

人工智能社会实验:伦理规范与运行机制

瞿晶晶¹,王迎春¹,赵延东²

(1. 上海人工智能实验室,上海 200000;2. 中国人民大学北京社会建设研究基地,北京 100872)

摘要:人工智能正推动社会进入智能化时代。社会实验被认为是系统性认识人工智能社会影响的重要途径,但其本身又叠加了多重风险,急需科学研判并形成共识性实验规范。本文从技术、社会以及二者双向互动的角度分析了人工智能社会实验的潜在风险,并结合生命医学、心理学、社会科学与人类学的代表性实验及其伦理规范,从实验对象权利保护、实验因素制约、实验过程可控性以及实验效益伦理性等维度提出 9 条实验规范。在此基础上,尝试性提炼形成了涉及主体参与、运行反馈和程序公正等层面的人工智能社会实验运行机制,以期助力实现人工智能稳健发展。

关键词:社会实验;人工智能;实验规范;技术治理

中图分类号:G312

文献标识码:A

文章编号:1005-0566(2022)11-0074-09

Artificial intelligence social experiment: Ethical norms and operating mechanism

QU Jingjing¹,WANG Yingchun¹,ZHAO Yandong²

(1. *AI governance research center, Shanghai AI lab,Shanghai 200000, China;*

2. *Beijing Social Construction Research Institute, Renmin University of China, Beijing 100872, China)*

Abstract: Artificial intelligence is pushing society into the era of intelligence. Social experiment is considered to be an important way to systematically understand the social impact of artificial intelligence, but it also superimposes multiple potential risks. It urgently needs research on the formation of consensus experimental norms. Thus, the present paper first analyzes the potential risks of artificial intelligence social experiments from the perspective of technology, society, and their interactions. Then, by combining representative experiments and ethical norms in life medicine, psychology, social science, and anthropology fields, we put forward 9 experimental specifications including the norms of protection of object rights, the conditionality of experimental factors, the controllability of the experimental process, and the ethics of experimental benefits. Finally, we try to raise an artificial intelligence social experiment operating mechanism that emphasized on the diverse subject participation, operational feedback, and procedural justice that aiming to support a healthy development of artificial intelligence technology in China.

Key words: social experiment; artificial intelligence; experimental norms; operation mechanism

随着人工智能技术的迅猛发展,人类正进入智能化社会。综合研判人工智能的社会影响,有效控制社会风险,提高治理水平,对于人工智能稳健发展极为重要。我国在人工智能领域具有数据

收稿日期:2021-11-28 修回日期:2022-06-24

基金项目:科技创新 2030“新一代人工智能”重大项目(2020AAA0105400);上海市科技创新行动计划软科学重点项目(21692102600);中国人民大学“双一流”跨学科重大创新规划平台“数字社会治理跨学科交叉平台”(297222109209)。

作者简介:瞿晶晶(1990—),女,江西南昌人,博士,上海人工智能实验室治理研究中心副研究员,研究方向为技术治理与科技创新。通信作者:王迎春。

资源丰富、应用场景广阔的先发优势,实践已经走在了理论指导的前面。在此背景下,我国不但要加快人工智能技术的应用,更要积极主动地关注人工智能技术给社会 and 个体带来的影响。目前,相关议题已有不少讨论,但尚未形成对人工智能社会影响的系统性认识,总体上还缺乏扎实的科学实证研究为科技政策循证提供支撑。通过组织开展长周期、跨学科的社会实验,可以客观记录、科学评估人工智能技术对个人和组织的综合影响,持续积累数据和实践经验,总结人工智能与经济社会协调发展的规律和趋势,为智能时代的政府治理科学循证提供支撑^[1]。

2019年9月5日,科技部发布《国家新一代人工智能创新发展试验区建设工作指引》强调:“人工智能社会实验是探索智能时代政府治理的新方法、新手段。”目前,全国77个国家智能社会治理实验基地和16个国家新一代人工智能创新发展试验区正在积极开展社会实验。2022年3月22日,我国发布《关于加强科技伦理治理的意见》进一步强调要完善科技创新体系,提升科技伦理治理能力。人工智能社会实验工作正是促进我国人工智能健康发展的重大任务布局^[2]。

然而,由于人工智能本身具有高度不确定性和较大伦理风险,人工智能以社会实验的新形式进入社会各类场景,对社会公众产生的影响可能长期持续并且无法有效调控,因此人工智能社会实验具有技术与实验系统的双重风险,实验的风险规避更具有挑战,需要更加重视实验的伦理规范,注重社会实验的风险控制运行机制^[1]。鉴于此,本文通过分析人工智能技术与社会实验的叠加风险,综合生命医学、心理学、社会科学与人类学、技术社会实验领域代表性实验或研究规范进行讨论,提出人工智能社会实验的实验因素制约性规范、实验对象权利保护性规范、实验过程可控性规范以及实验效益伦理性规范等4个方面9条实验规范,并基于全周期视角对人工智能社会实验的运行机制进行建构,旨在推动人工智能社会实验负责任的开展。

