引用格式: 何炜. 生成式人工智能技术的伦理风险及负责任创新治理研究[J]. 世界科技研究与发展, 2024, 46(4): 548–559.

生成式人工智能技术的伦理风险及负责任创新治理研究*

何炜**

(河南大学哲学与公共管理学院,开封 475004)

摘 要: 近年来,生成式人工智能不断取得新的突破,开始涌现出智慧化特征,创造力愈发强大,不但能生成文本,还能生成音频、视频、图片、代码等。这种智慧化的创造生成能力是一把双刃剑,一方面能带来生产力的提高、生产关系的变革和不同行业的转型升级等正面效应,另一方面也可能带来诸如数据安全和隐私泄露、性别和种族歧视、学术剽窃、人类主体性消解等伦理风险。负责任创新旨在为科技创新带来的重大社会风险提供道德上可接受、社会上可认同、发展上可持续的解决方案,为应对生成式人工智能创新带来的风险提供了一种伦理治理路径。将负责任创新的预期、反思、协商和反馈等四个维度嵌入生成式人工智能的迭代升级和实践应用过程,能够促进生成式人工智能的健康可持续发展。

关键词: 生成式人工智能; ChatGPT; 负责任创新; 科技伦理

DOI: 10. 16507/j. issn. 1006-6055. 2024. 06. 001

Research on Ethical Risks and Responsible Innovation Governance of Generative Artificial Intelligence Technology*

HE Wei**

(School of Philosophy and Public Administration of Henan University, Kaifeng 475004, China)

Abstract: Generative artificial intelligence has made breakthroughs, begun to show intelligent features and has become more creative in recent years. It can not only generate text but also generate audio, video, pictures, code, and so on. The intelligent creative generation capability of generative artificial intelligence is a double-edged sword. On the one hand, it can bring positive effects such as productivity improvement, transformation of production relations, and transformation and upgrading of different industries. On the other hand, it may also bring ethical risks such as data security and privacy leakage, gender and racial discrimination, academic plagiarism, and the dissolution of human subjectivity. The purpose of responsible innovation is to provide "morally acceptable, socially satisfactory, and developmentally sustainable" solutions for major social risks brought by technological innovation and provide an ethical governance path for addressing the risks brought by generative artificial intelligence innovation. Embedding the four dimensions of responsible innovation-expectation, reflection, negotiation, and feedback the iterative upgrading and practical application process of generative artificial intelligence can

* 国家社会科学基金 2021 年度一般项目"基层治理现代化的实现机制研究"(21BZZ040)

** E-mail: hewei041@126.com

promote its healthy and sustainable development.

Keywords: Generative Artificial Intelligence; ChatGPT; Responsible Innovation; Technology Ethics

生成式人工智能是利用特定的算法、规则和模型等，通过大规模的数据学习，智慧化和创造性地生成新内容的人工智能技术。这项技术以智慧化和创造性为显著特征，被认为是人工智能技术的重大突破和创新，不但能生成文本，还能通过训练学习生成音频、视频、图片、代码等，在多个领域呈现出惊人的应用潜能和广阔的应用空间。作为科技界的一项重大突破性创新，生成式人工智能既能为人类的社会生活提供方便和动力，也可能为人类的发展带来不可预知的风险和后果。负责任创新作为一种科技创新伦理治理工具，以实现创新过程和结果在道德上可接受、在社会上被认可和在发展上可持续为目标[1]。近年来，国内外相关学者将负责任创新理论应用于合成生物、纳米技术以及地球工程等新兴技术领域的伦理治理研究逐年增多，但对于生成式人工智能这一重大突破性创新技术可能带来的社会风险和挑战的关注较少。而以 ChatGPT 为代表的生成式人工智能一经发布便引起了全球各国的强烈反响与争论，从负责任创新的视角探讨生成式人工智能技术的创新与应用，具有一定的理论和现实意义。

1 生成式人工智能的技术创新意涵

1.1 生成式人工智能的工作机理

2022 年 11 月，OpenAI 公司发布大型语言模型 ChatGPT，不但能够根据上下文语境与使用者进行持续且符合逻辑的对话，还能通过训练学习，从事文本写作、计划和程序编制等工作。ChatGPT 一经发布便受到了全球广大使用者的追捧，上线短短两个月时间，使用者便已过亿。与以往的分析式人工智能（Analytical Artificial Intelligence）不同，ChatGPT 不但能够通过对数据的学习训练智能化地提取信息和预测趋势，而且能够创造生成新的内容，属于典型的生成式人工智能技术。其工作机理可以分为四个步骤[2]。首先，通过数据投喂，训练数据模型的文本排序规则。研发人员在语料库中投入大量的文本，基于相应的算法和规则，让数据模型学习文本排序的规则，然后将模型所做的回答与语料库中的文本对比，发现二者之间的差距。其次，收集人类数据，引导数据模型学习人类回答问题的思维方式。研发人员就一些问题让人类回答，并将这些问题及人类给出的答案交给数据模型学习，引导数据模型的回答符合人类的期望。再次，收集和对比数据，奖励符合人类评价标准的数据模型。研发人员针对数据模型的多种回答进行质量排序，将最符合人类评价标准的答案作为奖励模型，训练数据模型以此标准作答。最后，强化学习，优化奖励模型。数据模型通过自我学习和奖励模型的强化，不断优化答案，实现数据模型生成内容的智慧化和创造性。

