文章编号:1003 - 2053(2026)04 - 0746 - 11

生成式 AI 支持下社会科学知识生产的共同体德性

张 瑞,司孟韩
(国防科技大学军政基础教育学院,湖南长沙 410073)

摘 要:生成式人工智能技术已经嵌入社会科学知识生产研究的各环节,不仅重塑了社会科学知识生产的基本过程,更引领着智能化科研范式成为社会科学知识生产领域的主流研究范式之一。然而,生成式 AI 在处理庞杂的数据过程中也存在着数据确权不清、数据偏见、数据幻觉和数据傲慢等挑战,侵蚀着社会科学知识生产的共同体公共性、公正性、客观性及批判性等德性。为提升生成式 AI 在社会科学知识生产中的效能与质量,提高社会科学知识生产的共同体德性,需构建“政府 - 企业 - 学界”三元协同治理框架,以实现“知识共创,人机共生”的理想协作愿景。

关键词:生成式 AI;数据;社会科学知识;共同体;德性
中图分类号:B82 - 0 **文献标识码**:A

当前,生成式人工智能技术(Generative Artificial Intelligence)已经嵌入社会科学知识生产研究的各环节,在重塑社会科学知识生产基本过程的同时,也影响着社会科学知识生产的共同体德性。关于科学知识生产的共同体德性,目前学术界尚未有统一定义。1942 年英国科学哲学家迈克尔·波拉尼(Michael Polanyi)在《科学的自治》一文中首次提出并使用科学共同体(Scientific Community)概念,指由科学观念相同的科学家组成的集合体[1]。古希腊哲学家亚里士多德(Aristotle)在《尼各马可伦理学》中指出:“我们称那些值得称赞的品质为德性”[2]。综合以上观点,我们认为“社会科学知识生产的共同体德性”,不同于一般意义上的个人德性,而是指从事社会科学知识生产的共同体及其成员在研究中所应具备的品质,这种品质对于共同体本身的福祉和存在是必要的。美国社会学家罗伯特·金·默顿(Robert King Merton)提出了科学共同体的精神特质即德性,包括具有普遍性、公有性、大公无私和有根据的怀疑态度[3]。然而,当科学共同体延伸至社会科学领域时,其德性内涵发生了本质性拓展。与自然科学以客观存在的物质、现象及其规律为研究对象不同,社会科学以人类社会的组织结构、文化传统与价值系统为研究对象,这种特殊性使社会科学知识生产直接介入社会价值建构,要求社会科学知识生产者不仅需要遵循默顿提出的科学共同体德性,更应具有公共性、公正性、客观性以及批判性德性。

生成式 AI 技术是“具有文本、图片、音频、视频等内容生成能力的模型及相关技术”[4],核心是将海量训练数据中的潜在规律转化为生成新内容的算法框架,形成从统计模仿到涌现创新的创作链条。这一转化与构建过程,有力推动了社会科学知识生产与数据信息互动范式的重构,但同时也导致了社会科学知识研究模式的结构性变革,引发了社会科学生产者聚讼纷然。面对生成式 AI 技术主导社会科学知识研究范式重构的新趋势,需要厘清数据确权与共同体的公共性、数据偏见与共同体的公正性、数据幻觉与共同体的客观性以及数据傲慢与共同体的批判性之间的关系,以强化社会科学知识生产的共同体德性。

1 数据确权与共同体的“公共性”

“数据确权”是数据权属议题在系统层面的延展,旨在解决数据权利的归属问题,即哪些数据权利主体能够在何种程度上享有数据所附着的相关权

收稿日期:2025 - 03 - 05;**修回日期**:2025 - 06 - 23
基金项目:国家社会科学基金重大项目(22&ZD044);军队高层次科技创新人才工程科研项目(22TDRCJH - 02 - 005)
作者简介:张瑞(1995 -),女,博士研究生。司孟韩(1997 -),女,博士研究生,通讯作者,E - mail:smh123400@163.com。

