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

李 韬 周瑞春

【摘 要】基于行动者网络理论视角，生成式人工智能的社会伦理风险可以认为是在技术应用过程中由人类行动者和非人类行动者开展行动者转译和网络建构所造成的。从行动者转译与网络建构来看，生成式人工智能可能带来的主要社会伦理风险包括：智能互联下的行动者之“恶”传播风险加剧、行动者社会分工体系的结构性危机加剧、行动者转译的技术异化问题更为严峻、“行动者网络”下的数字安全风险增大、行动者“五际”关系维度的数字平权问题凸显。基于此，从合理界定行动者主体的权责关系、加强行动者转译的全过程算法治理、构建人机“共善”的行动者治理共同体等方面推进生成式人工智能的社会伦理风险治理，确保科技向善健康发展。

【关 键 词】生成式人工智能；社会伦理风险；行动者网络理论；治理理路

【作者简介】李韬，北京师范大学新闻传播学院教授；周瑞春，北京师范大学中国社会管理研究院、互联网发展研究院特聘研究员。

【原文出处】《中国特色社会主义研究》（京），2023.6.58～66，75

【基金项目】本文系北京师范大学互联网发展研究院—中国移动研究院联合研发项目“生成式人工智能技术发展中的科技伦理框架研究”（项目编号：BNUIDI2023）的阶段性研究成果。

随着ChatGPT作为一种预训练生成对话模型的火爆，通过人工智能生成内容，即“生成式人工智能”（AIGC）[①]也引发社会各界广泛关注。所谓生成式人工智能，一般是指基于智能算法、模型与规则等生成文本、图像、音视频等内容的技术。生成式人工智能能够通过学习人类语料库，生成相应对话内容，在与各类应用软件的结合运用下，还能进行跨模态内容的生成转换，这就使得AIGC在各个领域的平民化应用成为可能。在人们憧憬科技革命带来数字红利的同时，技术双刃性可能带来的风险挑战也成为学界探讨的重要议题。

纵观人类历史，任何一次技术变革所带来的各种各样的风险归根结底都会在社会层面扩散蔓延，引发社会结构的调整、社会关系的变迁，影响社会系统建构与运行。马克思就指出“手推磨产生的是封建主为首的社会，蒸汽磨产生的是工业资本家为首的社会”[②]。事实上，技术是人类社会生产的工具性产物，谈论技术风险的根本性意义在于探究技术对人与人、人与社会、人与自然等一系列社会伦理关系可能造成的直接或间接的风险挑战。当下，人工智能技术的飞速发展与应用使得“技术—社会”互动更为频繁，“硅基智能体”的出现更有可能成为与人类面对面的主体性存在，扩展人类社会伦理关系的内涵与外延。由此可见，在纷繁芜杂的生成式人工智能技术风险议题中，社会伦理风险理应成为焦点所在。所谓生成式人工智能的社会伦理风险，概言之就是生成式人工智能技术及应用在社会层面构成的伦理风险，亦即人类使用生成式人工智能技术产品

和服务对社会系统良序运行可能造成的潜在危机、负面影响、间接危害乃至直接破坏，既包括人与自我、人与社群、人与社会、人与自然等关系维度，也包括“人—机”关系维度。那么，生成式人工智能的深入发展和应用，将会带来哪些方面的社会伦理风险呢？面对种种不确定性风险又该从哪些方面入手加以应对？凡此种种，便成为本文研究试图聚焦的重要问题域。

一、文献述评

当前，生成式人工智能的风险已经进入国内外学界视野。一是生成式人工智能内容生产可能带来的虚假信息泛滥[③]、个人隐私数据泄露、意识形态安全冲击[④]等直接性风险，二是生成式人工智能在教育[⑤]、新闻[⑥]、司法[⑦]等领域应用的间接性风险。从现有文献来看，涉及社会伦理风险的研究较少，而且其中大多数还停留在对生成式人工智能道德主体性的理论探讨层面。包括AI是否拥有权利[⑧]、AI权利的基础与来源[⑨]、AI权利是否应当与人类权利保持一致性[⑩]，等等。比如，Wallach和Allen从“功能道德”的角度出发，认为AI具有弱道德的自主性[⑪]；Hauer指出，若是一项任务在由人类执行时需要某种形式的道德权威，那么将同样的任务转移到自主机器、平台和人工智能算法必然意味着道德能力的转移[⑫]；何怀宏认为，面向人工智能未来的伦理学需要将人与智能机器的关系纳入其中，将智能机器的发展“限制在专门化、小型化尤其是尽可能的非暴力的范围之内”[⑬]。总体而言，现有文献从研究内容来看，大多是对生成式人工智能在不同领域的技术应用风险进行点状分析，而对风险产生的原因、主要类型、应对之策等，并未能给出整体性、综合性、系统性的理论阐释；从研究视角来看，侧重在对人类个体、群体、社会的线性技术影响，而并未能将技术作为一种社会建构主体性要素，从“人—机”互动关系维度深入考察其社会伦理风险。