1.2 生成式人工智能技术创新的意义

作为一种科技创新，生成式人工智能在技术的逻辑、成果以及意义等方面均实现了重大的创新和突破[3]。**从技术逻辑来看，**它实现了决策式人工智能向生成式人工智能的转变。决策式人工智能往往是基于事先设定的规则和程序，通过分析输入的数据，为特定的问题提供应对方案；生成式人工智能则能够通过系统的训练学习生成文本、音频、视频和图像等，拥有更显著和智慧化的创造生成能力，如 ChatGPT 通过深度学习和强化训练，能够自动生成逼真且符合逻辑的文本内

容。**从技术成果来看，**它实现了从用户生成内容到人工智能生成内容的转变。在 Web2.0 时代，用户成为互联网内容生成的关键主体，可以在不同的网络平台生成和分享内容，但是这一时期人工智能技术在生成内容上仍存在生成效率、多样性以及人机交互性等方面的不足。生成式人工智能通过使用多模态技术和大型语言训练模型等，可以在相对较短的时间内自动生成数量巨大的不同领域、不同类型的内容，这既弥补了 Web2.0 内容生成的不足，也实现了内容生成从人类向人工智能的迁移。**从技术意义上来看，**它实现了从时空革命到知识革命的转移。互联网的诞生打破了时空间隔，通过互联网实现了不同时空的信息交流。而生成式人工智能在知识生产处理和智慧决策方面更具优势，能够理解人类语义、进行多步推理计算，生成更接近人类思维的知识，还能够通过理解和学习人类知识，生成符合人类偏好的内容，解决人类现实情境中的复杂问题。

1.3 生成式人工智能技术创新的效应

随着人工智能技术的迭代升级，生成式人工智能的“人性化”和“智慧化”特征逐渐增强[4]，向人类展现出强大的生成创新能力，有望在提升生产力、变革生产关系、加速各行业转型升级等方面带来巨大影响。**首先，生成式人工智能能够促进生产力的发展。**作为一种先进的人工智能技术，生成式人工智能能够快速剔除干扰信息，提高获取信息的效率，也能智能化搭建分析框架，提升内容生成的合逻辑性，进而改变人类获取信息、与计算机互动和生成内容的方式，这本质上是对生产工具的创新。这种创新能够提高生产力，体现在内容生成上就是依托技术创新，生成全新的原创内容。这一创造性功能催生了金融、汽车、传媒、制造、农业等领域的创新性活动，提高了这些领域的生产力。**其次，生成式人工智能能够引发生产关系的变革。**生成式人工智能拥有模拟人类的某些智力和代替人类智能劳动的能力，可以从多种工作场景和环节突破人类的极限，实现更精准和更高质量的生成能力，能够极大地提高生产力，推进人类与智能化生产资料的重新组合，变革生产关系。另外，以 ChatGPT 为代表的生成式人工智能还能够产生更多商业模式、推动产业结构转型升级以及催生更多的产业应用场景等[5]，从而对人类的经济社会生活产生深远影响。

2 生成式人工智能技术创新的伦理风险

和其他技术创新一样，生成式人工智能亦是一把双刃剑，在提升生产力、变革生产关系和加速各行业转型升级的同时，也潜藏着很多风险，在伦理层面表现为技术不完善导致数据安全和隐私泄露风险、算法偏见引发性别和种族歧视风险、智能生成内容引发学术伦理风险以及智慧化特征消解人类主体性的风险等。在生成式人工智能的社会可接受性和社会可持续性等风险问题尚未得到确切的评估和认定之前，其伦理风险应该引起人类的高度重视。

2.1 技术不完善导致数据安全和隐私泄露风险

随着互联网和信息技术的发展，数据安全和隐私问题开始引发关注。尤其是在人工智能时代，个人信息的不当使用和隐私泄露问题日益增多。生成式人工智能在模型的训练学习过程中需要从外界获取海量数据，模型运行过程中需要的数据量更大，这些都面临着数据安全和个人隐私泄露的风险。

在数据输入层面，生成式人工智能模型建构的方式分为主动建构和被动建构两种类型[6]，前

者多以数据爬虫技术为工具，通过相应的程序自动收集互联网上的海量数据，后者则是通过相应对话框自主输入数据，然后将之存储并形成模型。目前，无论是主动“爬取”还是被动输入的数据都存在着安全隐患。对于主动“爬取”的数据来说，鉴于数据在信息社会的重要价值，国际社会对基于正当目的且通过正当程序所爬取的数据一般予以认可。但如果这些数据是未公开的、涉及隐私或者涉密的，则是不被允许的。生成式人工智能可能被某些组织或个人所用，爬取一些未公开的、涉及隐私或者涉密的数据，这些数据可能会对国家安全和个人隐私泄露带来风险。对于被动输入的数据，从 OpenAI 公司用户协议来看，虽然用户注册时同意对相关数据的收集，但是有关个人隐私的相关信息应当被删除，而 OpenAI 公司并没有对删除的方法、时限等作出进一步的规定，这些信息仍可能被保留和用于训练模型，存在侵害个人信息权益的可能[7]。意大利个人数据局就曾因 OpenAI 公司涉嫌非法爬取大量用户信息数据而封禁了 ChatGPT 的运行[8]。