一、社会实验内涵:广义与狭义

(一)广义社会实验:一种哲学视角

广义的“社会实验”指的是“技术社会实验”这

一技术治理理论进路,其核心观点将技术本身发展以及社会应用的交互方式类比实验研究过程,认为社会是新兴技术创新发展的第二实验室:在现代技术动态演化过程中科技创新的治理机制已然面临“科林格里奇困境”,只有当技术真实地应用于社会之后才能有效测量其风险性与正确性,社会成为验证技术系统有效性的实验室。新兴技术的实验特性受到欧洲学界的广泛关注,学者们用“现实实验”“隐性实验”“集体实验”等概括技术的社会实验现象^[3,8]。Ibo Van de Poel^[4-5]基于已有研究将“以实验性思想将新技术引入社会这一动态过程”正式概念化为“技术社会实验”,并总结出技术社会实验的3个特性:第一,社会实验发生在开放的社会环境之中,实验对象广泛且不确定;第二,社会实验没有明确的实验设计人员,具有不可控、不可逆、不可预期等显著特征;第三,技术社会实验是动态的反身性过程,不仅表现在技术本身处于迭代发展的动态过程,而且表现在技术与社会交互方式也始终处于不断学习和发展过程中^[4-8]。

(二)狭义社会实验:一类科学方法

狭义的社会实验所指范畴为科学研究的方法论。实验方法在科学研究中的应用可以追溯至文艺复兴时代^[9]。通过在实验室里进行实验,自然科学家观测得到数据对科学假说进行检验,从而形成新的科学知识^[10]。随着西方社会快速的工业化和城市化进程,复杂的现实世界对因果机制的识别提出了更高的要求,而传统的观察性方法很难有效地控制干扰变量。因此,实验方法开始进入社会科学领域,并作为一种新型的政策研究和评估方法开始兴起,在社会心理研究、政策效果评价和技术的社会影响评估等方面获得了广泛的应用,日渐成为支撑决策的重要工具^[11-13]。一些著名案例包括“霍桑实验”^[14]、“联邦实验住房补贴计划项目”^[15]等,在联合国、世界卫生组织的扶贫项目中也多次采用社会实验方法支撑政策。

(三)人工智能社会实验:“实验中的实验”

人工智能是一门新兴的技术科学,已经构建了理论、方法、技术及应用系统,其主要旨趣在于研究、开发用于模拟、延伸和扩展人的智能。从广

义上看,人工智能进入社会的实验特性显著。主要体现在:一是人工智能技术在不断迭代发展。《2019 人工智能技术成熟度曲线》表明,人工智能发展刚进入普及应用阶段,尚处于从“不能用”到“可以用”的技术拐点,大部分人工智能技术还未进入成熟期,其安全性、可靠性等诸多方面还存在不足,技术本身在更新迭代之中。二是人工智能技术与开放的社会环境不断相互作用,还未形成成熟的应用范式^[1,16]。新的技术应用场景不断涌现的同时,人们对于人工智能的认知、接受程度和方式也在不断变化,因此人工智能社会交互也处于实验性过程中。

面对人工智能技术快速发展,科技创新治理需要相应变革,其关键在于对整个阶段社会综合影响进行全程跟踪监测、实时评估,形成敏捷且具有前瞻性的治理机制。苏竣等^[1]提出通过开展“人工智能社会实验”工作达成这一目标,强调“要利用政府和市场力量推动特定人工智能技术在社会治理中广泛应用,并通过采用设置实验组、对照组的实验方法对人工智能技术社会综合影响进行前瞻研判”以支撑政策的适应调整。由此可见,人工智能社会实验是在技术进入社会的实验性过程中开展技术治理探索的实验,可以被称为“实验中的实验”。

二、人工智能社会实验的双重风险

正因为人工智能社会实验是一种“实验中的实验”,其风险也具有双重性:一是技术本身的风险,二是实验作为方法体系产生的系统风险。

(一)人工智能的技术风险

(1)人工智能技术内生性风险。从技术要素角度:一是与数据相关的技术风险。人工智能数据具有“量大面广”的特点,算法训练依赖海量数据,数据涉及面非常广泛,同时数据采集则可能依靠多头采集,数据传输路径复杂等情况,增加了隐私及重要数据泄露的潜在风险^[17]。二是与算法相关的技术风险,即由于算法开发不合理产生的偏差。如专家主观性偏差融入算法设计导致模型潜存风险,以及深度学习模型存在逻辑难以解释的“黑箱”,均可能产生不确定性社会风险。值得注意的是上述数据、算法偏差往往是交互存在的,更

深化了人工智能社会风险的复杂性,增加了人工智能风险分析难度和应对措施的设计难度^[19]。

(2)技术与社会相互作用产生的外部性风险。人工智能技术外部性风险包括:①结构性风险,即人工智能作为新兴技术系统进入社会后重构传统基于人际互动而形成的组织与社会关系,对人类社会的运转模式、行为轨迹、社会网络、心理动态造成的综合影响^[1],对社会固有结构产生冲击,重构现有的社会经济秩序、弱化法律的调整功能、增加社会管理主体的治理成本等^[20-21]。②人工智能技术的非法利用。例如人工智能技术既可提升网络攻击的智能化水平,也具备深度伪造能力,不法者通过利用先进技术开展不法活动^[17]。

