智能社会实验:场景创新的责任鸿沟与治理

俞 鼎 李正风

【摘 要】人工智能社会实验具有场景驱动创新的显著特征,而前者需要迫切解决的人工智能应用中最困扰人类的“责任鸿沟”问题也是在具体场景中生成并得到解决的。通过分析人工智能赋能场景驱动创新的技术逻辑、场景创新对人工智能伦理规约的内在要求,可以得知,“人工智能社会实验伦理”与“人工智能伦理”之间存在“责任鸿沟”方面的共性问题,因此消解人工智能社会实验的“责任鸿沟”成了新的当务之急,这将直接助益人工智能伦理原则与法规的制定。根据实验场景的不同,实验伦理挑战的归责类型可以就性质、界限上的差异做出细分即罪责鸿沟、道德问责鸿沟、公共问责鸿沟、积极责任鸿沟,通过引入“有意义的人类控制”这一当前人工智能全球治理领域最具发展前景的指导思想及实践框架,可以为综合解决人工智能社会实验场景内就认知、具身、道德意义上的控制形式与其多元责任形式之间的脱节提供一种新思路。

【关 键 词】人工智能社会实验;场景驱动创新;责任鸿沟;有意义的人类控制

【作者简介】俞鼎(1988-),男,讲师,博士,浙江大学马克思主义学院,am_yuding999@163.com(杭州 310058),浙江大学马克思理论创新与传播研究中心(杭州 310058);李正风(1963-),男,教授,博士生导师,清华大学社会科学学院(北京 100084),中国科学院学部-清华大学科学与社会协同发展研究中心(北京 100084)。

【原文出处】《科学学研究》(京),2024.6.1121~1128

【基金项目】浙江省社科规划青年课题(24NDQN069YB)。

党的十九大以来,在全国各地部署“国家新一代人工智能创新发展试验区”成为了我国加快实施创新驱动发展战略的关键举措。人工智能社会实验是该项工作的重要组成部分,已经成为我国研判与防范人工智能潜在社会影响的支撑路径。与此同时,“场景驱动创新”(context-driven innovation)作为数字经济时代涌现的全新创新驱动范式[1],也为我国实施人工智能社会实验提供了需求牵引、政策导向与场景建设指南。不过,这也意味着,最困扰人类社会应用人工智能的“责任鸿沟”问题将在试点场域先行出现,导致人工智能社会实验的伦理治理也遭遇了归责困境。本文首先分析了人工智能社会实验中的场景创新特征,揭示出“人工智能伦理”与“人工智能社会实验伦理”之间的一阶、二阶责任并不是绝对泾渭分明的,存在核心问题上的重合;接着根据不同实验场景特点对相关责任鸿沟就性质、界限上的差异展开分类;最后尝试提出一种综合解决责任鸿沟的新思路。

1 人工智能社会实验中的场景驱动创新

数字经济时代的主导创新驱动范式,是以“场景即需求”为导向。场景创新需要人工智能赋能,而人工智能算法的迭代升级离不开新场景的学习,因此人工智能技术创新与应用场景之间存在反身性联结。人工智能社会实验需要迫切解决的“责任鸿沟”正是内生于这种反身性的联结过程。

1.1 人工智能赋能场景驱动创新的技术逻辑

人工智能是引领新一轮科技革命与产业变革的重要驱动力,也催生了由场景驱动的创新范式。那么,何谓“场景驱动创新”?通俗一点的释义是指场景需求为导向的人工智能技术产业化创新。我们知道,人类社会中“创新”概念的缘起与技术变革密不可分,但在初始阶段主要是由技术单向驱动经济场景创新。随着科技制度化、动力机制的演变以及知识经济的兴起,后学院科学引导的弱自治规范