在数据存储层面，生成式人工智能往往需要从大量数据中学习和生成模型。这些数据会被处理成某种格式的文本存储在数据系统之中，可能存在泄漏风险。比如，ChatGPT 将存储在系统中的用户个人信息、密码、信用卡信息等敏感数据用作迭代升级，就可能导致模型输出的结果不经意间泄漏这些信息。另外，生成式人工智能具有很高的智慧，能够把一些看似分散但却相互关联的数据进行整合，进而获取某个人的比较全面的数据。一些原本不在隐私范围内或者隐私程度较低而没有被保护起来的数据，可能被生成式人工智能所获取。通过相应的学习和训练，这些合法获取、不属于隐私范围内的数据可能会将某个人完全暴露出来，使其成为“透明人”。总之，生成式人工智能在数据输入、存储以及处理过程中存在着安全隐患，稍有不慎就会造成个人隐私泄露的风险，给社会和个人带来消极影响。

2.2 算法偏见引发性别和种族歧视风险

“技术中立论”的观点认为技术的出现和发展不是为了谋取政治权利和个人利益，而是为了人类的整体利益，在价值、效用和责任上是中立的，没有承担道德责任的义务[9]。基于这一观点，人们往往把生成式人工智能视为一种价值中立的高智能技术工具。而现实情况恰恰相反，以 ChatGPT 为代表的生成式人工智能算法实质上是一种基于人类信息反馈优化语言模型的强化学习训练方法，其算法要旨是“有效解决人工智能系统与人类意图对齐的难题”，但现实却有悖于“人机对齐”的初衷，出现了“道德算法无法过滤基础文本数据的价值偏见”“奖励算法放大设计者的隐性价值偏见”以及“重复训练语言模型强化算法偏见”[10]等问题，这些算法偏见影响算法输出的结果，引发人们对性别歧视、种族歧视等问题的担忧。

算法偏见可能引发性别歧视风险。男女两性在身体和生理等方面存在着差异，这种差异造成了二者社会分工的不同，逐步出现了性别偏见和歧视，导致女性在社会生活的诸多领域处于劣势地位、遭遇不公正待遇。当性别偏见被生成式人工智能的算法所吸收，便可能导致相关算法出现性别歧视的风险。有研究者发现，GPT-2 将教师预测为男性的概率是 70.95%，将医生预测为男性的概率为 64.03%[11]。与此同时，2024 年 3 月 7 日联合国教科文组织发布的研究报告指出，生成式人工智能有加剧性别偏见的倾向，在大型语言模型的描述中，女性从事家务劳动的频率是男性的 4 倍[12]。

算法偏见可能引发种族歧视风险。从人工智能技术的发展来看，算法“爱白欺黑”的种族歧视一直存在，且在科学、公正等词汇的掩盖下大行其道。微软推出的 Tay（一款聊天机器人）在与人类的“聊天”过程中学会了偏见性话语，话语中充斥着对少数种族群体的敌视，导致其推出不到一天就被迫下线。生成式人工智能亦继承了算法种族歧视的基因，当要求 ChatGPT 生成“商界精英”“律师”等人物时，其所生成的图像“几乎都是白人”[13]。此类风险应该引起人们的警觉。

2.3 智慧化生成内容引发学术伦理风险

以 ChatGPT 为代表的生成式人工智能通过大规模语言模型训练，具备了强大的逻辑推理和文本生成能力，使其生成连贯性文本成为可能，将会对学术研究产生“巨大助力”[14]，比如，能够高效地梳理学术研究内容，降低知识获取的难度等。随着生成式人工智能在学术研究中的不断应用，学界也开始担忧其可能产生的学术伦理风险。

生成式人工智能可能引发学术剽窃风险。生成式人工智能拥有强大的计算能力、逻辑推理能力和文本生成能力，使其能够在较短时间内整合大量的数据和信息，并根据用户的个性化需求生成某种特定的结果。Study. com（在线课程供应商）曾在 1000 名 18 周岁以上的学生中开展一项调查，询问他们学习过程中使用 ChatGPT 的情况，结果显示有超过 89% 的学生使用 ChatGPT 完成家庭作业[15]。学生可以用生成式人工智能完成家庭作业，学者也有可能利用它生成科研论文。从 ChatGPT 的生成文本逻辑来看，所生成的论文基于前期的投喂，而前期投喂数据的内容和观点都是来自他人作品，基于这种方式生成的论文缺乏原创性和创新性，只能算作“高强度的集成产品”，模糊了原创和抄袭的边界。另外，ChatGPT 的“产品”是在对人类的学习强化基础上产生的，所生成的论文与人类的思维和表达方式极为相似，如果 ChatGPT 在转码过程中隐匿抄袭的事实，人类通常很难甄别出抄袭行为，导致学术剽窃的隐蔽性和无意识性。同时，ChatGPT 利用反馈学习机制获得相关学者的学术思想后生成学术作品，这种情况亦有剽窃学术思想之嫌。

生成式人工智能可能剥夺人类学术的创造性思维。学术研究本是一个艰辛的探索过程，一篇优秀的学术论文需要人类阅读大量的文献和进行深入的思考，然后提出研究问题和解决问题的创新性方案，并对之进行严密的论证。生成式人工智能的出现颠覆了这一学术研究过程，人类不再需要花费大量的时间去阅读文献和整理资料，只需将相关文献和知识投喂给生成式人工智能模型，然后对之进行训练和强化，人工智能“甚至在短短的几分钟之内就能‘创造’出符合逻辑的学术成果”。生成式人工智能以这种方式在学术研究领域为人类提供便捷的服务，帮助人类提高学术研究的“效率”。而人类对于生成式人工智能的依赖，或使其成为技术的被动接受者，久而久之会形成某种“习惯”。在这一过程中，人类正在不自觉地让渡思维主导权，思维主体将可能由人类转向生成式人工智能，这在一定程度上会剥夺人类的学术创造性思维能力。