(二)社会实验的系统风险

实验方法既然用于对未知和不确定性科学知识进行探索、尝试,则本身作为方法系统存在一定的风险。从历史经验来看,以“人”为实验对象的实验风险主要有:①实验对象的生命安全风险。例如医学临床实验中,新的药物、手术方法、治疗方式在实验探索过程中均潜存对实验对象的生命安全风险。②实验的道德伦理性风险。在一些心理实验和社会实验中,心理学家、社会学家为了挖掘人类心理活动或者人类行为特质,采用一些极端刺激或者营造特殊环境,可能对受试者心理造成短期甚至长期不良影响。同时,如个人信息泄露或者实验结果的不妥当的曝光可能给实验对象以及相关利益人日常生活带来不利、不便,产生道德层面的风险,甚至造成严重的社会影响。③实验过程的失控风险。在实验实施时,由于实验设计、实验环境或者不确定性偶发因素导致实验过程与预期路径产生偏差,实验因此失败并产生不良后果。④实验效益的无用性风险。一些政策类的大型社会实验在耗费大量财物、人力、时间成本后,无法达到预期实验效果。如芬兰的“基本收入实验”由于在初期实验设计中对实验对象的选择过于单一,仅聚焦在长期失业者群体,实验对象地域分布过于分散,大量的低收入群体未被纳入,从而无法得出基本收入刺激的普适作用以及牵引刺激形成规模效应。同时由于税收政策的不匹配以及复杂的政治因素,在投入 2 270 万美元,持续两

年后,于2019年宣布实验终止^[22]。

值得注意的是,人工智能社会实验中的技术风险与实验的系统风险还会形成叠加效应,使得实验风险控制挑战性更大。如2012年,Facebook为研究算法对用户行为的影响,秘密地对68万名用户进行社会实验,涉及300万条个人状态,严重侵犯了用户的隐私和知情权利,直到研究团队将成果发表后,该实验才被披露^[23]。由此可见,人工智能技术的隐蔽性和扩大性造成实验系统风险的影响更为广泛和深刻^[24],大大增加实验风险应对的难度。

三、人工智能社会实验规范与运行机制

实验规范能够有效指导实验的设计和过程,有效控制风险且符合伦理道德。面向“人类”实验的规范起源于医学领域的人体实验,随着科学技术的发展,其它领域的科研实验也出台了相应的规范。梳理已有的面向“人类”实验或者科学研究的规范,挖掘内容普遍性意义,对人工智能社会实验规范的制定具有参考价值。本文类比自然科学实验组成四要素(实验对象、实验要素、实验过程、实验效益),基于生命医学研究领域世界医学组织和国际医学委员会发布的《人体生物医学研究国际道德指南》、国家药品监督管理局会同国家卫生健康委员会组织修订的《药物临床试验质量管理规范》、心理学研究领域中国心理学会于2007年发布的《中国心理学会临床与咨询心理学工作伦理守则》、社会科学与人类学领域欧盟颁布的《社会科学与人类学伦理指南》以及技术社会实验领域Ibo Van de Poel等^[6]的《技术社会实验伦理框架》等现有伦理规范文献,从实验对象的保护性规范、实验因素的制约性规范、实验过程的可控性规范以及实验效益的伦理性规范进行归纳,结合人工智能社会实验的特点,尝试提出人工智能社会实验的实验规范(见表1)。

(一) 实验对象权利保护性规范

(1) 保护人工智能社会实验对象的隐私权利

隐私保护在《人体生物医学研究国际道德指南》《中国心理学会临床与咨询心理学工作伦理守则》《药物临床试验质量管理规范》《社会科学与人类学伦理指南》《技术社会实验伦理框架》里均被作为首要原则,是面向“人类”的科学研究基本准

则(A2、B2、C2、D2、E1)。在新科研范式兴起的背景下,大数据的隐蔽性、传播性、扩大性对隐私保护提出了新的挑战。因此,人工智能社会实验更需要充分保障实验对象的隐私权利:一是坚持“最小采集原则”,采集在满足研究需求前提(例如随机性与代表性)的最小样本量,同时避免采集研究不必要的敏感信息(如身份证、轨迹、生物特征等)。二是数据传输、存储、使用以及开放分享时要做好脱敏、加密处理,预防个人信息泄露,数据调用以及传输途径要做到可追踪,实现数据管理规范化、流程化。三是涉及安全等级较高的数据集时要防范技术对原数据的还原所导致的隐私泄露。