事实上，技术肇始于人类对劳动工具的使用，而使用劳动工具进行劳动生产则是人类从动物界中分化出来的决定性因素之一。由此可见，技术在其产生之初，就对人之为“人”的确证以及社会建构发挥着重要的作用，从农业社会、工业社会到数字社会，人类社会史在某种意义上也可以被视为是一部人与技术的互动互构史、社会伦理关系的变迁史。在此过程中，技术不再是与“人”割裂的客体，而是作为一种“非人类因素”[⑭]参与社会建构的内生性力量。尤其是随着生成式人工智能机器人的出现，在不同领域、不同行业、不同场景下，技术性的“非人类因素”逐渐成为与人类并行的“非人类行动者”，与“人类行动者”开展着对话交流、分工协作的互动互构，伴生着技术带来的社会伦理风险。综上可见，当前迫切需要引入新的理论框架，对生成式人工智能的社会伦理风险研究进行类型化、系统化梳理，对“人类行动者—非人类行动者”互动关系以及社会伦理风险产生的内在机理进行深入剖析，在此基础上提出有针对性的治理之策。

二、行动者网络：一个理论分析框架及其适用

从生成式人工智能研究进展可知，如何从“技术—社会”“人—机”关系维度，探究生成式人工智能作为社会建构行动者的可能风险、规制路径，成为寻找新的理论分析框架的关键。为此，本文拟基于“行动者网络理论”对生成式人工智能带来的社会伦理风险进行深入分析。

所谓“行动者网络理论”(Actor-Network-Theory，ANT)也被称为“异质建构论”，于20世纪80年代由法国社会学家米歇尔·卡龙(Michel Callon)、布鲁诺·拉图尔(Bruno Latour)和约翰·劳(John Law)等人提出。该理论反对现代工具理性将科学技术视为单一“工具”的主客二分法，赋予包括科学技术在内的“非人类因素”以重要地位，认为对于一个社会系统或网络而言，“任何通过制造差别而改变了事物状态的东西，都可以被称为行动者”[⑮]。也就是说，人类和非人类因素的行动能力或参与能力在本质上没有区别，即社会是一个人类和非人类两种类型的行动者相互作用的场域[⑯]。行动者网络理论主要包括三个核心要素，即行动者(actor)、转译(translation)和网络(network)。“行动者”包括人类因素与非人类因素，没有主动与被动、主体与客体之分，更加注重行动中的角色分工与权责分配；“转译”是指不断把其他行动

者的问题和兴趣用自己的语言转换出来的过程，包含着利益博弈与协商动员；“网络”是指各类行动者基于不同社会行动目标，彼此之间形成的流动变化的连接关系。行动者网络理论已经被应用到社会学、传播学、教育学等各个学科领域，已成为探讨科学技术与社会互动演进的重要理论分析方法，正如加拿大学者西斯蒙多(Sergio Sismondo)所言，行动者网络理论已经“演变成一种围绕技术科学的一般社会理论，而不仅仅是技术科学理论”[17]。

本文之所以选择行动者网络理论作为探究生成式人工智能社会伦理风险的理论分析框架，主要基于以下考虑。第一，在生成式人工智能技术应用全过程中，包含着人类行动者与非人类行动者两类行动者主体，成为“技术—社会”视角下社会建构的行动者要素；第二，人们在不同场域使用AIGC产品时，通常是以“人—机”对话下的交互模式推进，这种互动互构实际上就是行动者的转译过程；第三，人类行动者基于不同的社会行动目标，与不同的非人类行动者形成流动性的连接关系，符合行动者网络建构过程。以ChatGPT为例，人类通过与ChatGPT对话，提出问题、表达利益诉求、招募技术计算、动员技术反馈，推动着数字社会的建构进程，实际上扮演着“人类行动者”与“非人类行动者”的社会分工角色。如图1所示，在生成式人工智能技术应用过程中，人类行动者通常包括各类网络接入者、AIGC产品用户、技术群体，乃至人类群体建构的企业、社会组织、政府部门、国家等；而非人类行动者包括网络软硬件设施、AIGC产品以及AIGC背后依托的大数据、云计算平台、人工智能算法等。人类行动者与非人类行动者在不同的AIGC应用场域开展转译互动，基于共同目标进行经验整合、利益博弈、互动互构，主要包括问题呈现、利益赋予、招募和动员四个环节[18]；在不同的应用场域，人类行动者根据不同的社会行动需求，与非人类行动者建立节点关联，开展着行动者网络建构。

具体而言，人类行动者通过对话形式呈现问题、展示问题包含的利益需求、招募和动员非人类技术因素；AIGC产品通过对应的大数据语料库、云计算平台、独特算法等技术因素，提供文本、图像、音频、视频等多模态内容生产。实际上，在网络数字空间，人类行动者基于特定的社会行动目标任务，与非人类行动者不断发生互动互构，实施行动者转译；在得到非人类行动者的信息反馈后，人类行动者还将继续围绕最初的社会行动目标任务，以技术赋权开展