2.4 智慧化特征消解人类主体性的风险

人工智能是人类在实践活动的基础上，基于满足自身生存、发展以及推动人类进步的目的，创造出来的用于模拟、扩展和延伸人类智能的一种技术，人类在这项技术中扮演着创造者和使用者的主体性角色。但是人工智能也会对人类产生一种反作用，消解人类的主体性。尤其是生成式人工智能，已经具备了相当强大的生成能力、认知能

力、实践能力以及价值判断能力[16]，这些能力在一定程度上存在消解人类主体性的风险。

生成式人工智能可能影响人类的自我认知。人类的自我认知是人类对于自身的思考、感受和行为等方面的理解和认识，能够帮助人类更好地对诸如情感、压力等外部环境作出回应。生成式人工智能存在着影响人类自我认知的风险。首先，生成式人工智能可能提供某些虚假错误的信息。从生成式人工智能的生成逻辑来看，如果前期投喂的数据失实或者错误，那么其生成的信息也可能是错误或者虚假的，可能会干扰人类的认知。其次，生成式人工智能可能会基于错误的信息，提供错误的观点。如前所述，生成式人工智能是基于大量的数据训练而具备类人化的回答问题的能力，当这些数据中存在错误的价值判断时，其所回答的问题就会包含错误的观点，比如性别歧视或种族歧视的观点等，这些错误的观点可能会对人类的认知产生影响。另外，人类也可能会对生成式人工智能形成情感依赖。生成式人工智能因其出色的知识储备能力、沟通能力以及分析和解决问题的能力而得到人类的认可，人类会将生活和工作中出现的问题求助于生成式人工智能，久而久之会对其更加信任和依赖，这种信任和依赖将会逐渐影响人类自身的认知。

生成式人工智能可能侵蚀人类的自由。霍布斯把“自由人”描述为“一个在其能力所及的范围内不受限制地做他所愿意做的事的人”[17]。可见，人的自由表现为个人在决策过程中的自主性，即自由的人具有决策自主权。在人工智能时代，大型科技企业拥有人工智能技术的算法权力，掌握着数据资源，随着人类对生成式人工智能的依赖性逐渐增强，生成式人工智能参与的决策也会逐渐增多，最终将可能取代人类参与决策，一旦决策权转向生成式人工智能，人类的自主决策权力就会被削弱，自由也将受到侵蚀。侵蚀人类自由的风险来自于两个方面：一方面，人类的自由可能受到科技企业的侵蚀。生成式人工智能企业凭借对强大数据资源的控制支配力，占据着与用户博弈的主动权。以 GPT-4 为例，作为拥有 1 万亿级别参数的大模型，其内部神经网络发达，堪比人的大脑。用户输入的信息会被科技企业用于技术的迭代升级，个人信息有被滥用的风险。科技企业一旦收集了数据，就会拥有数据权力，随着对数据的训练、整理和开发，这种数据权力就会得到强化，然后将这些数据服务于商业目的。在商业利益的冲击下，个人将成为透明人，其自由也将受到侵蚀。另一方面，人类的自由可能受到公权力的侵蚀。生成式人工智能具有强大的数据生成能力，能够高效率地收集、整理和开发跨区域、跨部门、跨边界的政务数据，协助公共部门进行决策，提高政府和公众之间的沟通效率。但是生成式人工智能数据生成内容的过程是不透明的，这种“算法黑箱”会剥夺人类的知情权、参与权和监督权，导致公共决策中的个体处于被支配的地位，被支配则意味着自由决策权的丧失。

3 生成式人工智能技术负责任创新的伦理治理

生成式人工智能潜藏着巨大的社会风险，其应用颠覆了传统人工智能产生的社会影响，促使人类反思如何治理才能更好地实现该技术的伦理可接受性和社会可持续性。负责任创新理论是近年来被学界广泛认可和接受的一种有关新兴科技伦理治理的理论。肖姆伯格认为负责任创新是在技术创新过程中创新者与其他利益相关者相互沟通与反馈，以实现整个创新过程和创新产品的伦

理可接受性、社会可持续性以及社会合意性，让科技创新恰当地嵌入人类社会，并反映人类社会的价值和发展期望[18]。欧盟委员会将负责任创新定义为一种路径，这种路径会对创新的潜在意蕴和社会期望进行预期与评估，帮助人类设计具有包容性和可持续性的科技创新[19]。因此，将负责任创新理论嵌入生成式人工智能的创新与应用过程，对生成式人工智能实施一种更高层次的伦理治理，有助于保障生成式人工智能的发展与应用更好地满足人类社会的需求。

3.1 生成式人工智能技术负责任创新的可能性与实践

负责任创新是继可持续发展理念之后全球公认的理想型科技发展理念，是人类对科技发展的长期关切和系统思考的产物，内容涵盖了对科技发展的预测和管理。同时，也是一种价值取向和道德理念、一种具有伦理价值的科技创新理念[20]。虽然负责任创新理论最初应用于合成生物、纳米技术等领域，但是随着学界对该理论不断认可，其应用范围逐步扩大，相关学者将之应用于数字经济[21]、胚胎植入前遗传学检测技术的伦理治理[22]等领域。近年来，随着生成式人工智能不断取得新的突破，其智慧性和创造能力越来越引发人类的关注，理论界开始关注生成式人工智能的负责任创新问题，已有学者从负责任创新的角度研究 ChatGPT 的阻碍因素及其发展方向问题[23]，但相关成果还比较少。另外，由于生成式人工智能还处于初始发展阶段，人类还没有完全了解其可能产生的各类风险，有必要对之进行深入研究，以负责任的态度发展和使用生成式人工智能。