(2) 保障人工智能社会实验对象的自主性权利

保障知情同意权利亦是科学研究的基本伦理准则(A1、B1、C1、D1、D3、E2~E4),人工智能社会实验过程中应当尽可能地通知可能涉及到的实验对象,然而在实际实施中将面临两方面挑战:一是开放场景中很难精准识别实验对象的范围;二是在一些情况下如果严格实行知情同意社会实验将无法开展。如在开放街道上进行人脸识别技术相关社会实验,实验对象为街道上行人,如若对每一个行人进行知情同意的规范操作,成本以及操作性的挑战很大。借鉴《社会科学与人类学伦理指南》中有关隐蔽性实验的规定,人工智能社会实验在进行实验前,若无法进行或者全部覆盖知情同意规范操作,可采取相关灵活方式。主要方式:一是要求匿名收集数据,不能涉及个人敏感信息;二是收集内容需要谨慎考虑,要尊重当地文化习俗;三是数据收集应尽可能公开、透明,最好有符合相应资质的第三方在场见证。

此外,要确保实验对象可以对社会实验的开始、进行、退出、终止以及实验披露有话语权和知情权。关于人工智能社会实验过程中的数据存储、使用以及时效,要告知实验对象或对其保持透明(B2、E2~E4)。人工智能社会实验需要构建与实验对象或者公众沟通的有效渠道,发布负责人的联系方式,对于在线开展的社会实验,可以考虑开发知情同意、个人信息管理程序或是采用去中心化的解决方案来保障公民知情同意权利。

表 1 各领域研究规范与人工智能社会实验规范

规范	《人体生物医学研究国际道德指南》	《药物临床试验质量管理规范》	《中国心理学会临床与咨询心理学工作伦理守则》	《社会科学类与人类学伦理指南》	《技术社会实验伦理框架》	人工智能社会实验规范
实验对象权利保护性规范	A1 尊重受试者自主权利;所有以“人体”为受试对象的生物医学研究,必须征求受试者知情同意 A2 保护隐私;必须采取科学合理的数据管理措施,保护受试者隐私数据安全	B1 受试者的权益保护和生命是首要考虑因素,优先于科学和社会获益。伦理审查与知情同意是保障受试者权益的主要措施	C1 知情同意:保障寻求专业服务者开始或维持与心理师专业关系的自由选择权利,在进行专业服务前以及过程中均有权充分了解服务过程和理论取向 C2 隐私权和保密性:心理师对保护寻求专业服务者的隐私具有主体责任	D1 自主原则、知情同意;尊重研究对象的自主权利,保障其个人自由和知情同意权利,研究过程中应保持对研究对象的诚实和信息透明 D2 隐私保护:保护隐私权利,做好对个人信息的保护措施 D3 尊重原则:尊重人的尊严,保障研究对象人格完整性	E1 保护实验对象数据隐私安全 E2 实验对象被告知 E3 实验对象可以对开始、进行、终止实验有话语权 E4 设计实验对象可以退出社会实验的机制和渠道	1. 保护人工智能社会实验对象的隐私权利 2. 保障实验对象的自主性权利
实验准入的制约性规范	A3 伦理审查机制:涉及人类受试者的研究方案须经由一个或一个以上的科学伦理审查委员会对其科学价值和伦理的可接受性进行详实地审查	B2 试验方案应当清晰、详细、可操作。试验方案在获得伦理委员会同意后方可执行 B3 试验药物的制备应当符合临床试验用药物生产质量管理规范。试验药物的使用应当符合试验方案 B4 参加临床试验实施的研究人员,须受过相关专业教育、培训,并取得相应的职业资格认证,同时具备在临床试验中相应的工作经验	C3 研究人员的专业性:专业服务人员需要持有专业证件上岗服务	D4 伦理的自我审查:特别要注重对于研究可能产生“意外的/潜在的/偶然的”违反伦理的结果自我推测与审查 D5 外部审查机制:遵循各国、各自领域以及机构的伦理审查机制与法律规范	E5 缺乏获取技术危害知识的其他合理手段时才开展社会实验,即开展社会实验前,已经通过所有的合理方法确定技术危害 E6 经民主合法化机构批准	3. 尽可能采用较成熟的人工智能技术与产品进行社会实验 4. 实验方案需要通过相应的审查机制认可
实验过程的可控性规范(数据和方法)	A4 明确受试对象参加研究的持续时间,确保实验提前中止以及受试对象提前退出的可能性	B5 临床试验的质量管理体系应当覆盖临床试验的全过程,全过程中实验操作不得违背相关法律法规,要确保受试者生命安全受到保护、确保试验数据和结果具有可靠性 B6 要符合药物临床试验质量管理规范(GCP)的标准规定,包括方案设计、组织实施、监查、稽查、记录、分析、总结和报告 B7 研究者在临床试验过程中应当遵守试验方案,凡涉及医学判断或临床决策应当由临床医生做出 B8 谨慎保存、管理可鉴别受试者身份信息的相关记录。所有临床试验的纸质或电子资料应当被妥善地记录、处理和保存。记录数据须具备精准性、可解释性、可论证性	C4 伦理问题处理:心理师应在日常专业工作中努力践行专业伦理规范,在专业工作中应当遵守有关法律和伦理规范。心理师应努力解决伦理困境,和相关人员进行直接而开放的沟通,在必要时向督导及同行寻求建议或帮助	D6 研究人员专业性:要求研究人员具备足够专业水平,对研究中潜在的伦理问题能够保持敏感 D7 方法、数据规范性:对研究方法、数据进行规范处理 D8 充分考虑特殊场景中的敏感问题	E7 对应用场景延伸及拓展进行前期规划,防止造成大规模的伤害,增进对“意外事件”的学习能力 E8 灵活设立实验,避免技术锁定效应 E9 避免缺乏系统弹性的技术社会实验 E10 尽可能合理地控制风险 E11 社会实验的设立、执行、监测、评估、调整以及终止均有清晰的责任主体 E12 确保能够终止实验的可能性	5. 科学规划人工智能社会实验场景 6. 对人工智能社会实验评估测量、方法、数据进行科学设计与规范管理 7. 明晰责任:社会实验的设立、执行、监测、评估、调整以及终止均有清晰的责任主体
实验效益的伦理性规范	A5 参加研究的受益和风险:所有的以“人体”为受试对象的人体生物医学研究,要最小化使用风险,同时保证潜在的利益和风险得到了合理的平衡	B9 临床试验应当权衡受试者和社会预期的风险和获益,只有当预期的获益大于风险时,方可实施或者继续临床试验 B10 应当回避实验相关利益人之间重大利益冲突	C5 心理测量与评估:应正确理解心理测量与评估手段在临床服务工作中的意义和作用,充分考虑被测量者或被评估者的个人特征和文化背景,恰当使用测量与评估工具来促进寻求专业服务者的福祉	D9 促进公平和多样性:注重对弱势群体效益分配,注重环境保护和可持续发展	E13 对潜在的危害和社会效益公平分配 E14 伤害可逆性,能够弥补或者降低伤害 E15 合理预期实验的社会效益 E16 避免将弱势群体纳入社会实验对象,或者采取措施保护社会实验中的弱势群体	8. 人类有益性 9. 社会公平性