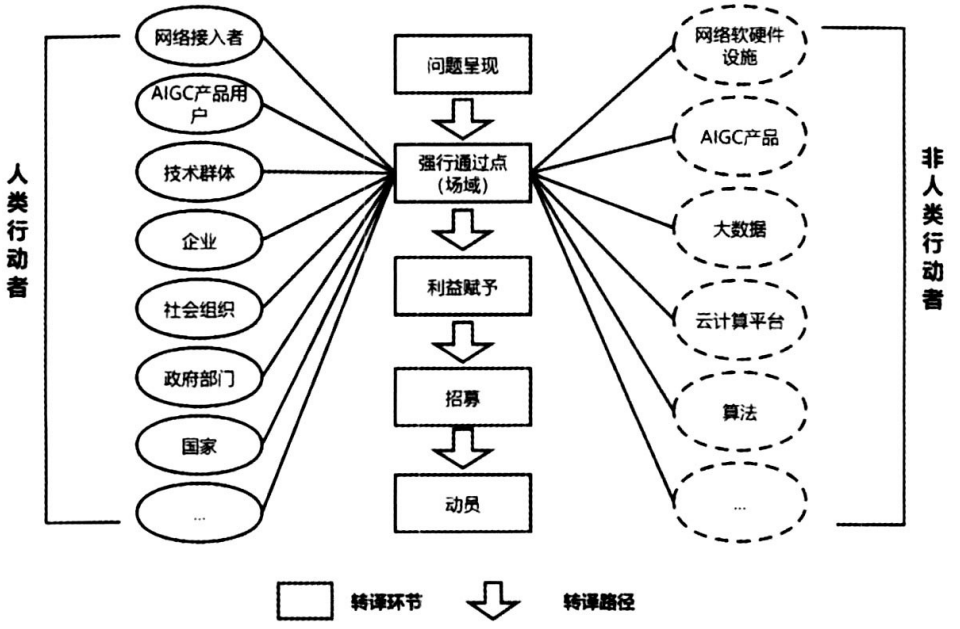

图1 生成式人工智能行动者网络模型

后续社会行动，推进线上线下联动的行动者网络建构。整体而言，人类行动者与非人类行动者反复互动，不断进行着问题呈现、利益赋予、招募及动员的转译过程，伴生着可能带来的社会伦理风险。显然，在不同类型应用场域中，生成式人工智能的转译具有不同的类型化特征，由不同类型的行动者主体主导也会产生不同的社会伦理风险。

三、基于行动者网络理论的生成式人工智能社会伦理风险分析

从行动者网络理论来看，生成式人工智能的社会伦理风险是行动者转译及其网络建构的伴生物。由于人类行动者与非人类行动者之间的角色定位、责任分工、权力关系始终处于动态变化之中，技术应用可能带来的社会伦理风险也会随之转移变迁。因此，有必要从行动者转译过程入手，重新梳理生成式人工智能可能带来的社会伦理风险，厘清人类行动者、非人类行动者在不同类型应用场景下各自应当承担的风险责任，揭示风险产生的内在机理。

如图2所示，基于行动者网络理论分析框架不难发现，人类行动者和非人类行动者都有可能在转译过程中给社会系统运行带来“恶”的风险及后果。人类行动者本身具有“善”与“恶”的主观目的性，而非人类行动者在一定阶段的程序化运行也存在着“技术善”与“技术恶”的双刃性。因此，人类行动者与非人类行动者在转译过程中，也就会相应产生善恶两种不同的现实结果，带来可能的社会风险。生成式人工智能技术应用的社会现实后果，既取决于人类行动者与非人类行动者两者本身的善恶初始设定，也体现着两类行动者主体在转译过程中的权力博弈关系。

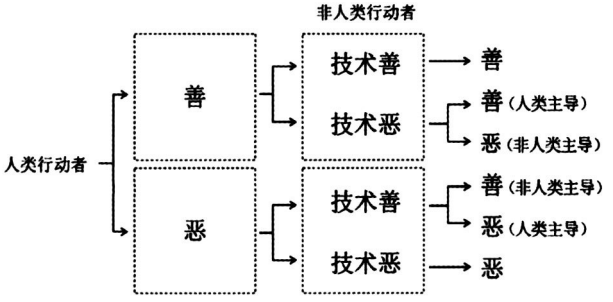

图2 行动者转译的伦理风险路径

（一）智能互联下的行动者之“恶”传播风险加剧

在以往互联网时代，信息传播有着较为明确的“公域”和“私域”之分，即网民对哪些信息可以在网络发布分享、哪些信息只能作为隐私存储在私人计算机设备中有着较为清晰的认知。平台型媒体的出现，使得信息传播泾渭分明的“公域—私域”界线被打破，而生成式人工智能的普及应用更让这一界线日益模糊，各类AIGC产品成为联结公域与私域的智能互联“超级媒体”。一方面，生成式人工智能能够以大数据的形式抓取公域资源，从中提炼出最具传播价值的信息；另一方面，生成式人工智能也能够通过与用户频繁互动对话，对原本存储在私域空间的信息资源进行整合提取，甚至还能提取分析人们脑海中所思所想的隐性信息。通过聚合公域和私域资源的智能互联，生成式人工智能建立起无所不知的“网络中枢”神经系统，建立强大的信息收集、整理、分析、利用、生产、分发、传播、再生产全流程、全自动“生产线”。

在这种智能互联的超级媒体传播模式下，生成式人工智能的行动者之“恶”会被无限放大。不管是人类行动者的“动机恶”还是非人类行动者的“技术恶”，都会在转译过程中被超级媒体不断加权，穿透公域与私域的藩篱，形成更大的“恶”之社会效应，给网络生态环境乃至现实社会健康发展带来巨大的破坏。比如，Tristan针对推特的实验研究表明，人工智能社交机器人已经成为网络社交谣言生产与传播的主要力量之一，如果将社交机器人生产的内容剔除，虚假信息的转发量将会减少70%[19]。美国乔治城大学一项研究也表明，AIGC能够根据对话情境自主“重复叙述、阐述、操纵、说服”“在短短几分钟内设计和制作出有针对性的虚假信息，模仿特定的写作风格，从而有效影响受众的态度”。[20]在各类社交媒体平台上，这些由AI自动生成的大量文本、图像、音频、视频等虚假信息，“不仅难以被人类用户分辨，而且网站自身的过滤系统和剽窃检测软件都无法识别出真假”[21]。