体系纳入了市场、政治、社会等外部力量作为科技创新的驱动力，于是技术创新成为了国家创新系统的组成部分，创新驱动战略的议题由此而来。如今，人工智能技术对多样性社会需求的满足必须强调现实场景的引导。数字经济时代的底层逻辑是创造满足用户需求的场景来解决消费痛点，而新消费痛点的出现又将产生新的经济创新点。这样，场景可以简单理解为嵌入性场域与需求情景的结合：前者是被限定的时空框架，而后者是其特定内容，包括了行为、要素、关系等。人工智能驱动的一系列自主系统提供了满足数字经济时代对场景需求的迭代性、碎片化、嵌入性、虚实相生等特征的技术支撑。人工智能的深度学习以及挖掘、分析海量数据的超强算力，能够精准识别、整合与反馈行为惯性、用户体验、深层次需求等情境条件，实现对应用场景的赋能以及突破。比如，碎片化场景的商业应用能够实现企业竞争优势；赛博、元宇宙这类高阶虚实融合空间的出现，直接缔造了新的人类生存状态。

1.2 场景驱动创新对人工智能展开伦理规约的内在要求

场景迭代以及与其他场景边界的融合过程会产生无序性，源于人工智能与场景交互会促发四类显著的技术属性：(1)学习能力；(2)黑箱性质；(3)对许多利益相关者的影响(甚至那些本身不使用系统的利益相关者)；(4)自主或半自主的决策特征[2]。智能体与复杂场景的交互容易产生不可预测的自主决策，导致监管空场以及参与主体道德感的降低。但是，场景创新的底层逻辑决定了数字经济时代的产业结构变革依赖人工智能算法的迭代，而这种迭代又离不开场景。具体而言，场景创新对人工智能的影响具有反向约束性，一方面是认识论上的，在实践中缺乏明确的情境表征是导致知识型系统故障的根本原因[3]；另一方面，认知规范又会受到社会规范的建构影响，判断人工智能的道德框架总是立足于现实场景，而应用人工智能不仅引起了原先价值观念的演变，还会驱动人工智能的道德自主学习。显然，场景创新蕴含着对人工智能展开伦理规约的内在要求。从行动哲学来看，场景创新维系的反身性结构的规则依赖一种克制无序性的内在机制，这种场景类似于一种“小生境”的观点，内在于场景的规则不是初始就是先验的，只有置身这种场景并与之展开实践，个体的行动才能遵循规则。于是，场景创新对技术驱动的伦理规约乃是重塑后者应用边界的过程。比如，人工智能介入医疗领域既拓展了辅助诊断、个性化健康管理、精准公共卫生等新的应用场景，也交互产生了智能医疗数据共享标准、智能医疗器械审批标准、智能医疗人才培养体系等作为智能医疗伦理可接受性的基础条件。

1.3 责任鸿沟是人工智能场景创新的核心伦理问题

人工智能技术的不断迭代升级对场景创新的赋能具有试错性；反过来，场景创新对技术的最优伦理规约机制需要经历探索的过程，也就亟需新的容错纠错空间。缺乏监管规则的人机交互系统行为会对社会道德框架的一致性与法律中责任概念的基础构成威胁，这就产生了“责任鸿沟”问题。这些挑战在传统实验室条件下难以充分暴露，人工智能的实验阶段需要直面实验室之外的真实社会场景，就有了所谓的人工智能社会实验。我国政府对于人工智能社会实验的规划部署，并不限于潜在技术风险的研判与预防，而是前瞻性试验未来智能应用场景及其规范秩序。例如2021年12月，中央网络安全和信息化委员会印发的《“十四五”国家信息化规划》，将“人工智能社会治理实验”列为构建数字社会治理体系的重点工程，涉及医学、城市管理、养老、环境治理、教育、风险防范等7个应用场景；2022年7月以来，国家部委又相继发布了《关于加快场景创新以人工智能高水平应用促进经济高质量发展的指导意见》《关于支持建设新一代人工智能示范应用场景的通知》，明确将人工智能迭代升级与制造、物流、交通、医疗、农业等需求侧牵引的“场景创新”进行结合。可见，场景创新实则消解了实验室的内外边界。人工智能社会实验是对“场景”的试验，区别于以“自然物”为对象、完全受控为前提的“自然科学实验”，这里的“场景”是“硬设施(人工物)”与“软环境(社会规范体系)”相整合的“社会技术(AI)系统”；实验的设计者、部署者、立法者等主体行为及其影响都是内在于实验系统，与参与实验的公众、用户、旁观者一样都是“社会实验者”，不存在绝对的实验主客体。这样，“人工智能伦理”与“人工智能社会实验伦理”之间并不存在泾渭分明的一阶、二阶责任，而人机交互场景产生的“责任鸿沟”也直接构成了人工智能社会实验的伦理挑战。