(二) 实验因素的制约性规范

(1) 尽可能采用较成熟的人工智能技术与产品进行社会实验。借鉴《药物临床试验质量管理规范》,医疗产品在进入临床实验前应以其他形式进行过相关产品测试(例如动物实验)以及质量管理规范(B3),人工智能社会实验应秉持无伤害原则,社会实验开展应在有较好的人工智能技术应用实践基础、相关技术成熟度较高、技术水平具备规模应用的条件下开展(E5)。人工智能技术、产品在进入社会实验前应当尽可能在实验室完善且对已知的社会风险采取规避措施,并通过人工智能相关技术标准、产品质检、产品市场应用许可等。对于还未有规范或者标准可以依据的技术和产品,应开展技术风险前瞻性评估,不仅仅是对技术的效能方面,对其技术应用与社会交互的方式影响,例如商业模式、场景边界等也需进行前瞻性评估。对于社会实验过程中不可预计的风险,需跟踪应用状况,发现风险后积极探索应对措施。

(2) 实验方案需要通过相应的审查机制认可。在社会实验实施之前,所有的实验方案应在自我审查后提交给相关机构或组织进行外部审查(A3、B2、C2、D4、E5)。审查内容应包括社会实验研究团队的专业性(B4、C3、E6)、符合相应的法律规范的底线要求(B3、D5)、对潜在的风险的论证和审查(C4)等,审查机制应做到程序化(A3、B2),审查主体应是具有资质的合法组织(A3、E6),例如相应级别的科学伦理审查委员会。

(三) 实验过程的可控性规范

(1) 科学规划人工智能社会实验场景。人工智能社会实验与具体的社会情境高度融合:一方面,人工智能社会实验对于场景具有高度依赖性,科学规划场景环境、合理应用场景数据对于规避实验的系统风险起到关键作用;另一方面,人工智能的社会实验与场景中的“人”互动密切,因此要提前对应用场景做好风险控制的规划,确保社会实验场景中的“人”不受伤害(A4)。如无人驾驶社会实验需要对实验场景进行提前规划,明确实验持续时间、地点以及参与人员,测试车辆应在封闭道路、场地等特定区域进行充分的实车测试,对

于场景中的测试路段、测试时间、测试项目、测试规程均需要提交详细规划方案、风险分析及应对措施(参照《智能网联汽车道路测试管理规范(试行)》)。同时,要关注到实验对“人”产生的心理性伤害,特别要避免一些极端实验给实验对象造成例如创伤后应急综合症等精神疾病的严重后果(C4、D8)。

(2) 对人工智能社会实验评估测量、方法、数据进行科学设计与规范管理。人工智能社会实验具有长时间周期、宽空间区域的特征,实验过程中的技术风险以及实验系统风险具有动态性以及不确定性,因此对人工智能社会实验进行全程跟踪监测、实时评估,从而保证风险敏捷控制显得尤为重要(B5)。一是可以采取定量方式进行评估:从生命周期各个过程提取实验的风险因素,构建风险测量指标与风险管理模型,同时要设计实验中止、终止的措施,当风险程度超过警戒阈值时启动风险警报(A4、B6、C4、E10、E12);其次,可以通过实证研究,采用大数据方法,系统梳理人工智能社会应用风险,构建人工智能社会实验风险关键控制点判定图谱^[26],对社会实验研究方法、数据进行规范处理(B6~B8、C6~C7、E7),在关键点上进行干预机制设计。二是同时可采取定性方式,根据具体的实验特点以及实验环境进行分析,形成对定量方法的补充,防止特殊情况下风险被忽视的可能,并考虑设置备选方案,实现人工智能社会实验敏捷可控(E8、E9)。