利[5]。古罗马权利分割思想对数据确权制度的建立具有重要的启示意义。罗马法将物权分为自物权和他物权,自物权人可根据需要在一定范围和期间内出让部分权能,为他人设立他物权。由于他物权派生于自物权,且权利内容受特定目的限制,故又被称为派生物权或限定物权[6]。现代所有权理论关于其本质要素的讨论也表明,所有权的权能是可以分割的。德国学界基于权利分割思想(Abspaltungsgedanke),提出了源权利(Quellrecht)和限制权利(beschränktesRecht)两组概念,源权利人可以从源权利中分离出有限权利,并将其授予其他人。以所有权为例,源权利人可以将所有权的使用权能与变价权能分离,创设出用益物权与担保物权,以实现物的使用价值与交换价值,达到物尽其用[7]。党的十九届四中全会明确指出“数据”是数字经济深化发展的核心引擎,是最具时代特征的生产要素[8]。其中,数据确权是完善要素市场化配置体制机制中的重要一环,是数据市场建设和数字经济持续发展的基础性制度[9]。2022 年 12 月 19 日,中共中央、国务院正式对外发布了《数据二十条》,指出要“保障其投入的劳动和其他要素贡献获得合理回报”,并提出建立保障权益、合规使用的数据产权制度[10]。通过加快数字确权的推进,能够在为数字经济发展提供制度性回应的同时,也为社会科学知识生产提供数据流通和利用的规则依据,使知识产权保护下的“知识”真正成为全社会共同进步的源泉。

科学服务于全人类,这是科学公共性本质决定的。社会科学知识生产共同体的公共性德性是指社会科学研究和知识产生、传播、应用过程中的公共特征和社会维度。数据确权与社会科学知识生产共同体的公共性之间存在着相互促进和依赖的关系,因为它涉及如何在保护个人隐私和促进数据共享之间找到平衡点,以及如何在不同利益相关者之间公平地分配数据权利和利益。一方面,数据确权为社会科学知识生产共同体的公共性培养提供了法律和制度基础。明确的数据确权规则可以为数据权利提供法律保护,也可以为数据流通和利用活动提供规则依据,有助于促进数据生产、流通与利用的有序进行,提升数据要素配置的市场化程度,从而有效满足社会科学知识生产对大量数据的需求。另一方面,社会科学知识生产共同体的公共性要求数据确权政策能够支持数据的开放和共享。为了确保社会科学知识生产共同体充分而自由地使用科学成果、开辟知识边界,需要有明确的数据确权机制,以便于这一共同体能够合法地访问和使用这些数据。需要注意的是,适当的数据确权政策可以促进数据的合理利用,支持社会科学知识生产共同体对社会科学知识的深度挖掘,而过度的确权则限制了研究数据的获取和使用。因此,在制定数据政策时,需要综合考虑促进知识共享和保护数据权益的双重目标。

数据确权作为数据共享与利用的重要前提,其存在的制度性、技术性问题对社会科学知识生产共同体的公共性德性带来了挑战。首先,由于数据的无形性、来源复杂性、应用场景丰富性等特点,使数据的所有权、使用权和流通权难以界定和掌握,导致科学知识生产共同体在研究中很难把握好数据与隐私之间的关系。生成式 AI 是具有创造性的,其创造过程需要摄入海量数据,这些数据或来自开源数据集、或通过互联网爬虫来获取,亦或通过向数据公司购买获得,一些机构在训练时违规收集大量个人信息,或使用没有合法来源的数据训练,给个人信息泄露和数据安全埋下隐患[11]。如果研发机构在训练 AI 时未能获得训练材料提供者的充分授权,就很可能产生知识产权纠纷,尤其是对社会科学知识的文字产出更为容易产生争议。比如,全球最大的媒体集团之一新闻集团(News Corp)旗下道琼斯公司日前就发布了声明,对 OpenAI 在未经道琼斯授权的情况下就违规使用《华尔街日报》的行为进行了指责[12]。其次,数据确权中缺乏对数据的合法和道德使用的规定,影响社会科学知识生产共同体研究的社会性。数据确权的过程中,涉及复杂的法律问题,如数据的归属、使用和共享的法律界限,增加科学知识生产共同体研究的法律风险和成本。社会科学知识生产共同体因研究目的而收集数据,但如果没有明确的指导原则和监管机制,这些数据可能被盗用,引发诉讼、损害研究者的声誉和研究机构的公信力等问题。例如,在一项研究中,如果研究者收集到的相关信息被盗用,并用于商业广告、政治宣传或其他非预期目的,公众会感觉研究收集的信息可能存在泄漏风险,将导致公众对参与研究持保留态度,减少研究者进行数据收集的机会,从而限制社会科学研究的范围和深度,不仅损害了数据主体的权益,也破坏了研究的诚信和公共信任。最后,数据确权不清晰限制了知识的自由流通和共享,导致数据资源集中在有能力获取和处理这些数据的个体或组织手中,这可能使研究结果偏向于特定群体的观点和利