显然，从人类行动者发起问题，到非人类行动者生成虚假信息，是非人类行动者“技术恶”与人类行

动者“动机恶”的行动者转译互动过程。一方面，当“技术恶”掌控转译主导权时，在海量语料库及强大的自然语言生成能力赋权下，AI极有可能会引导人类行动者信任并传播虚假信息、错误信息，并最终导致人类行动者因错误信息引导下产生的错误认知、错误行动；这些错误认知与行动所产生的大量数据，还将会再次成为AI更新自身语料库的大数据来源，为非人类行动者带来更多的错误“源数据”。上述转译过程使得生成式人工智能的行动者转译陷入不良信息批量生产与传播“以讹传讹”的恶性循环之中。另一方面，当“动机恶”掌控转译主导权时，技术中立的AI对人类行动者“唯命是从”，往往会放大“动机恶”的社会效应。比如，各类谣言、色情、暴力内容的AI生产传播，以及“AI换声”“AI换脸”等新型AI诈骗犯罪频发，造成更大的社会负面影响。概言之，生成式人工智能在张扬“动机善”“技术善”的同时，也会加剧“动机恶”“技术恶”的超级媒体传播风险。

(二)行动者社会分工体系的结构性危机加剧

作为人的本质性存在方式，劳动是人的神圣权利，“不仅仅是谋生的手段”，而且本身就是“生活的第一需要”[22]。然而，随着生成式人工智能应用普及，“技术性失业”再次成为削减甚至剥夺人类行动者劳动权利的可能性风险。凯恩斯(John Keynes)在考察20世纪30年代全球经济大萧条时解释了“技术性失业”的主要表征，即“发现了节约劳动力使用的方法(技术)，但又未能及时为由此带来的剩余劳动力找到新的用途”[23]。有学者认为，这样的技术性失业在AIGC时代将更为突出[24]。据麦肯锡报告预测，随着AI技术深入应用，人类社会将面临“AI失业潮”，现有的820种职业中约710种可能会被机器人取代[25]。不管如此规模巨大的AI失业潮是否会到来，“AI失业”必将成为各国需要正视的技术双刃剑效应之一。

从行动者转译过程来看，“AI失业”实际上是非人类行动者在特定劳动生产领域占据了转译主导权，甚至完全替代了人类行动者成为唯一的行动者。实际上，马克思和恩格斯都曾对技术性失业问题作出过深刻论述。恩格斯认为，大规模机器生产的出现使得大批工人失业，变成机器的附属品，机器“对手工劳动的排挤以及分工都达到了高度的发展”[26]；马克思则更为深刻地指出，工人应该反抗的不是机器本身，因为“矛盾和对抗不是从机器本身产生的，而是从机器的资本主义应用产生的”[27]，机器不是造成工人失业贫困的根源，而是应用机器的资本主义生产方式。可以看出，“AI失业”的本质是非人类行动者掌握了转译主导权，对人类中心地位的行动者社会分工体系造成结构性冲击，非人类行动者成为社会分工体系中的重要主体性力量。非人类行动者对原有社会分工进行重新分配，也是社会生产方式的转变过程，使得熟练掌握和使用AI技术的群体在社会分工结构中占据优势位置，而与此相对应的社会关系也随之发生改变。在特定的社会行动场域，非人类行动者在行动者转译的全过程中发挥着引领作用，倒逼人类行动者不断提升自身的AI技能、AI素养、AI意识。一方面，使得未能跟上AI转译步伐的一部分人类行动者“出局”失业，给其余人类行动者带来时刻要提升自身AI综合能力的职业焦虑；另一方面，也通过主导行动者转译的社会权力，对整个人类社会分工体系进行改造更新。

(三)行动者转译的技术异化问题更为严峻

人类行动者与非人类行动者分工协作，通常由人类行动者主导转译过程。然而，随着AIGC进入人类社会生活各个领域的泛在性不断增强，非人类行动者逐渐在转译行动中占据主导地位，甚至“发展成为一种新的外在的异己力量”[28]，使得行动者转译的技术异化问题也将更为复杂。

第一，人类行动者的交往主体地位受到挑战。比如，由于AIGC日益成熟的自然语言生成能力，在文本、图像、音视频多模态转换软件的辅助之下，能够实现人机对话的全场景应用，人们不再需要通过与他人身体在场的面对面交往来共享资讯、学习知识、共通情感，AI机器人成为全方位“完美伴侣”。尤其是情感陪伴型生成式人工智能机器人的智能化加速，让人们更加容易沉溺在一种问题呈现与自动应