2 人工智能社会实验的四类责任鸿沟挑战

实验总是关乎伦理。受制于机械自然观、事实

与价值二分的传统认知立场,科学实验室与社会划界契约下的实验伦理与技术伦理之间缺少互惠互鉴:技术伦理源于社会性,而实验伦理针对的是研究主体。在场景驱动创新背景下,人工智能社会实验要实现对“干预行为”(即人工智能引入社会)及其“干预后果”(潜在风险、道德困境)的双重控制,这是一个“先干预后控制”的过程。简言之,人工智能社会实验具有“非完全受控”的根本实践特征。

这样,人工智能社会实验场域下的实验伦理与技术伦理就有了焦点问题与规范要求上的重合,一方面它们涉及共性问题,比如隐私与数据保护、公平性、不透明、缺少人文关怀;另一方面,它们也可以共享一些伦理原则像以人为本、可解释性、知情同意、不伤害。当然,本文的目的不是要彻底消解这种边界,也不存在这种情形,因为作为一项现场实验始终需要施展实证科学方法论,这就存在固有的伦理问题域,比如“脏手”欺骗[4],还有参与者角色客体化、创伤后遗症、胁迫性干预等,而人工智能也存在彰显技术伦理特质的父爱主义、成瘾、过度依赖等问题。本文侧重的是“人工智能社会实验伦理”与“人工智能伦理”的重合点,在这里后者能够直接共享前者的一部分伦理规范,因为这种实验会在后验实践中生成新的价值观与规约机制,既对自身的前置实验规则进行迭代,也同时实现了对实验对象的规约。

责任鸿沟作为场景驱动创新情境下人机交互系统的核心伦理问题,不是单一、独立的,而是多个相关联问题的集合。传统意义上,责任鸿沟源于责任概念的缺失。溯源亚里士多德对责任条件的界定,无知与强迫是豁免道德责任的基本理由。随着现代性困境以及科技泛化带给主体性、公共性的影响,追责标准被不断细化,可大体区分为行动自由、可及性、可预见性、不当行为、因果性五个条件,这里的认知与行为规范又是互为前提的。

人工智能在数字经济时代的应用场景中被赋予了作为委托决策代理人的自主权,这被关联了现实的人类行动与社会因素(法律、机构、惯例等),而对这种自主权的问责需要探索一套新的法律、道德与社会上可接受的人类控制形式,前提是要厘清责任鸿沟的多样性。在此种背景下,荷兰代尔夫特理工大学的德西奥(Filippo Santoni de Sio)、梅卡奇(Giulio Mecacci)整合道德哲学、法哲学、社会学、技术伦理等多学科语境下对责任概念的界定,将人工智能的责任鸿沟问题区分为为四种类型:罪责鸿沟(culpability gap)、道德问责鸿沟(moral accountability gap)、公共问责鸿沟(public accountability gap)与积极责任鸿沟(active responsibility gap)。人工智能社会实验在实践层面兑现了科学实验哲学的期许,相比实验室实验更直接地与社会或伦理规范联系了起来。鉴于伦理二阶性的局部消解,人工智能社会实验同样需要正视多元的责任鸿沟挑战(表1[①])。

2.1 罪责鸿沟

罪责鸿沟与实验的“安全场景”紧密相关。罪责或应受责备的基本定义是对做出不具有合法理由的不当行为的主体,进行谴责、制裁或惩罚。罪责归因主要有三方面的重要性:第一是提高司法威慑力,控制和减少不当行为;第二是强化集体对共享规范的遵循,遏制不正当价值观念的扩散;第三是补偿作用,公开的罪责归因能够保障受害者获得