益,忽视了更广泛社会的需求和多样性。

社会科学知识生产共同体的公共性德性,强调了数据作为公共资源的特性,需要数据确权在保障公共利益的同时,考虑到数据的共享和开放。为增强社会科学知识生产共同体的公共性,需要采取以下措施:第一,采取政府主导,多方协同的方式,制定数据全生命周期管理政策并监督执行。数据治理政策的制定和执行是确保数据确权过程中保护公共性的关键。首先,国家数据局等监管机构应牵头建立跨部门协作机制,贯彻实施科技部 2025 年 1 月 24 日颁布的《科学数据安全分类分级指南》,在收集阶段要求研究机构获取伦理委员会批件,明确敏感数据识别标准与使用边界。其次,中国标准化研究院需联合高校智库制定数据处理技术规范,例如,规定社会调查数据的匿名化处理必须达到 k - 匿名度 ≥ 5 的技术标准。最后,监督执行层面,网信办可建立数据流通溯源平台,对政务数据开放进行实时合规监测。

第二,采取政产学研联动机制,促进数据共享和开放。数据共享和开放是实现社会科学知识生产共同体的公共性的核心原则,为此需要在加强数据隐私保护的同时促进数据的共享和开放。政府平台方面,国家自然科学基金委可施行政策激励的方式,要求受资助项目在结题后 6 个月内将脱敏数据上传至国家哲学社会科学文献中心。企业方面,呼吁头部企业承担企业社会责任,积极开放非竞争性训练数据集,如 2025 年 1 月 11 日,腾讯研究院联合 Bottle-Dream 在上海举办了“AI 向善语料库 · 行动者”会议,并正式启动“AI 向善语料库”2025 年共创计划 Pro 版[13]。同时,鼓励企业进行技术创新,增强数据隐私保护。如商汤科技等人工智能软件公司可提供自适应脱敏 API 接口,可动态调整数据扰动强度以平衡效用与安全。学校及学界层面,鼓励学界加入技术响应。高校数据中心应开发轻量化数据沙盒系统,允许跨机构研究者在虚拟环境中分析加密数据。社会科学知识生产共同体层面,应采取最小化原则,仅收集实现研究目的所必需的数据,并确保数据主体了解其数据如何被使用。同时,鼓励社会科学研究者在保护隐私的前提下,共享数据集和研究成果,如资助机构和学术期刊可以要求研究者在申请资金或发表文章时,说明他们的数据共享计划。多方联合发力,通过建立数据共享平台和标准,促进研究数据的广泛传播和利用,从而加速社会科学知识的积累和创新。