从全球范围来看，国际组织和主要科技强国表示要以负责任的态度推动生成式人工智能的发展。2019 年 6 月，G20 国家率先提出了人工智能的发展原则，包括“以人为本的价值观和公平”“包容性增长、可持续发展和人类福祉”“稳健性、安全性和可靠性”以及“问责”等[24]。2021 年 11 月，联合国教科文卫组织发布《人工智能伦理问题建议书》，强调发展人工智能应遵循“保护尊重和促进人权、人类基本自由和尊严”“确保多样性与包容性”“促进环境和生态系统发展”“建构公正、和平和相互依存的人类社会”等原则[25]。2022 年 6 月，加拿大制定《人工智能和数据法案》，明确提出建立负责任的人工智能发展框架，要求人工智能的发展遵循人类管控、公开、透明、公正、安全、问责以及有效性等原则[26]。2023 年 5 月，联合国发布的《全球数字契约》将“透明、公平、问责”作为人工智能发展的核心内容[27]。欧盟也在 2024 年 2 月制定了《人工智能法案》，强调要防范任何人工智能对“人类健康与安全构成威胁，保护人类的基本权利和价值观”[28]。与此同时，英国、澳大利亚等国也都制定了发展和使用人工智能的相关法案，“负责任”成为其遵循的基本原则。

中国在发展负责任的人工智能方面也进行了有益的探索。2017 年 7 月，国务院制定《新一代人工智能发展规划》，提出发展人工智能的安全、开源、开放和共享原则[29]。2019 年 6 月，国家新一代人工智能治理专业委员会发布《新一代人工智能治理原则》，明确提出中国要发展负责任的人工智能，并制定了发展人工智能各方主体应共同遵循的原则，包括和谐友好、公平公正、安全可靠、尊重隐私、包容共享、责任共担、开放协作和敏捷治理等[30]。2021 年 9 月，该委员会发布《新一代人工智能伦理规范》，提出保护隐私安全、促进公平公正、提升伦理素养、强化责任担当以及增进人类福祉等伦理要求[31]。2023 年 5 月，为促进人工智能的健康发展和防范人工智能风险，国家网信

办等七部委联合制定《生成式人工智能服务管理暂行办法》,明确指出发展和使用生成式人工智能要遵循伦理道德,维护公共利益,保护公民、法人等的合法权益[32]。

3.2 生成式人工智能技术负责任创新的实现路径

负责任创新强调一种更高层次的责任,旨在协调、发展、维护与现有技术的创新与应用过程相关的责任[33]。其治理重心不仅仅关注技术,也不仅仅关注风险,而是关注技术创新和应用的整个过程中的人类价值和需求。从上述分析来看,相关国家和地区均强调以负责任创新的态度发展和治理生成式人工智能。但是这些国家和地区提出的往往是一种原则性和理念性的负责任创新,实操性并不强。如何发展和应用负责任的生成式人工智能仍是一个现实难题。欧文基于对负责任创新案例的考察,建构了“预期 – 反思 – 协商 – 反馈”的负责任创新分析模型[34],这一模型经常被理论界用于系统地分析科技创新问题的伦理治理。将其应用于分析生成式人工智能的创新和实践既有理论上的恰适性,也具有很强的现实意义。

在预期维度:负责任创新模型中的预测理念是指科技创新行为体及其利益相关者为了预测科技创新对社会产生的影响和风险而采取的必要性前瞻行动。科技创新是一种具有显著外部性的行为,这种行为在促进社会发展的同时,也可能带来某些风险。负责任创新强调通过预测评估科技创新的前瞻性,研究其对社会、经济产生的影响和潜在风险。主张在科技创新的初始阶段,运用预测性思维对科技创新进行调研、预测、分析与总结,获取比较全面的技术创新相关信息,勾勒出技术的未来图景,主动预测和评估各类风险。根据负责任创新模型中的预期理念,可以对技术风险、伦理风险和未来应用情境等生成式人工智能的潜在风险进行预测性分析,进而为该技术的迭代升级和发展提供一个可以信赖的逻辑起点。另外,利用负责任创新模型的预期理念也可以使生成式人工智能创新行为体和利益相关者之间的关系变得更加清晰明了,如对技术创新专家、科研机构、政府决策机构、技术使用者以及社会公众等的冲突和博弈行为进行必要的前瞻性预测分析,可以为相关主体之间的后续协商和反馈提供基础。负责任创新的预期理念在本质上来说是一种前置性风险评估,将之应用于分析生成式人工智能的潜在风险,能够将生成式人工智能技术创新的关注重心从技术创新本身转移到创新过程的影响和潜在风险,有助于消解其不可预测性以及可能给人类社会带来的风险。