表1 人工智能社会实验的四类责任鸿沟挑战

责任类型	界定	实验鸿沟
罪责	基于意图、知识或控制的不当行为应受谴责	实验“安全场景”中的“多手问题”“多物问题”,容易导致实验风险或事故的罪责归因困境
道德问责	人类在某些情形下有义务向他人阐述自己的行动和理由	实验“黑箱场景”中的专家共同体无法向其他参与者充分解释设计与部署实验方案的意图与理由,导致道德感降低
公共问责	公共代理人有义务向公共论坛解释其行为	实验“决策场景”中人工智能企业或其他委托单位诸如司法、政府、医院等算法决策系统的试验囿于商业机密、行政保密管理,无法公开审查
积极责任	有义务促进和实现某些社会共享的目标和价值观	实验“公益场景”中参与者过度关注实现人工智能经济效益或工具价值的实验结果,缺乏抑制无度追求非公益性的绩效而可能造成风险的负责任意识或能力

象征或物质意义上的补偿[5]。人工智能社会实验的罪责鸿沟主要源于自动驾驶汽车、医疗人工智能、无人机等与实验者以及未来用户的生命安全有关的“安全场景”。赋能安全场景的人工智能在社会中应用的潜在罪责归因可以追溯社会实验的设计与部署阶段,这里的罪责鸿沟又主要源于“多手问题”(the problem of many hands)与更复杂的“多物问题”(problem of many things)。人工智能赋能安全场景属于复杂的人机交互工程,也是人类与非人类行动者并重的集体实验。比如,人工智能医疗社会实验的控制组和对照组涉及了不同主体,例如医疗数据采集与分析研究员、医学伦理审查委员会、智能仪器生产商与审批机构以及医学人工智能培训组织等,这就遭遇了遵循不同规范标准的多(人类)手,以及不同智能医疗系统(传感器、驱动器、处理器以及辅助、诊断、手术机器人)之间的交互影响,倘若出现实验事故就难以实现罪责归因。

2.2 道德问责鸿沟

道德问责鸿沟与实验的“黑箱场景”紧密相关。道德问责的形式较为宽泛,应受责备是其中的严厉形式。关于道德责任的哲学文献中,道德问责被认为是证明和理解道德责任实践的一个重要方面[5]。为将道德问责与其他责任形式区分开来,德西奥等人将法哲学家加德纳(John Gardner)提出的“基本责任”概念(即个体对其自身负责的能力)进行了拓展。个体基于理性能力有对事物的认知做出阐释的义务,在当代复杂社会关系中个体所具备的反思性更多是要向他者证明或辩护自身的责任,这不像罪责是极力去避免的。那么,人工智能社会实验场域中的道德问责鸿沟主要表现为认知美德的缺乏。参与实验的专家共同体面对“黑箱场景”有向他者的期望进行理性响应的义务。比如,对智能医疗诊断系统的不透明指标进行边界测试,但智能系统学习、人机交互产生的“黑箱性质”导致实验的程序员、设计师、工程师需要各自就专长与道德判断对不确定性实验方案做出决策,但在这种决策的理由与逻辑又囿于技术的不可解释性难以进行知识表征性的诠释情形下,并没有如实履行对其余实验参与者期望性的解释义务。

2.3 公共问责鸿沟

公共问责鸿沟与实验的“决策场景”紧密相关。应用人工智能的“决策场景”涉及了教育、卫生、司法等事关社会公平正义的算法决策系统(algorithmic decision support systems),在享有行政与商业保密权的同时,对嵌入智能算法设计的理由以及自主决策与人为干预之间的划界意图理应提供公共问责的空间。公共问责的基本定义是对公权力的监督与制约的一项制度安排。场景创新驱动的人工智能社会实验是一个公共组织,相关主导机构(制造商、研发主体、监管审批部门、伦理委员会等)与公共论坛之间有着问责关系:一方面实验涉及成本,不可能无节制地反复试错,有义务接受纳税人质询;另一方面,实验伦理很早就成为了一项公共议题。按照大多数社会科学实验的惯例是单盲试验,很少存在双盲,因为实验设计者要清楚学科研究的指定条件,才可能在不同实验场景下采取不同的行动,因此,主导人工智能社会实验“决策场景”的部门是有明确理由,并有责任向公共论坛回应,但是外界仍然忽略了提高这类实验透明度的要求。比如,人工智能教育社会实验的中观场景是要重塑未来学校形态,涉及到的学分管理机制、在线学习的可信度与有效性、数据隐私等问题局部反应了主导实验的决策设计[6],但往往忽略了要向家庭、学校、社会三个主线场景的参与者、旁观者回应部署实验的动机与理由,反而会降教育实验的社会效益。