第三,建立社会科学知识生产共同体主导,社会力量参与的伦理审查机制。伦理审查机制的建立是确保社会科学知识生产过程中数据确权活动遵循伦理标准的重要保障。正如南非比勒陀利亚大学数据科学教授武科西 · 马里维特(Vukosi Marivate)所言,AI 治理是一项团队运动,道德决策和责任应由全社会共同承担[14]。首先,高校应履行学界核心职能,学校学术委员会负责常规审查,重点评估研究设计中的知情同意流程。其次,专业机构扩展延伸职能。如中国社会科学院需组建跨学科伦理审查特别委员会,对涉及民族、宗教等敏感领域的研究进行听证式审查。伦理审查委员会应该负责审查研究提案,确保研究设计、数据收集和使用过程符合伦理准则。这些准则应包括对参与者隐私的保护、数据的公正使用以及研究结果的透明报告。为了提高伦理审查的有效性,委员会成员应包括来自不同学科的专家、法律顾问。并且,需定期对伦理审查流程进行评估和更新,以适应新的研究方法和数据技术的发展。最后,政府保障方面,需要司法部建立专家库为委员会提供法律意见,特别是对数据跨境流动中的主权问题提出合规建议。同时,借鉴马里维特提出的“团队运动”理念,提倡社会共治机制,即公众代表可通过数字孪生平台参与模拟审议,对社区画像研究等公共性项目行使监督权。

2 数据偏见与共同体的“公正性”

“数据偏见”是指在数据收集、处理与分析的各个环节中遭遇边缘化、覆盖面不全或样本具有偏向性等问题。生成式 AI 基于带有数据偏见的数据生成的结果,会陷入错误性、不完整性的泥沼,进而引发误导性主张的危机[15]。数据偏见产生的逻辑根植于认知偏差、社会文化背景、技术限制、经济利益、政策和法规等多个层面。经济利益可能驱使某些数据被优先收集,政策和法规限制可能影响数据的可用性和多样性,研究者或数据科学家的个人信念和偏见可能影响数据的选择和处理,社会文化中的刻板印象和历史不平等也可能在数据收集和处理中被编码。如美国平等就业机会委员会观察到,“高科技”行业中白人比例过高,这一数据如果被收集用于 AI 模型训练,则模型可能产生白人更擅长从事高科技行业的结果,而对其他非白人群体产生歧视。

生成式 AI 在模型训练过程中的复杂性与阶段性,使数据偏见存在于多个环节。

在数据收集阶段,数据来源的单一性使数据集缺乏对所有相关群体的全面代表。生成式 AI 在数据收集时主要选取某一特定群体或地区具有代表性的数据,尤其是在文化和语言方面,导致模型在处理其他群体或地区的数据时可能表现不佳。耶鲁大学的一项研究揭示了这一点,基于 AI 的皮肤病学算法在诊断不同种族人群时,准确性存在显著差异,这背后的原因,正是训练模型所采集的数据大部分来自白人,导致算法在应对其他种族时产生了偏见[16]。另外,对数据进行采样和选择的偏差同样导致了数据偏见。数据采集过程中,如果采样方法不当,可能导致某些群体的数据被过度代表或不足代表。例如,在线调查可能更多地吸引年轻人,而忽视老年人的声音。同时,在数据收集选择的过程中,研究人员可能无意中选择了符合自己预期的数据,忽略了其他可能同样重要的数据,导致收集数据不全面。例如,在情感分析任务中,不同标注者对同一条评论的情感标签可能会有不同的理解,从而导致标注数据的不一致。

在数据预处理阶段,数据清洗和特征工程的不当处理可能导致某些信息被错误地忽略或过度强调,从而产生偏见。数据清洗和格式化过程中对缺失值、异常值处理不当,去重逻辑缺陷,可能导致充斥着重复数据、无意义数据,从而产生数据偏见。同时,特征工程中的特征选择和数据转换步骤,如果未能恰当执行,则会导致 AI 的选择会强化固有的偏见或引入新的偏见。

在模型训练阶段,算法本身的设计、数据不平衡问题,以及训练过程中的偏差都可能导致模型学习并放大这些偏见。普林斯顿大学 Kapoor 团队便指出,至少 17 个领域和数百篇论文中存在可重复性危机[17],这背后,正是基于伪研究数据所训练的模型,其生成的研究结果也具有误导性[18]。