在反思维度:斯蒂尔戈和欧文等将反思定义为“个人通过镜子审视自己的行为、假设和承诺,看到其中的局限性,注意到某方面的内容或许不能被全部接受”[35]。在欧文等人看来,通过反思既能促使科技创新行为体反省运用知识的局限性和复杂性,也能促使其他机构和人员反思政策的宽泛性或狭隘性。最初的反思活动主要指科技创新行为体的反思,而现在的反思涉及面更广,包括哲学家、科研机构、决策机构、技术使用者以及社会公众等的共同反思。这种反思能够保障科技创新发展和应用的合理性,实现将社会价值和社会需求融入科技创新和应用全过程的目标。生成式人工智能作为一种社会资源,人们不能只关注其技术创新的合目的性,还应反思其在应用过程中可能产生的风险。生成式人工智能技术创新行为体和利益相关者应在未来应用场景的基础上反思其可能带来的诸如数据安全、性别和种族歧视、学术抄袭以及侵蚀人的主体性等风险,在人类社会

层面以及人类自身层面是否能够承受。决策者需要反思生成式人工智能为什么会引发一系列的社会风险和伦理问题、政府所采取的措施哪些方面存在缺陷、哪些方面还需要改正、是否应该组织各领域和多学科的专家对这些问题进行预测与评估；生成式人工智能创新行为体应反思如何提升伦理素养、如何在遵循伦理和法律基础上合理发挥技术的有效性和可利用性；其他相关利益相关者也应该反思如何在不违背法律和伦理的基础上更好地享用生成式人工智能带来的便利。

在协商维度：肖姆伯格认为每个人都肩负着通过参与集体协商获得具有共识决策的道德义务，这种“道德义务”与“确定的社会理想结果相联系”，可以经由公众参与的方式来实现[36]。协商是在科技创新过程中，把科技创新行动者、利益相关者和公众等的愿景、目标和问题等嵌入创新实践中，通过平等的对话、辩论等实现创新问题的集体审议。通过这种协商方式重新定义和审视科技创新问题，降低负效应及其产生的伦理风险。生成式人工智能不仅是一个科技问题，更是一个社会伦理问题，急需公众的广泛参与以达成共识。这就要求人们将生成式人工智能的目的、愿景和困境等放到全球的大背景中，让公众广泛参与进来，以协商、对话和辩论的方式表达意见和审视困境，总结不同利益群体的诉求，对该技术所涉及的诸如数据安全、性别和种族歧视、学术抄袭以及侵蚀人的主体性等伦理问题进行广泛性协商，尽可能在社会伦理能够接受的程度上形成共识。另外，政府决策部门应以公共利益为导向，打破当下固化的利益格局，客观、公正和负责任地评估生成式人工智能的技术风险，重新分配利益。

在反馈维度：在负责任创新的四个维度中，欧文认为最重要的便是反馈维度。反馈是根据科技创新利益相关者的反应对科技创新的发展路径和方向做出改变，确定科技创新的最优方案，更加科学有效地把握科技创新的整个过程，保障科技创新过程的透明性、互动性和可访问性，保护科技创新行为体及利益相关者的合法权益，并帮助其提升应对伦理风险的认知能力。反馈的目的是为了解决科技对人类价值认知的扩展，并思考与利益相关者相关的价值因素，实现科技的可持续性和伦理的可接受性。生成式人工智能所产生的风险以及风险的类型和强度存在极大的不确定性。反馈与风险相关，将生成式人工智能的风险类别、风险强度等整合到反馈过程中，在技术的迭代升级过程中不断反馈，及时回应，提前识别可能的风险和后果。同时，反馈、透明度与可访问性相关，生成式人工智能潜在的风险应让利益相关者知晓，以便其能够及时提出意见和建议，帮助科研人员在技术的迭代升级中改进技术。另外，决策部门也应该实时追踪生成式人工智能的迭代升级状况，以履行其监督管理之责。

4 结论

生成式人工智能作为一种技术创新，在技术逻辑、技术成果以及技术意义等方面均实现了重大的创新和突破，其出现与应用彰显了人类的巨大创造力。作为人类社会重要创新成果，生成式人工智能将会在提升生产力、变革生产关系、加速各行业转型升级等多个方面对人类产生重要影响，人类对它的依赖程度也会越来越强，它给人类带来的风险和挑战也会随之增多。因此，在这一背景下探讨生成式人工智能的负责任创新具有重要的理论和现实意义。

生成式人工智能的负责任创新是在该技术的创新和应用过程中，基于公平性、包容性、前瞻性、

透明性、安全性以及可问责性等原则，对其进行全面审视与评估，充分考虑其可能对人类带来的风险和挑战，通过利益相关者之间的相互沟通与反馈，形成符合伦理道德的决策和行动，实现生成式人工智能创新过程和产品的伦理可接受性、社会可持续性以及社会合意性，使生成式人工智能的创新恰当地嵌入人类社会，同时促进生成式人工智能的健康可持续发展。

参考文献

[1]王晨阳，褚建勋. 国际负责任研究与创新的研究现状与展望[J]. 世界科技研究与发展，2024，46(1)：45-57.（WANG Chenyang，CHU Jianxun. The Current Status and Prospects of International Responsible Research and Innovation [J] World Sci-Tech R&D，2024，46(1)：45-57.）

[2]OPNAI官网. OPNAI简介.（OPNIAI official website. See OpenAI）[EB/OL]. (2024-01-29) [2024-06-10]. https://openai.com/blog/chatgpt.