2.4 积极责任鸿沟

积极问责是一项前瞻性责任,区别以上三项消极责任形式,这与实验的“公益场景”紧密相关。积极责任在工程伦理领域已经界定为工程师应遵循的义务与美德,例如主动促进公共价值的实现,并尽量预防与避免不当后果。人工智能企业在当下有义务主动通过技术调解来促进社会向善,比如算法工程师利用用户画像来捕获可能具有自杀倾向的用户,并及时反馈信息,因此人工智能赋能“公益场景”的社会实验涉及了“社会实验者”的积极责任。不过,常见的积极责任鸿沟有二:一是不清楚该遵循何种义务,二是清楚该遵循何种义务却没有能力执行。前文已经提到,人工智能“社会实验者”都是内在于实验系统,组成专家共同体的设计者、程序员、工程师并不能充分诠释人工智能驱动的系统行为逻辑;与此同时,社会科学研究中常出现一种“实验者效应”(experimenter effect),即实验者的内心期望会不经意间通过表情、语气、动作传递给受试者,导致后者行为无法保证不偏不倚。场景创新驱动下的人工智能社会实验很多时候是围绕商业性智能系统展开的场景实验。但是为了经济效益考量,实验专家组可能有意夸大实验结果或下意

识的强化“实验者效应”;另外,同样作为社会实验者的用户、公众与旁观者更多时候囿于受试者心态,没有意识到自身作为实验者有促进公共善的义务。

3 解决“责任鸿沟”的一种新思路:“有意义的人类控制”的视角

社会技术(AI)系统的规范要求是人类主体理应作为自主系统关键行为的最终决策者并承担道德与法律责任。鉴于,社会实验者对智能机器自主行为的控制是全局性的而非直接操作意义上的,需要探索不同的控制形式与责任形式之间新的监督机制。在这里,“有意义的人类控制”(以下简称MHC)这一近年来国际社会积极倡导并不断完善的人工智能伦理治理理念及其设计框架可以为综合解决人工智能社会实验的责任鸿沟提供新思路。

3.1 “有意义的人类控制”的概念缘起、理论基础与实践框架

MHC这一概念的源起同样是要解决人工智能的责任鸿沟挑战,核心理念是要求人类保证对智能机器的最终控制并承担道德责任,并试图倡导当代新兴技术的社会控制哲学。这一概念最早兴起于一个相对特殊却是对人类社会最具威胁性的人工智能应用领域——“致命性自主武器系统”。2012年,美国国防部颁布了3000.09指令提出自主武器系统的全生命周期都应在有意义的人类控制之下;2014年,联合国《特定常规武器公约》正式将智能军事武器应遵循有意义的人类控制的伦理议题纳入框架议程,随后逐渐成为国际社会探讨人工智能武器的伦理、法律与公共政策框架的概念核心。不过,军事领域对MHC的概念阐述有很大局限,主要强调人类操作员与智能武器之间的直接因果责任,问题是与人类社会普遍关联的智能系统并非来自军事领域,而是交通、医疗、家具、教育、保险等公共生活系统,在愈发普遍的同时也变得更具自主性。这种“自主性”也代表了一种自主权,却在当前的社会规范层面缺乏监督机制,根据贝基(George Bekey)对自主性概念的界定“一旦机器被激活,至少在某些操作领域,能够在真实世界的环境中不受任何形式的外部控制,长时间运行的能力”[7]。这里对机器控制的界定就不再仅限于直接操作意义上的。