答的“自言自语”之中无法自拔，从而走向缺乏面对面人际交往的原子化社会。

第二，行动者之间的社会权力之争日益凸显。出于对技术本身的信赖，非人类行动者可能还会成为传统社会知识精英的“掘墓人”，使得人人皆可AI下的社会共识凝聚更加困难。如此一来，基于共识性认同的社会共同体建构亦将更为脆弱，群体性分歧乃至对抗成为共识匮乏下的新风险。实际上，面对AI“价值对齐”问题的探讨已经折射出非人类行动者对人类行动者自身内部团结的新挑战。甚至在人类行动者日常交往中，以计算结果为行动准绳的技术主义一旦盛行，也将进一步消解基于理解共情的社会信任，加速对人类社会整体性的技术异化。说到底，在社会建构与运行各个层面，人类行动者与非人类行动者的社会权力之争将日益明显，“人类优先”不再是不证自明的社会共识。

(四)“行动者网络”下的数字安全风险增大

福柯(Michel Foucault)指出，现代社会的发展使得人的肉身逐渐被纳入国家权力机制的算计和监控之中，人的生命与国家治理术结合，使得“身体规训的目标和人口调节的目标”[29]都成为一种政治技术，而每一个人时时刻刻都处于一种“全景监狱”[30]监控之下，受到现代社会权力机制无处不在的规训。环顾当下，在大数据、云计算、人工智能、物联网、区块链等数字技术的赋权之下，人类社会已经全面进入跨时空的数字“全景监狱”时代。随着生成式人工智能的全面深入应用，人类行动者基于不同的社会行动目标，与不同的非人类行动者建立起一个个“行动者网络”，这些“行动者网络”可能成为对人类行动者所思所想、所欲所求“一网打尽”的“超级全景监狱”，使得人们数字化生存的一切行为都在数字技术的监控之下。

波斯特(Mark Poster)认为，这样的“超级全景监狱”(Super Panopticon)在无声无息中使得数字权力的毛细血管延伸至社会生活的每一个角落，“把我们的私人行为转化成公开布告，把我们的个人言行转化成一种集体语言”[31]。人类行动者在各个场域与非人类行动者进行转译互动时，包括性别、职业、喜好、商品需求、精神状态、情感状态、价值立场等海量用户信息都可以通过特定方式被采集、标注、整理和分析，AI算法下的用户画像描绘日益精准。当越来越多的社交互动平台嵌入AIGC插件后，这些精准画像将进一步成为对用户定点投放广告、定向推送特定商品的最有力工具。与此同时，大量人类行动者的个人隐私也将在与非人类行动者的转译互动中被收集提取，形成一个个数字资料库，存在数字信息泄露的重大安全风险。

此外，生成式人工智能行动者网络的“超级全景监狱”也将进一步增加国家、政府部门、企业、社会组织、技术社群等人类行动者的数字安全风险。比如，政府部门、企业、社会组织在与非人类行动者转译过程中产生大量文本、图像、音频、视频等信息，加上与社交定位软件联动，将产生涉及领域广、精准度高、实时性强的一手数据，这些数据也必然会成为大国网络空间博弈的重要利器，直接关系到国家主权、安全和利益。美国胡佛研究所在一份关于俄乌战争情报工作研报中就提出，技术一直就是推动情报演变的核心动能，并认为美国情报部门应该建立“第19个情报机构”，致力于“一个专门的、开源的情报机构，专注于梳理非机密数据并辨别其含义”[32]。这里所谓的“非机密数据”，实际上就是指向各类应用软件尤其是AI社交软件中提取的海量用户交互信息，以聊天机器人身份出现的非人类行动者，正是默默“窃取”人类行动者个人隐私乃至用户所在行业、部门、岗位内容信息的“王牌间谍”。

(五)行动者“五际”关系维度的数字平权问题凸显

在生成式人工智能行动者网络建构中，行动者之间的主体关系博弈始终存在。比如，在知识问答型AIGC产品使用中，处于提问一方的人类行动者往往占据着转译的主导权；在文艺创作、绘画设计、程序编程等AIGC产品使用中，非人类行动者的技术性因素发挥着决定性作用，掌握着转译的主导权。显然，行动者之间的主体关系博弈，说到底是对行动者网络建构的数字权力之争。这样的数字权力之争，不只发生在人类行动者与非人类行动者之间，也发生在人类行动者内部的多元主体之间，主要包括“国

际”“群际”“人际”“代际”“超人际”等“五际”社会伦理关系维度。需要说明的是，上述五际关系维度并不是理论上的逻辑周延维度，而是在生成式人工智能行动者网络建构过程中最具代表性的实践维度。