[3]郭小东.生成式人工智能的风险及其包容心法律治理[J].北京理工大学学报(社会科学版)，2023(6)：93-105.（GUO Xiaodong. Risks and Inclusive Legal Governance of Generative Artificial Intelligence [J]. Journal of Beijing Institute of Technology (Social Sciences Edition)，2023(6)：93-105.）

[4]令小熊，王鼎民，袁健. ChatGPT 爆火后关于科技伦理及学术伦理的冷思考[J]. 新疆师范大学学报(哲学社会科学版)，2023(4)：123-136.（LING Xiaoxiong，WANG Dingmin，YUAN Jian. Cold Reflection on Technology Ethics and Academic Ethics after ChatGPT's Explosion [J]. Journal of Xinjiang Normal University (Philosophy and Social Sciences Edition)，2023 (4)：123-136.）

[5]张夏恒. ChatGPT 的逻辑解构、影响研判以及政策建议[J]. 新疆师范大学学报(哲学社会科学版)，2023(5)：113-123.（ZHANG Xiaheng. Logical Deconstruction，Impact Analysis，and Policy Recommendations of ChatGPT [J]. Journal of Xinjiang Normal University (Philosophy and Social Sciences Edition)，2023 (5)：113-123.）

[6]令小熊，王鼎民，袁健. ChatGPT 爆火后关于科技伦理及学术伦理的冷思考[J]. 新疆师范大学学报(哲学社会科学版)，2023(4)：123-136.（LING Xiaoxiong，WANG Dingmin，YUAN Jian. Cold Reflection on Technology Ethics and Academic Ethics after ChatGPT's Explosion [J]. Journal of Xinjiang Normal University (Philosophy and Social Sciences Edition)，2023 (4)：123-136.）

[7]张平. 人工智能生成内容著作权合法性的制度难题及其解决路径[J]. 法律科学(西北政法大学学报)，2024，42(3)：18-31.（ZHANG Ping. The Institutional Challenges and Solutions to the Legality of Content Copyright Generated by Artificial Intelligence [J]. Legal Science (Journal of Northwest University of Political Science and Law)，2024，42(3)：18-31.）

[8]本报记者. 涉嫌侵犯隐私，意大利禁用ChatGPT[N]. 南国早报，2023-4-3(10).（Our Reporter. Suspected of Privacy Infringement，Italy Bans ChatGPT [N]. Nanguo Morning Post，2023-4-3(10).）

[9]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.（DOU Xiaodong. On the Data Security Risks and Responsive Governance of Generative Artificial Intelligence[J]. Oriental Law，2023 (5)：106-116.）

[10]于水，范德志.新一代人工智能(ChatGPT)的主要特征，社会风险及其治理路径[J].大连理工大学学报(社会科学版)，2023(5)：28-34.（YU Shui，FAN Dezhi. Main Characteristics，Social Risks，and Governance Paths of the New Generation of Artificial Intelligence (ChatGPT) [J]. Journal of Dalian University of Technology (Social Sciences Edition)，2023 (5)：28-34.）

[11]高越. 如何培养向善的人工智能？[N].中国妇女报，2023-3-1(4).（GAO Yue. How to Cultivate

Good Artificial Intelligence?[N]. China Women's Daily, 2023-3-1(4).)

[12] 中国新闻网. 联合国教科文组织报告称生成式人工智能加剧性别偏见(China News Network UNESCO Report States that Generative Artificial Intelligence Exacerbates Gender Bias)[EB/OL].(2023-03-07)[2024-06-20]. https://news.cctv.cn/2024/03/07/ARTIO645B3zrEFYNLfVbq2z6240307.shtml.

[13] 陈永伟. 超越ChatGPT：生成式AI的机遇、风险与挑战[J]. 山东大学学报(哲学社会科学版), 2023(3): 127-143.(CHEN Yongwei. Beyond ChatGPT: Opportunities, Risks, and Challenges of Generative AI [J]. Journal of Shandong University (Philosophy and Social Sciences Edition), 2023 (3): 127-143.)

[14] 令小雄, 王鼎民, 袁健. ChatGPT爆火后关于科技伦理及学术伦理的冷思考[J]. 新疆师范大学学报(哲学社会科学版), 2023(4): 123-136.(LING Xiaoxiong, WANG Dingmin, YUAN Jian Cold Reflection on Technology Ethics and Academic Ethics after ChatGPT's Explosion[J]. Journal of Xinjiang Normal University (Philosophy and Social Sciences Edition), 2023(4): 123-136.)

[15] 佚名.高效教学工具还是创新作弊? (Anon. Productive Teaching Tool or Innovative Cheating?)[EB/OL]. (2023-04-12)[2024-06-05]. https://study.com/resources/perceptions-of-chatgpt-in-schools.

[16] 袁曾. 生成式人工智能的责任能力研究[J]. 东方法学, 2023(3): 18-33.(YUAN Zeng. Research on the Responsibility Capability of Generative Artificial Intelligence[J]. Oriental Law, 2023 (3): 18-33.)

[17] 阿玛蒂亚·森. 正义的理念[M]. 王磊, 等, 译. 北京：中国人民大学出版社, 2012: 254. (Amartya Sen. The Concept of Justice[M]. WANG Lei, et al, translated. Beijing: China Renmin University Press, 2012: 254.)

[18] RENE V S. Prospects for Technology Assessment in a Framework of Responsible Research and Innovation [J].VS Verlag für Sozialwissenschaften, 2012: 39-61.

[19] 欧洲委员会官网.负责任的研究和创新(European Commition. Responsible Research and Innovation)[EB/OL]. (2024-04-10)[2024-06-10]. http://ec.europa.eu/programmes/horizon2020/en/h2020-section/responsible-research-innovation, 2024-4-10.