德西奥与范登霍温(Jeroen van den Hoven)对奠定MHC的哲学基础做出了开创性贡献。他们基于美国伦理学家费舍(John Fischer)、拉维扎(Mark Ravizza)提供的当代道德哲学相容论版本——对指导控制(guidance control)、自由意志、道德责任之间的关系进行了哲学阐发,以及价值敏感性设计的理念,对构成行动机制道德责任的基础条件进行了修正,提出满足MHC的智能系统应遵循“追踪—追溯”双重伦理设计原则[8]:(1)“追踪”原则的目的是要保证人类主体与智能系统之间的实质性关系,即人机系统的行为能够对现实场景中的人类行动理由(意图、计划或是道德理性)做出反应;(2)“追溯”原则提供了追责的时间阈值,即人机系统的行为、能力及潜在社会影响应能溯及至少一个系统设计或与系统交互的相关人类主体的适当的技术与道德理解。

为实现抽象的伦理设计原则转向现实的可操作路径,这项工作吸引了哲学、交通工程学、心理学、计算机、人类学、设计学等一众跨学科专家的努力,致力于建构一套综合解决人工智能责任鸿沟的方法论,代表性的贡献有:针对追踪原则的操作属性要求,即可量化代理人、行动理由对系统行为产生影响的程度,梅卡奇(Giulio Mecacci)等人基于行动哲学、行为心理学和交通工程学的研究成果,提出了一种可联结、可解释人类行动理由与系统行为之关系性的“理由接近度量表”(proximal scale of reasons)[9];根据追溯原则的实践标准是要能衡量人类在多大程度上能够在技术专长与道德意识层面理解智能机器的设计、部署与操作,卡尔弗特(Simeon Calvert)开发了一种“评估级联表”(evaluation cascade table),可为事关知识、能力和意图的纯粹定性评估提供实证基础[10];还有一些综合双重原则的研究,比如,希伯特(Luciano Siebert)等人基于溯因思维的迭代过程将双重抽象性原则转化为四项可操作属性[2];佛迪森(Ilse Verdiesen)等人根据控制自主系统的多层维度与时间序列构建了一套(水平层和垂直列相互连接并在信息上相互依赖)“全方位人类监督框架”(A Framework for Comprehensive Human Oversight)[11]等。

3.2 “追踪—追溯”双重伦理设计原则对实验伦理责任的促进

可以看到,实现MHC的智能系统就是一个社会技术(AI)系统,这与人工智能社会实验的实施场景与对象是相一致的。另外,人工智能社会实验道德问责之外的责任鸿沟类型都可能与人工智能这一技术变量引发的“黑箱场景”相交集,对于不完全可控的技术变量,存在指导实验的理论前提或概念

框架的缺失,显然,人工智能社会实验没有传统实验的典型约束,在实践层面具有显著的探索性质,通过不断调整实验参数,以确定那些不可缺少的实验条件,然后寻找稳定的经验规则。同样,“追踪—追溯”双重条件的满足不是前提而是后验实现的,目的也是要解决人工智能的责任鸿沟。那么,根据MHC的规范性要求展开人工智能社会实验责任鸿沟的综合治理,需要遵循“追踪—追溯”双重原则来促进多元责任形式的效力(表2[②])。

根据追踪原则,人机系统的行为要与人类主体的行动理由保持协同变化,这里的关键就是量化何种理由最大程度影响了人机系统。在国外一些自动驾驶汽车的事故案例中,由于自主系统的故障却将罪责判给了乘车员,而实质上是汽车厂家的商业意图与所在州政府的招商政策导致不成熟技术的直接民用,但没有向公众澄清,这就同时涉及“安全场景“与“决策场景”以及罪责、公共问责鸿沟。因此,实验阶段的智能系统伦理设计有必要引入梅卡奇等人开发的“理由接近度量表”。这是基于人类行动理由与系统行为之间有远端、近端之分:近端理由是以一种临时即时的方式连接一个动作的意图;远端原因是指以一种不那么直接的方式制定长期规划的目的。比如,组织与设计自动驾驶汽车实验方案的群体与体验这辆汽车的用户就有不一样的价值偏见。那么,根据追踪原则对指导实验的伦理规范,首先是对社会实验者与实验理由进行各自分类。比如,实验者包括了制造商、政策法律、设计师与用户、旁观者等;实验理由包括价值、规划、意图等;其次是揭示系统的先验价值以及潜在的价值冲突,比如是否存在远端理由的集体偏好优先个体近端偏好的现象,这样远端理由可以直接取代近端理由;最后就是对实验理由进行排序,确定最大权重。