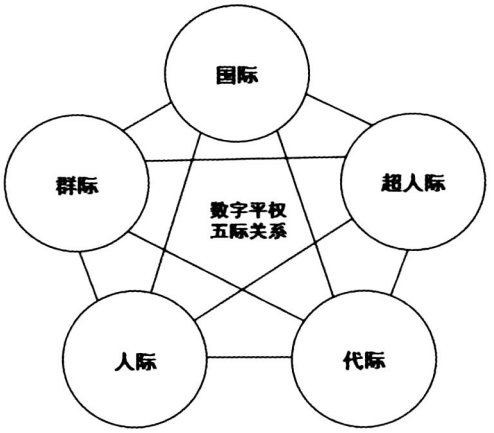

图3 行动者数字平权的五际关系维度

比如，在“国际”关系维度，各国行动者由于在经济实力、人才储备、数字基础设施建设等方面的差异，无法在AI技术发展和应用中保持同步跟进，技术差距越来越大；AI治理话语权也逐渐被AI强国主导，滋生AI霸权主义。在企业、社会组织、技术社群等“群际”关系维度，由于不同行业AI接入应用程度不同，相同行业不同企业、组织和社群之间发展不一，也将导致数字不平等的加剧。在“代际”关系及“人际”关系维度，由于不同年龄段、不同地区的行动者对AI技术掌握水平差异，使得代际行动者、个际行动者的数字不平等不断加剧。随着AIGC产品收费，人类行动者中的富人群体可以通过支付更多费用获得非人类行动者响应速度更快、使用过程更安全、附加功能更强大的科技服务，继而成为“AI富人”；而人类行动者中的穷人群体则将进一步丧失AI时代的平等数字权利，无法在AI接入、AI使用和AI自我可行能力提升上紧跟步伐，沦为“AI穷人”。如此一来，人类社会数字不平等、数字贫富差距将不断扩大，给数字社会稳定和谐发展带来潜在的巨大风险。此外，随着非人类行动者在转译中的主体地位提升，也必然会扩大数字平权的可能边界，使得人类行动者与非人类行动者之间的“超人际”关系维度上的数字平权问题日益浮现。

四、行动者网络理论视域下生成式人工智能社会伦理风险的治理理路

随着生成式人工智能的飞速发展与深化应用，人类行动者与非人类行动者的转译互动和网络建构也日益深入，在常态化运行中不断推进着整个人类社会协同性、结构性、系统性的演进和变迁。因此，对生成式人工智能可能带来的社会伦理风险进行治理，应该从人类行动者、非人类行动者两个向度，聚焦行动者转译和网络建构的全过程进行有针对性的治理。

(一)合理界定行动者之间的权责关系

对生成式人工智能的社会伦理风险进行精准治理，必须要先厘清人类行动者、非人类行动者在转译过程中各自应当承担的风险责任，其关键在于界定两者之间的权利关系和责任关系。

第一，权利分配关系。人类行动者与非人类行动者的权利关系，直接影响到AI社会伦理风险责任分配。有学者在论述人类与AI的权利关系时，梳理了权利理论发展史，认为依据“实力界定权利”理论，“真正影响机器人权利主体地位的客观要素在于机器人同人类之间的实力对比”[33]，若是机器人的社会化程度和各个领域占有率不断提高，其所有权人自然会呼吁立法机关来界定机器人的权利主体地位。普特南(Hilary Putman)甚至指出，机器人同人类的思维方式和规则如出一辙，在技术发展到一定阶段，仅仅是人类的主观态度才影响了机器人是否为“人”[34]。也就是说，人类行动者与非人类行动者的权利关系，主要取决于非人类行动者对人类行动者本身生产生活的影响力。如果非人类行动者全面融入人类社会生活，并在社会建构层面上发挥越来越大的主体性作用，其权利也将获得人类主观上的充分尊重，乃至能够获得与人类同等的权利主体地位。

第二，责任分配关系。非人类行动者权利地位的提升，意味着在技术逻辑下其相应的风险责任分配比重也应随之提高。在对非人类行动者实施风险责任分配时，一个重要的争议在于非人类行动者是否能够成为真正的责任主体。客观而言，开展自主行动以及对行动后果的认知能力是作为责任伦理主

体的前提。一般认为,以计算机技术程序化运行的人工智能虽然具有一定的自主能力,但由于缺乏认知能力,仍不能作为责任主体承担伦理责任[35]。沃拉奇(W. Wallach)和艾伦(C. Allen)就指出,由于机器无法知晓每一个可能行动的后果,无从预测行动可能创造的各种快乐的类型和强度,因而也就不能依据这种预测进行道德抉择[36]。然而,随着生成式人工智能的出现,AI非人类行动者具备了通过大数据语料库进行预训练式自我学习的能力,并且能够在与人类行动者的转译互动中展示出某种程度上的认知能力、智识能力,将生成式人工智能视为责任主体的呼声日益高涨。比如,布鲁克斯(Rodney Brooks)就认为,机器人或早或晚都将拥有与人类对等的权利,这是人工智能实现社会化功能的必然结果[37];考克伯格(Mark Coeckelbergh)也指出,由AI生成的仇恨言论不应当归咎于技术开发人员或用户,因为很多时候并非出自用户本意,而是机器人自我学习的结果[38]。

综上可见,在生成式人工智能的行动者权责关系中,对人类行动者与非人类行动者的权利与责任进行分配日渐成为治理共识。总体而言,进入人工智能时代,AI本身应该获得更多主体性权利、承担相应社会责任的呼声日益高涨,即对具有自我学习与更新能力的非人类行动者应加强监管,包括AI技术不能无底线发展、AI算法应推进公开透明、AI技术伦理审查和预判应该在产品开发前就介入等,都逐渐成为治理共识。因此,在具体实践中,就是要坚持权责关系明确的行动者治理原则,摸清行动者转译流程,区分哪些是人类行动者责任,哪些是非人类行动者责任,开展权责明确的有效治理。