在模型应用阶段,确认偏误和反馈循环可能会进一步巩固和传播已有的偏见。这一过程经过几个环节:一是初始偏见,若在训练数据中存在隐性偏见,如历史招聘中女性晋升率低,基于这一数据偏见训练的模型可能会学习到“女性能力不足,男性更适合这一岗位”;二是确认偏误,标注时更关注男性简历中的“领导力”关键词,忽视女性简历中的同等描述;三是反馈循环,人工智能模型推荐更多男性候选人进入面试环节,成功入职的男性数据强化模型对“技术岗位 = 男性”的认知,女性候选人因推荐减少而获得的机会减少,系统内女性成功案例数据枯竭,由此形成恶性循环。

社会科学知识生产共同体的公正性德性涉及如何在社会科学研究中确保所有参与者和受影响群体的权益得到公平对待。这包括确保研究过程中的包容性,避免歧视和偏见,以及确保研究结果能够真实反映社会现象,而不是仅仅反映出特定群体或利益相关者的视角。因此,需要警惕数据陷阱现象,需要量化它们的偏见,警惕它们在准确性上的偏差[19]。生成式 AI 中的数据偏差对社会科学知识生产共同体的公正性影响表现在以下三个方面:

一是选择。传统的数据收集或者搜索软件具有选择性偏见的缺陷,即边缘群体在数据资源池属于失语状态,听者和说者双方都平等地面临解释资源缺乏的不利[20],这一缺陷同样体现在生成式 AI 模型中。生成式 AI 模型依赖于大量的数据进行训练,但是如果这些数据在收集和训练中就存在偏见和不公正性,(例如,某一群体或文化的样本过多),生成的模型就会继承这些偏见,在数据库中选择带有偏见的内容,并在生成的文本中表现出来,从而导致生成文本结果偏向特定群体的特征。而边缘群体发展出来的代替性解释资源通常无法被纳入占主导地位的解释资源之中[21]。这种不利偏差直接关系到社会科学知识生产共同体的公正性,影响社会科学知识生产的普遍适用性和公正性。

二是解释。生成式 AI 的训练数据存在偏见,导致其生成的内容不可避免地带有偏见,误导社会科学研究者对现象的解释,从而影响社会科学知识生产共同体的公正性。生成式 AI 的技术“黑箱”特质,虽然使社会科学研究得到了解决问题的工具,但是却无法充分解释解决问题的原理。因而,在表征层面,数据偏见导致生成式 AI 产生选择性解释,将社会现象简单化、表面化、程序化的现象,并通过影响社会科学知识的生产、传播和应用(特别是作为资源分配、社会服务以及政策制定的依据)进一步放大数据偏见。公共知识系统构成公共解释资源池,公共知识意味着一个社会的信息共享体[22]。而数据偏见则进一步扩大当前社会公共解释资源的鸿沟,构成对公共解释资源中边缘群体的伤害,在社会实践中造成不公正的结果。尽管生成式 AI 通过海量的数据处理从而实现自身的迭代升级,并以此为