(3) 明晰社会实验各主体责任。人工智能社会实验的设立、执行、监测、评估、调整以及终止等环节要确保主体责任清晰,确保每个环节能够责任落实,问责有效(D11)。人工智能社会实验主体包括政府、学术专家、企业、公众等多元主体,因此对于主体关键性责任应进行明确定义。如政府主体责任应在于顶层设计、组织、推动社会实验工作,给予社会实验顺利开展提供政策、人才等保障。企业主体责任在于场景以及技术的支持,为社会实验提供研究场景以及开放技术研发应用过程。学术专家则应在设计实验方案、构建关键测量指标、采用科学测度方法、数据收集分析等方面

充分发挥专业性,确保实验的科学性、有益性以及伦理性。其次应在制度和管理上进行规范和保障。一是在实验前,对实验的组织形式、管理流程,各主体的工作内容、职责范围等情况进行专业调研和评价。建立完善项目负责人制、定期督查制、定期报告制、适时查新制、目标责任制(B5~B7)。二是在实验过程中,从研究内容的任务分解、任务工作描述、任务责任分配、任务工作关系、任务工作时间估计等方面进行工作分解和进度计划的制定,并采取机制进行过程把控,以保证课题按计划进行(B5~B8)。

(四) 实验效益的伦理性规范

(1) 有益性。首先,人工智能社会效益设计强调技术有益性原则。技术有益性是技术创新的根本目的,人工智能的社会实验目的亦是强调技术在实验性过程中形成对人类社会有益的应用范式。例如阿西洛马人工智能原则中的共同繁荣原则、分享利益原则以及公共利益原则^[28]均强调人类共同利益,而不是某一单一组织、机构、国家甚至地区的单边利益(B9~B10、E15)。其次,人工智能社会实验要具有支撑智能时代治国理政的功能,在人工智能社会实验过程中积累人工智能治理的新知识,总结现有的技术或者技术社会应用系统中的短板并及时反馈到技术开发方,助力技术更新迭代、健康发展。同时,针对现有治理体系的知识缺失和不足提出新的干预措施,升级迭代人工智能治理系统,为技术研发者、社会、政府有效防范人工智能普及的潜在社会风险提供参考。

(2) 公平性。一是人工智能社会实验要注重对弱势群体的保护(D9、E16)。弱势群体包括但不限于儿童、难民、非常规移民、性工作者、认知受损人群、少数民族与宗教人士、创伤后人群以及与研究团队依从关系人群。首先,对于一般性社会影响规律的研究要尽量避免将弱势群体纳入社会实验对象范围。其次,关注弱势群体声音可能是人工智能社会实验应有之义。因此,当实验对象必须包含弱势群体时,要综合群体特性与文化背景,将安全风险最小化,同时要保证团队中有具

备相关专业知识的人员进行专业防控把握(C5)。如针对儿童的教育社会实验,应有儿童心理专家全程参与实验过程。要充分注重弱势群体感受,避免弱势群体在实验中感受到歧视或不被尊重。二是要注重实验利益的公平分配,维护研究的回报公正、分配公正(A5、E13)。要明晰实验对象参与实验的利益回报形式和价值额度,在没有报酬情况下要清晰地告知实验对象。三是要明确实验产生伤害的补偿应对方案,在参与实验前应向实验对象明确补偿形式及价值额度(E14)。

四、人工智能社会实验负责任运行机制

实验伦理审查的运行机制也是降低实验风险的关键。在人工智能社会实验运行过程中,有几点值得关注。

一是发挥专家力量,注重多元主体联动。人工智能社会实验是指利用政府和市场的力量推动特定人工智能技术在社会治理中广泛应用,强调多元主体的联动进行,包括政府、企业、场景方、专家团队、公众等多元实验主体^[1]。人工智能社会实验是在延续我国“政策试点”的“先试点后推广”的实验思维基础上,强化突出路径的科学性和伦理性。因此,“专家力量”应成为社会实验主体间联动的关键。首先,专家应参与政府顶层设计,帮助政府科学统筹、有序推动,形成科学的研究体系和工作体系,根据实验需要提供相应的资源保障和政策支撑;其次,专家应深入场景方和企业调研,帮助场景及企业充分发挥自身数据优势,围绕具有应用基础的产品以及应用场景科学搭建实验环境,保障实验能够顺利开展的数据、基础设施、技术等基础条件完备;再次,实验的科学设计、实施、监测、分析以及凝练出政策反馈均需要跨学科专家团队提供基础性的智力支撑,保障研究任务的科学开展、有序完成;最后,实验的伦理性也需要相关专家进行研判、审查,保障实验在具备科学价值同时,其全过程是符合伦理道德和人类福祉的。此外,要充分发挥公众、媒体、行业协会等第三方的监督协同作用,广泛吸纳多方声音,实时分析反馈,最终应形成统一领导、发挥优势、科学有序、社会监督的多主体参与的工作体系,推进人工