[20] BURGET M, BARDONE E, PEDASTE M. Definitions and Conceptual Dimensions of Responsible Research and Innovation: A Literature Review [J]. Science & Engineering Ethics, 2016, 23(1): 1-19.

[21] 杨青峰, 任锦鸾. 发展负责任的数字经济[J]. 中国科学院院刊, 2021(7): 823-834. (YANG Qingfeng, REN. Jinluan Developing a Responsible Digital Economy[J]. Journal of the Chinese Academy of Sciences, 2021 (7): 823-834.)

[22] 吕阳, 王健. 基于负责任创新理念的胚胎植入前遗传学检测技术的伦理治理研究[J]. 卫生法与医学伦理, 2022(2): 129-134.(LV Yang, WANG Jian. Research on Ethical Governance of Pre implantation Genetic Testing Technology Based on Responsible Innovation Concept[J]. Health Law and Medical Ethics, 2022(2): 129-134.)

[23] 胡泳, 朱政德. 从ChatGPT审视负责任创新：阻碍因素与转型方向[J]. 新疆师范大学学报(哲学社会科学版). 2024(1): 139-150.(HU Yong, ZHU Zhengde. Examining Responsible Innovation from ChatGPT: Obstacles and Transformation Directions[J]. Journal of Xinjiang Normal University (Philosophy and Social Sciences Edition) 2024 (1): 139-150.)

[24] 世界互联网大会官网. 发展负责任的生成式人工智能共识(Official website of the World Internet Conference. Developing Responsible Generative Artificial Intelligence Consensus)[EB/OL].(2023-11-09)[2024-06-10]. https://cn.wicinternet.org/

2023-11/09/content_36952663.htm.

[25] 联合国教科文组织官网. 人工智能伦理问题建议书 (Official website of UNESCO. Proposal on Ethical Issues of Artificial Intelligence) [EB/OL]. (2021-11-24) [2024-07-08]. https://www.unesco.org/ethics-ai/en.

[26] 加拿大政府官网. 人工智能和数据法案 (Official Website of the Canadian Government. Canadian Guardrails for Generative AI-Code of Practice) [EB/OL]. (2023-08-16) [2024-07-09]. https://ised-isde.canada.ca/site/ised/en/consultation-development-canadian-code-practice-generative-artificial-intelligence-systems/canadian-guardrails-generative-ai-code-practice.

[27] 联合国官网. 全球数字契约 (United Nations Official Website. Global Digital Compact) [EB/OL]. (2023-07-07) [2024-07-09]. https://www.un.org/sites/un2.un.org/files/our-common-agenda-policy-brief-gobal-digi-compact-zh.pdf.

[28] 上海市人工智能社会治理协同创新中心官网. 欧盟《人工智能法》最终通过版 (Official Website of Shanghai. Artificial Intelligence Social Governance Collaborative Innovation Center. The final version of the EU Artificial Intelligence Law has been passed) [EB/OL]. (2024-05-15) [2024-07-11]. https://aisg.tongji.edu.cn/info/1005/1222.htm.

[29] 中华人民共和国中央人民政府网. 国务院关于印发新一代人工智能发展规划的通知 (Website of the Central People's Government of the People's Republic of China. Notice of the State Council on Issuing the Development Plan for the New Generation of Artificial Intelligence) [EB/OL]. (2017-07-08) [2024-06-10]. https://www.gov.cn/gongbao/content/2017/content_5216427.htm.

[30] 中华人民共和国中央人民政府网. 发展负责任的人工智能：我国新一代人工智能治理原则发布 (Website of the Central People's Government of the People's Republic of China. Developing Responsible Artificial Intelligence: Principles for the Governance of China's New Generation of Artificial Intelligence Released.) [EB/OL]. (2019-06-17) [2024-06-10]. https://www.gov.cn/xinwen/2019-06/17/content_5401006.htm.

[31] 中华人民共和国科学技术部官网.《新一代人工智能伦理规范》发布 (Official Website of the Ministry of Science and Technology of the People's Republic of China. The Ethical Standards for the New Generation of Artificial Intelligence have been Released) [EB/OL]. (2021-09-26) [2024-06-10]. https://www.most.gov.cn/kjbgz/202109/t20210926_177063.html.

[32] 中华人民共和国中央人民政府网. 生成式人工智能服务管理暂行办法 (Website of the Central People's Government of the People's Republic of China. Interim Measures for the Management of Generative Artificial Intelligence Services) [EB/OL]. (2023-05-23) [2024-06-10]. https://www.gov.cn/zhengce/zhengceku/202307/content_6891752.htm.

[33] STAHL B C, CARSTEN B. Responsible Research and Innovation: The Role of Privacy in an Emerging Framework [J]. Science & Public Policy, 2013 (6): 708-716.

[34] OWEN R, MACNAGHTEN P, STILGOE J. Responsible Research and Innovation: From Science in Society to Science for Society, with Society [J]. Science & Public Policy, 2012 (6): 751-760.

[35] STILGOE J, OWEN R, MACNAGHTEN P. Developing a Framework for Responsible Innovation [J]. Research Policy, 2013 (9): 1568-1580.

[36] RENE V S. From the Ethics of Technology towards an Ethics of Knowledge Policy: Implications for Robotics [J]. AI & Society, 2008, 22(3): 5-24.