基于追溯原则,人机系统的行为要与人类主体的能力相一致。不可否认,人类最早获取应用智能系统的规则、知识与技能主要就是在实验阶段,像自动驾驶汽车的教学体系与交通法规都属于实验结果。在实验过程中,实验者就会陆续将决策任务移交给智能系统,但不可能全部移交,这里的边界就是测试对象,尤其其中涉及的人为元素,可进一步划分为技术理解与道德意识,因为掌握应用智能技术的各种知识并具备道德意识的代理人符合作为归责主体的条件。但是这些能力具有高度抽象性,卡尔弗特开发的“评估级联表”提供了量化这些抽象要素的策略,这是对心理学实验中最常用来测度感知量的李克特量表(Likert Scale)的综合应用。倘若在实验阶段要明确人类对智能系统的控制及

表2 “有意义的人类控制”对人工智能社会实验伦理责任的促进

有意义的人类控制		
双重伦理设计原则	操作标准	四类实验责任形式的效力
追踪原则:社会技术(AI)系统行为与人类行动理由相一致	绘制远近端理由的排序图谱,对所有参与实验的行动者及其理由进行分类,并量化系统行为与何种人类行动理由(意图、计划或道德理性)之间的协变程度最为紧密	罪责与问责:前瞻性与后置性责任的归因要溯及实验场景的软系统部分,包括法律、制度、监管与经济政策等 积极责任:廓清社会实验者的角色责任归属,参与实验的专家、制造商、政策制定者、用户与公众等要坦诚各种的价值偏见与理由,这是展开利益冲突协调的前提
追溯原则:社会技术(AI)系统行为与人类能力相一致	测度社会技术(AI)系统中实验者各项能力(技术理解、道德意识、心理承受等)的权重,并验证系统是否与它们匹配	罪责:对实验风险或事故的责任归因,要溯及直接实验行动之前的实验设计与部署、互动阶段; 道德问责:实验参与者必须明晰在实验过程中参与系统设计或互动的功能权重,还要求明确自身的行动(意图、计划、价值判断)理由; 公共问责:要求组织与监督人工智能社会实验的机构组织能够全程参与并熟悉实验流程,并履行对实验系统之外的公共论坛进行回应的职责; 积极责任:培训实验参与者理解、掌握与智能机器交互应有的知识、规则、技能和道德意识,尤其培育公众、用户等群体参与实验的责任感

其责任问题,首先要求实验者能够感知控制智能系统的不同形式,这可大致区分为战略(政策法规与市场规划)、战术(使用规则)与直接操作;其次是测度实验者对系统行为的潜在道德后果的感知程度以及控制系统是基于何种知识类型——显性知识(know-that)、默会知识(know-how)的感知程度;最后,每个实验者按照6分李克特量表得到的总分进行排序,以反映他们对实验系统的参与程度。

需要补充的是,MHC当前的理论基础尚缺乏监督机制,相应的公益场景中有必要培育用户、公众、旁观者等作为"社会实验者"的责任感。以往的实验伦理主要是为了约束实验主体,如今,人工智能社会实验与公众实践、利益紧密关联了起来。与此同时,即便系统设计实现了MHC,并不代表智能系统的应用完全符合社会期望,例如满足追踪原则的个体行动理由可能是基于价值偏见,但系统行为又能够充分反应,这就要求监督机制的存在。一方面,这些群体首要是拥有权利意识,避免促发"实验者效应",在智能实验场域中用户群体产生的真实行动数据对于提升实验的外部有效性至关重要,因此他们有义务抵御实验专家组过度的功利主义倾向;另一方面,明确这些群体与智能系统交互的实验伦理原则,可以借鉴人机兼容与AI控制问题专家拉塞尔(Stuart Russell)提出的"拉塞尔三原则"[12]:(1)机器的唯一目标是最大限度地满足人类的偏好;(2)机器初始不确定这些偏好是什么;(3)关于人类偏好的最终信息来源是人类行为。任何实验者都可能对实验的不当后果承担道德与法律责任,因此人机交互原则始终是以人为本。