智能社会实验系统性开展。

二是注重实验运行形成反馈闭环。人工智能社会实验要遵循反馈性和反身性原则,从全过程视角对人工智能社会实验的运行机制进行架构,支撑人工智能负责任发展。实现从社会实验前对人工智能新技术或产品的实验室结果技术判定和社会影响评估,达成风险控制方案设计后,进入社会实验,形成持续跟踪、评估、学习、反馈的一个闭环治理机制。单一的实验治理循环结束后会实现持续循环,即新知识将反馈到新的治理制度之中,而新的治理制度也会进入新知识和新理论的产生过程中,通过该闭环实现人工智能技术以及

技术与社会交互方式的不断迭代(见图1)。

三是注重研究的程序公正。公正的研究程序和审查评价程序是展现科学研究的过程价值,通过设置相应顺序、方式、步骤与规则,可以保障科学研究的目标能够代表各主体的利益。与人文社会科学研究相同,人工智能社会实验研究程序和审查评价程序也应体现“主体平等、实质参与、价值中立、程序理性、程序公开、程序自治”六大原则^[25],构建人工智能社会实验公正的程序,保障具体开展中的过程可控性、决策可控性、一致性、无偏性、准确性、可修正性、代表性以及伦理性^[26],实现人工智能社会实验研究负责任开展。

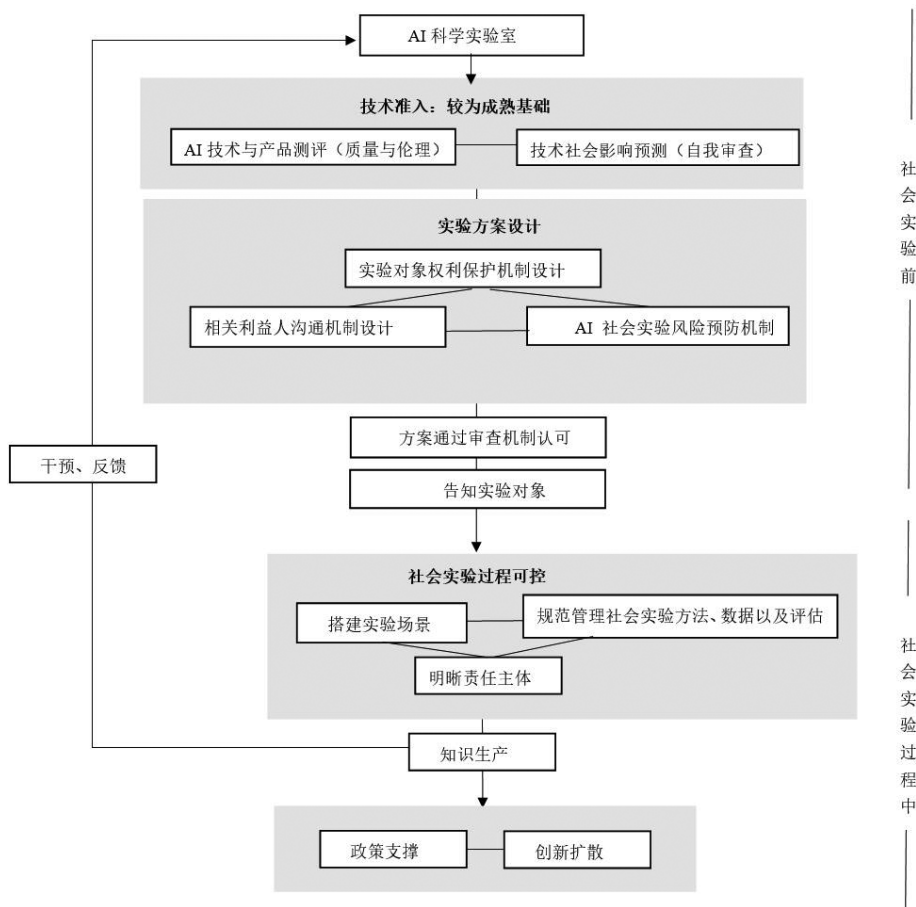


图1 人工智能社会实验运行闭环

五、结语

新一轮科技和产业革命已经来临,人工智能越来越广泛地和人类的生产生活活动相结合,从各方面开始重塑人类社会。人工智能社会实验采取科学方法对技术潜在风险进行系统研究和论证,通过研究相关伦理、社会问题保障我国人工智

能技术稳健发展是其重要内容。囿于技术内生性风险与实验系统风险,实验将面临隐蔽性、扩大性等新挑战,伦理问题在具体开展过程中会趋向复杂化。因此,开展人工智能社会实验需要特别重视伦理问题^[1],确保实验能够负责任地进行。在此背景下,本文借鉴多个领域已有实验规范,结合