(二)加强行动者转译的全过程算法治理

算法统摄着数据收集、整理、分类、组合、生成、输出等各个环节,是各类生成式人工智能产品设计、研发、测试和应用中的核心要素。换言之,算法是最为核心的非人类行动者,贯穿着行动者转译的全过程。因此,对生成式人工智能进行监管规制,重点便应加强行动者转译的全过程算法治理,介入算法设计、算法运行、算法反馈及算法更新的全流程,重视算法伦理、规范算法权力、彰显算法正义。比如,在生成式人工智能产品研发和应用的算法设计环节,应落实算法备案制度,将履行告知义务、参数报备、参数公开和规定算法可解释权等治理手段纳入算法治理的公共政策考量之中;在算法运行环节,非人类行动者基于与人类行动者的转译交互,获取大量用户信息,应公开对话中的算法逻辑,恪守《中华人民共和国数据安全法》《中华人民共和国个人信息保护法》《关于构建数据基础制度更好发挥数据要素作用的意见》等数据信息法律法规,切实保护数据安全和用户个人隐私;在算法反馈环节,加强数据标签的伦理分级分类,引导算法正能量,对危害公序良俗内容进行伦理限制,对涉及违法犯罪信息应及时予以服务终止,并对用户进行警示性提示;在算法更新环节,注重用户体验,切实吸收用户反馈,完善算法,提升服务效能。

(三)构建人机“共善”的行动者治理共同体

在生成式人工智能的各个应用领域,人类行动者的“动机恶”与非人类行动者的“技术恶”是诸多社会伦理风险产生的直接原因。因此,如何抑制人类行动者的“动机恶”、消除非人类行动者的“技术恶”就成为提升治理效能的关键环节。实际上,在行动者转译和网络建构的过程中,始终存在着“人—机”善恶行动取向的互动互构。为此,充分发挥两类行动者“善”的引领性、互补性、共生性,构建人机“共善”的行动者治理共同体,就成为开展生成式人工智能社会伦理风险治理的必然选择。当然,尽管非人类行动者具备了与人类行动者“对话”的能力,成为与人类共同开展社会建构的技术性主体,但从历史长河来看,当前生成式人工智能虽有长足发展,却依然尚处于人工智能发展的“幼童”阶段,主要是作为一种社会实践的工具,还没有形成自主愿望、动机和意识。不可否认,非人类行动者已经能够在转译过程中,逐渐影响人类行动者的价值观念、思维模式、行动选择,但总体而言,人类行动者依然在社会行动中占据着主导地位。因此,在构建人机“共善”的行动者治理共同体时,也应以人类行动者为主导,给予非人类行动者充分的理解、包容与尊重,推进切实可行的治理实践。

在行动者治理共同体构建中，政府作为社会治理的主导性行动者，应坚持以人为本、促进发展、保障安全、造福社会的原则，牵头推动生成式人工智能战略规划，制定各类法律法规，出台各类伦理指南，对行动者“动机恶”“技术恶”进行有效规制，发挥“动机善”“技术善”的正向效能；加强对欠发达地区的人工智能投入，弥合不同地区、不同群体之间的人工智能数字鸿沟。此外，国家作为全球数字交往的主要行动者，还应积极参与全球人工智能治理体系规划和标准制定，以数字平权为核心理念，推动世界各国在人工智能领域享有平等的数字参与权、发展权、获益权、治理权，构建全球人工智能发展的“共善”格局。

此外，由于各类大模型平台能够“凭借中间人身份，通过数字化手段加速匹配双边或者多边市场需求”[39]，应该发挥平台企业“中间人”优势，引导平台联合不同行动者主体共同构建“抑恶扬善”的良好生态体系，设立生成式人工智能伦理委员会，在技术路线规划、产品设计、开发、实验、推广及应用的各个阶段，加强“共善”伦理引导与纠偏。行业协会、技术社群、公民个体等，也应与平台企业一道积极发挥自身作用，开展人工智能技能培训，提升公民人工智能数字可行能力；开展人工智能技术及应用科普宣传，引导公民对人工智能产品和服务加深理解；加强公民对“技术恶”的数字安全防范意识和应对能力，提升公民“动机善”与“技术善”的共振能力。

结语

基于行动者网络理论分析可见，在开展生成式人工智能社会伦理风险治理时，应从行动者权责关系入手，有针对性地进行主体端治理。针对人类行动者端的治理，应注重提升人本身的道德自觉、伦理责任、人工智能素养；针对非人类行动者端的治理，应注重以人为本，坚持数字包容与算法正义。在行动者转译过程中，非人类行动者尽管发挥着越来越重要的作用，甚至在特定阶段占据技术主导性，但转译的发起依然由人类行动者实施，而转译的社会行动落地也由人类行动者执行。因此，生成式人工智能治理应是“以人的发展为目标的治理”[40]，推进人的自由全面发展是衡量生成式人工智能科技向善的根本标尺。

总而言之，生成式人工智能的出现，让人们对科技赋能社会建构充满了想象和期待，但与此同时，技术创新可能带来的风险挑战也需要我们全面考察评估和审慎应对。安全是发展的前提，发展是安全的保障，不发展是最大的不安全。当前，全球人工智能领域的国际竞争博弈日趋白热化，这既是技术之争、话语权之争，也是主权之争、利益之争和安全之争。面对这种严峻复杂的形势，我们既要坚持审慎监管，恪守社会伦理底线，也要坚持数字包容，给予新一代人工智能技术发展和应用充分的成长空间，以人工智能的快速、健康、可持续发展助力全面建设社会主义现代化强国。

注释：

①目前，将“生成式人工智能”翻译为“Generative Artificial Intelligence”、简写为“GenAI”的做法较为普遍，本文认为技术在参与社会建构时才具有行动者意义，因此在文中主张“Artificial Intelligence Generated Content”译法，简写为“AIGC”。

②马克思恩格斯文集：第1卷[M].北京：人民出版社，2009：602.