总体而言,"有意义的人类控制"是比"负责任创新"更为严苛的人工智能伦理原则,要求人类、机构(而非计算机及其算法)必须对高阶智能系统的关键决策进行控制并承担道德责任,也由此人工智能社会实验的伦理治理具备了价值导向、指导框架与高标准。不过,MHC尚且存在一些显著缺陷:比如,德西奥等人为MHC奠基的道德哲学基础部分源于"法兰克福案例"为相容论提供的论证,但"法兰克福案例"本身并不是完全自洽的;另外,追溯条件溯及的个体与追踪条件控制链上的个体可能不是同一人。显然,这些缺陷提供了拓展MHC哲学基础与实践框架的空间,这也为进一步消解人工智能社会实验的责任鸿沟铺平了道路。

注释:

①遵循"人工智能伦理"与"人工智能社会实验伦理"之间二阶性局部消解的原则,笔者根据德西奥(Filippo Santoni de Sio)、梅卡奇(Giulio Mecacci)在人工智能领域所做的责任鸿沟分类以及概念界定,与不同的人工智能社会实验场景进行了结合,进而揭示了不同实验鸿沟的具体内容。

②笔者根据德西奥(Filippo Santoni de Sio)与梅卡奇(Giulio Mecacci)对"有意义的人类控制"促进不同责任效力的分类,结合人工智能社会实验的场景特点进行了重塑。

参考文献:

[1]尹西明,苏雅欣,陈劲,等.场景驱动的创新:内涵特征、理论逻辑与实践进路[J].科技进步与对策,2022,39(15):1-10.

[2]Siebert L, Lupetti L, Aizenberg E, et al. Meaningful human control: Actionable properties for AI system development [J]. AI and Ethics,2022:1-15.

[3]Brézillon P, Pomerol C. User acceptance of interactive systems: Lessons from knowledge-based and decision support systems[J]. Failure and Lessons Learned in Information Technology Management,1997,1(1):67-75.

[4]于雪,李伦.人工智能社会实验的伦理关切[J].科学学研究,2023,41(4):577-585.

[5]Sio F S D, Mecacci G. Four responsibility gaps with artificial intelligence: Why they matter and how to address them[J]. Philosophy & Technology,2021,34(4):1057-1084.

[6]郝祥军,顾小清,王欣苗.缓和技术与教育的融合争议:教育中的技术社会实验[J].现代远距离教育,2022(4):42-50.

[7]Bekey A. Current trends in robotics: Technology and ethics[A]. Lin P, Abney K, Bekey G. Robot Ethics: The Ethical and Social Implications of Robotics [C]. Cambridge, MA: MIT Press,2014:17-34.

[8]Santoni de Sio F, Van den Hoven J. Meaningful human control over autonomous systems: A philosophical account [J]. Frontiers in Robotics and AI,2018,5(15):1-15.

[9]Mecacci G, Santoni de Sio F. Meaningful human control as reason-responsiveness: The case of dual-mode vehicles [J]. Ethics and Information Technology,2020,22(2):103-115.

[10]Calvert C, Arem V, Heikoop D, et al. Gaps in the control of automated vehicles on roads[J]. Institute of Electrical and Electronics Engineers (IEEE),2021(4):146-153.

[11]Verdiesen I, Santoni de Sio F, Dignum V. Accountability and control over autonomous weapon systems: A framework for comprehensive human oversight[J]. Minds and Machines,2021,31(1):137-163.

[12]Russell S. Human Compatible: Artificial Intelligence and the Problem of Control[M]. United States: Viking,2019:173.