技术与实验风险分析,尝试构建人工智能社会实验规范,并在此基础上提炼人工智能社会实验运行机制,为学界后续实验在防范潜在社会风险方面提供参考,也为各个试验区有序开展相关实验提供伦理规范方面的理论支撑。诚然,由于人工智能社会实验是采用新兴研究方法为政府决策提供循证支持的开创之举,一些实验风险需要在具体的实践中才能显露。因此,未来研究需要以探索完善新科研范式下的科研伦理机制为追求,持续跟进社会实验研究工作,观察、积累相关经验,为规范开展人工智能社会实验做出新的贡献。

参考文献:

- [1] 苏竣,魏钰明,黄萃. 社会实验:人工智能社会影响研究的新路径[J]. 中国软科学,2020(9):132-140.
- [2] 邓晖. 清华举办首届全国人工智能社会实验学术会议[EB/OL]. (2019-10-15) [2021-07-04]. https://news.gmw.cn/2019-10/15/content_33235663.htm.
- [3] KROHN W, WEYER J. Society as a laboratory: the social risks of experimental research[J]. Science and public policy, 1994,21(3):173-183.
- [4] 刘玉强. 社会实验:作为一种技术治理路径——以氟氯烃的社会引入为例[J]. 自然辩证法研究,2017(4):63-67.
- [5] VAN DE POEL I, VERBEEK P P. Ethics and engineering design[J]. Science technology & human values, 2006(3):223-236.
- [6] VAN DE POEL I, ROYAKKERS L. Ethics, technology, and engineering: an introduction[M]. Chichester: Wiley-Blackwell,2011:161-194.
- [7] HERBOLD R. Technologies as social experiments: the construction and implementation of a high-tech waste disposal site[M]// RIP A, MISA T J, SCHOT J. Managing technology in society: the approach of constructive technology assessment. London: CENGAGE Learning Business Press,1995:185-198.
- [8] 刘玉强,齐昆鹏,赵公民. 大型社会技术系统的实施验证——“事故”作为一种学习资源[J]. 系统科学学报,2018,26(4):35-40.
- [9] YATES F. The early history of experimental design[C]// SRIVASTAVA J N. A Survey of statistical design and linear model. Amsterdam: North-Holland, 1975:581-592.
- [10] LIST J A, RASUL I. Field experiments in labor economics[J]. Handbook of labor economics, 2011(4):103-228.
- [11] ADDAMS J. Ten years' experience in Illinois[J]. The annals of the American academy of political and social science, 1911,38(1):144-148.
- [12] BROWNR. Artificial experiments on society: Comte G C. Lewis and Mill[J]. Journal of historical sociology, 1997,10(1):74-97.
- [13] GROSS M, KROHN W. Society as experiment: sociological foundations for a self-experimental society[J]. History of the human sciences, 2005,18(2):63-86.
- [14] MAYO E. The human problems of an industrial civilization[M]. New York: Macmillan Co,1933:53-74.
- [15] OAKLEY A. Public policy experimentation: lessons from America[J]. Policy studies, 1998,19(2):93-114.
- [16] 谭铁牛. 人工智能的历史、现状和未来[J]. 求是,2019(4):39-46.
- [17] 中国信息通信研究院. 人工智能数据安全白皮书[R]. 北京:中国信息通信研究院,2019:1-45.
- [18] 何积丰. 安全可信人工智能[J]. 信息安全与通信保密,2019(10):4-8.
- [19] 李仁涵. 人工智能技术属性及其社会属性[J]. 上海交通大学学报(哲学社会科学版),2020,28(4):19-22.
- [20] 胡元聪,廖娟. 人工智能的负外部性及其经济法规制[J]. 大连理工大学学报(社会科学版),2020,41(3):71-79.
- [21] 魏薇,景慧昀,牛金行. 人工智能数据安全风险及治理[J]. 中国信息安全,2020(3):82-85.
- [22] WYATT T. Universal basic income trial in Finland fails to help unemployed people back to work[EB/OL]. (2019-02-08) [2022-06-22]. <https://www.independent.co.uk/news/world/europe/finland-universal-basic-income-trial-unemployment-experiment-trial-a8769621.html>.
- [23] BBC. Facebook emotion experiment sparks criticism[EB/OL]. (2014-06-30) [2022-06-22]. <https://www.bbc.com/news/technology-28051930>.
- [24] 李修全. 人工智能应用中的安全、隐私和伦理挑战及应对思考[J]. 科技导报,2017,35(15):11-12.
- [25] 谭春辉. 人文社会科学研究评价程序公正探讨[J]. 重庆大学学报(社会科学版),2013(5):93-99.
- [26] LEVENTHAL G S, KARUZA J, FRY W R. Beyond fairness: a theory of allocation preference[C]//MIKULA. Justice and social interaction: experimental and theoretical contributions, from psychological research. New York: Springer - Verlag,1980:167-218.

(本文责编:王延芳)