会维度层面，生成式人工智能作为驱动经济增长的新引擎，正深刻重塑产业格局，但同时可能加剧就业结构失衡与社会分化风险。治理框架应聚焦包容性发展目标，通过完善社会保障体系、实施职业再培训计划等措施，构建技术普惠机制，缩小数字技能鸿沟。此外，需加快构建适应新技术特征的法治监管体系，完善数据产权界定、算法责任认定等法律规范，防范技术滥用引发的系统性风险[15]。

在技术维度上，生成式人工智能的快速迭代需要对治理机制进行动态调整。治理策略应兼具战略前瞻性与实施灵活性，通过建立跨学科协同创新平台、深化国际技术治理合作等路径，构建开放包容的技术发展生态。同时，需完善技术伦理准则体系，将社会责任融入技术标准制定，探索商业价值创造与社会价值实现的平衡机制。例如，可建立技术影响评估–伦理审查–持续改进的闭环治理流程，通过设立创新容错机制、推行技术影响沙盒测试等制度创新，保障技术演进与价值引领的协同发展。

5. 结论

本研究聚焦生成式人工智能的伦理风险及化解路径，全面剖析了其在多方面的表现、产生原因，并提出针对性化解策略。当前亟需解决的伦理困境呈现多种特征：技术理性对交往理性的压制造成了公共领域的“再封建化”风险，知识–权力结构下的数据霸权加剧了社会不公，文化编码偏差加剧算法歧视

的群体性蔓延，数字资本异化作用加剧了社会阶层分化，技术准入壁垒又造成了数字资产分配的不平衡。化解这些交织的伦理风险，不能仅依赖技术修补，而须在“责任伦理”的指导下，重建被技术模糊的责任链条，培育具备批判意识的理性公众，从技术、法律、伦理、国际合作等多维度入手，才能有效化解风险，推动生成式人工智能技术健康、可持续发展，使其更好地服务于人类社会。

参考文献

[1] 贺苗, 褚星波. 生成式人工智能的伦理风险与规范选择[J]. 自然辩证法研究, 2024, 40(11): 94-101.

[2] 王锋. 人工智能对道德的影响及其风险化解[J]. 江苏社会科学, 2022(4): 44-50+242.

[3] 陈鹏. 人工智能人格权确认的道德风险和法律困境[J]. 西部法学评论, 2018(6): 1-8.

[4] 范佳丽. 生成式人工智能风险预防探究[J]. 合作经济与科技, 2024(10): 190-192.

[5] 李扬, 李晓宇. 康德哲学视点下人工智能生成物的著作权问题探讨[J]. 法学杂志, 2018, 39(9): 43-54.

[6] 段伟文. 前沿科技的深层次伦理风险及其应对[J]. 人民论坛·学术前沿, 2024(1): 84-93.

[7] 孙那, 鲍一鸣. 生成式人工智能的科技安全风险与防范[J]. 陕西师范大学学报(哲学社会科学版), 2024, 53(1): 108-121.

[8] 李颖. 生成式人工智能 ChatGPT 的数字经济风险及应对路径[J]. 江淮论坛, 2023(2): 74-80.

[9] 李健. 出版视域下生成式人工智能创作物的权利归属与制度应对[J]. 出版发行研究, 2023(3): 41-47.

[10] 郭传凯. 人工智能风险规制的困境与出路[J]. 法学论坛, 2019, 34(6): 107-117.

[11] 邹瞳, 张景玥. 生成式人工智能技术的伦理风险防治[J]. 湖北第二师范学院学报, 2023, 40(11): 58-63.

[12] 谭九生, 范晓韵. 算法“黑箱”的成因、风险及其治理[J]. 湖南科技大学学报(社会科学版), 2020, 23(6): 92-94.

[13] 王悠然. 人工智能治理需具有前瞻性[N]. 中国社会科学报, 2022-07-15(003).

[14] 雷浩然. 论生成式人工智能科技伦理的敏捷治理[J]. 华北水利水电大学学报(社会科学版), 2025, 41(1): 50-58.

[15] 兰立山. 技术治理的伦理风险及其应对之策[J]. 道德与文明, 2023(5): 159-167.