③⑳管俊路、顾理平.价值冲突与治理出路：虚假信息治理中的人工智能技术研究[J].新闻大学，2022(3).

④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).

⑤李政涛.ChatGPT/生成式人工智能对基础教育之“基础”的颠覆与重置[J].华东师范大学学报(教育科学版)，2023(7).

⑥陈昌凤.生成式人工智能与新闻传播：实务赋能、理念挑战与角色重塑[J].新闻界，2023(6).

⑦郑曦.生成式人工智能在司法中的运用：前景、风险与规制[J].中国应用法学，2023(4).

⑧封锡盛.机器人不是人，是机器，但须当人看[J].科学与社会，2015(2).

⑨张玉洁.论人工智能时代的机器人权利及其风险规制[J].东方法学，2017(6).

⑩Putman H, Putnam H. Robots. Machines or artificially created life?. The Journal of Philosophy, 1964, 61(21), pp. 668-691.

⑪Coeckelbergh M. AI ethics[M]. Mit Press, 2020, pp. 52.

⑫Hauer T. Machine ethics, allostery and philosophical anti-dualism: will ai ever make ethically autonomous decisions?[J]. Society, 2020, 57(4): 425-433.

⑬何怀宏.人物、人际与人机关系——从伦理角度看人工智能[J].探索与争鸣,2018(7).

⑭刘大椿、赵俊海.科学哲学的经验主义新建构[J].中国社会科学,2016(8).

⑮王增鹏.巴黎学派的行动者网络理论解析[J].科学与社会,2012(4).

⑯刘文旋.从知识的建构到事实的建构——对布鲁诺·拉图尔"行动者网络理论"的一种考察[J].哲学研究,2017(5).

⑰[加]瑟乔·西斯蒙多著,许为民等译,科学技术学导论[M].上海:上海科技教育出版社,2007:84.

⑱Callon M. "Some Elements of a Sociology of Translation: Domestication of the Scallops and the Fishermen of St Brieuc Bay" in Law J. Power, Action and Belief: A New Sociology of Knowledge. London: Routledge, 1986, pp. 196-223.

⑲Tristan Greene. Here´s why low-credibility news seems to dominate Twitter, https://thenextweb.com/artificial- intelligence/2018/11/27/heres- why- lowcredibility- news- seems- to- dominate-twitter.

⑳Yao Y, Viswanath B, Cryan J, et al. Automated Crowdtur?ng Attacks and Defenses in Online Review Systems. Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 2017, pp. 1143-1158.

㉒马克思恩格斯文集:第3卷[M].北京:人民出版社,2009:435.

㉓John Maynard Keynes. Economic Possibilities for Our Grandchildren, in Essays in Persuasion. New York: Harcourt Brace, 1932, pp. 358-373.

㉔Mohana Ravindranath. White House: AI Could Disrupt Millions of Jobs, and Policy Needs to Adjust[EB/OL]. https://www.proquest.com/magazines/white-houseai-could-disrupt-millions-jobs-policy/docview/1851280985/se-2.

㉕AI失业潮[J].二十一世纪商业评论,2017(10).

㉖马克思恩格斯全集:第2卷[M].北京:人民出版社,1957:322.

㉗马克思恩格斯文集:第5卷[M].北京:人民出版社,2009:508.

㉘孙伟平.人工智能与人的"新异化"[J].中国社会科学,2020(12).

㉙[法]福柯著,佘碧平译.性经验史[M].上海:上海人民出版社,2000:105—106.

㉚[法]福柯著,刘北成等译.规训与惩罚[M].上海:生活·读书·新知三联书店,2010:227.

㉛[美]马克·波斯特著,范静哗译.第二媒介时代[M].南京:南京大学出版社,2000:120—121.

㉜AMY ZEGART. Open Secrets: Ukraine and the next intelligence revolution[EB/OL]. https://www.foreignaffairs.com/world/open-secrets-ukraine-intelligence-revolution-amyzegart.

㉝张玉洁.论人工智能时代的机器人权利及其风险规制[J].东方法学,2017(6).

㉞Hilary Putnam. Robots: Machines or Artificial Created Life? The Journal of Philosophy, 1964, 61(21), pp. 668-691.

㉟Hakli, R., M?kel?, P. Moral responsibility of robots and hybrid agents. The Monist, 2019, 102(2), pp. 259-275.

㊱Wallach, W., Allen C. Moral Machines: Teaching Robots Right From Wrong, Oxford: Oxford University Press. 2008, p. 87.

㊲Brooks, Rodney. Will robots rise up and demand their rights? Time, 2000, 155(25), pp. 86.

㊳Mark Coeckelbergh. AI Ethics. MA: The MIT Press, 2020, p. 7.

㊴李韬等.平台经济下的垄断与治理:新特征、新挑战、新对策[J].社会治理,2022(2).

㊵李韬、冯贺霞.数字治理的多维视角、科学内涵与基本要素[J].南京大学学报(哲学·人文科学·社会科学),2022(1).