科学知识生产者提供所需的高质量答案,但是基于数据偏见而产生的这种服务是带有某种倾向的一种偏见选择,极容易导致社会科学知识生产中公正性的消解。

三是互动。解释过程往往包含解释者和被解释者之间的互动。在社会科学中,某一社会现象的解释往往需要经过多轮修正,尤其是当涉及复杂的社会问题时,解释过程中的反馈机制能够帮助研究者更好地理解受众的需求和经验,进而对解释框架进行调整。然而,因不同主体占有和使用数字技术的客观差异,导致了数字鸿沟的出现。在此形势下,一种极端的情况是所有被解释者都没有机会参与互动,还有一种常见的情况是部分被解释者没有机会参与互动过程,比如部分群体在数字社会中解释资源占有的空白,进而导致收集的数据所代表的群体不具有全面性,即存在数据偏差。根据第 54 次《中国互联网络发展状况》统计报告显示,截至 2024 年 6 月,我国网民规模近 11 亿人,较 2023 年 12 月新增网民 742 万人,互联网普及率达 78%[23]。但是需要注意的是,尽管我国互联网普及率在不断提高,但仍有接近四分之一的人还未接入互联网。这一过程存在的问题是,互联网接入不再是属于某一类特定群体的福利,相反,未接入互联网的群体逐渐被社会边缘化,在数据收集过程中被忽视,经验和需求无法通过 AI 的解释框架得到反映,造成社会公共解释资源的缺失,剥夺了边缘群体的认知贡献,偏见性的内容可能会在知识传播过程中被优先传播,而其他观点则被忽视,导致知识传播的不均衡。基于此,数据偏见影响社会科学研究结果的准确性、多样性和包容性。从而导致公众对社会科学知识生产共同体的研究结果的信任和接受度。

数据的客观性、中立性、全面性以及数据库的标准化程度,都将影响生成式 AI 算法决策的结果。因此,为了提升社会科学知识生产共同体的公正性,需要采取以下措施:第一,要明确数据标准,重视数据信息系统的建设。目前,美国、英国、新西兰等国家均重视数据信息系统的建设,如美国国家人文基金会(National Endowment for Humanities)被立法授权“开发一个实用的关于人文、学者、教育和文化团体及其受众的国家信息和数据收集系统”[24]。通过加强数据信息系统建设,能够为专家、学者、社会公众等相关服务对象提供更加广泛的数据信息获取途径和高质量的数据共享服务,从而提高社会科学研究中数据的客观性和全面性。首先,由政府主导建立标准制定,扩大 AI 训练数据文本的可选择空间。例如,国家标准化管理委员会应牵头制定《社会科学 AI 训练数据采集规范》,明确方言数据、少数民族文化数据等特殊语料的采集比例。其次,学界协同实施的多元化数据采集清单。在数据采集过程中,高校社科数据中心需建立动态抽样框架,采用空间分层随机抽样技术,确保城乡样本、职业群体等维度的均衡覆盖,以保证数据的全面性,并通过系统化的数据储存与精细化的管理,确保数据的准确性、完整性和一致性,为社会科学知识生产研究者提供全面的视角。最后,企业应开发智能数据标注平台,遵循 FAIR 原则(可查找、可访问、可互操作和可重复使用),内置数据代表性实时监测模块,当样本偏差超过预设阈值时自动预警。例如,地球化学图书馆(EarthChem Library),围绕业界标准整合数据集格式,与不同领域保持高度一致,并能在互联网上便捷地获取[25]。

第二,提高 AI 驱动的社会科学生成解释的互动性、开放性和对称性。在社会科学知识生产中,研究者需要企业研发结合不同历史时期或文化背景,与不同社会群体沟通交流,来揭示社会现象的多维度因果关系。因此,应对复杂的社会现象时,需要研发多模态交互式解释系统,允许用户通过自然语言对话修正模型认知偏差。学术界在机制的设计上要建立解释资源众包平台,邀请少数民族语言使用者参与提示词工程,动态优化文化敏感词库。网信办要制定《生成式 AI 社会研究交互规范》,强制要求人口统计学研究模型每 10 万次推理后必须进行边缘群体反馈校准以及在项目评审中需政府和学界协同增设“人机互证”环节,要求申请者提交传统质性研究与 AI 生成解释的交叉验证报告。通过提高生成式 AI 的反馈机制,能够不断地修正生成式 AI 对社会现象的了解,帮助研究者更真实地理解普遍社会现象和边缘群体的需求和经验,进而对生成式 AI 的解释框架进行调整,在一定程度上弥补了边缘群体在数字社会中解释资源占有的空白现象。

第三,建立公平性审查体系,识别和减少 AI 系统中潜在偏见。首先,政府需建立 AI 系统的数据收集、处理和使用的执行法律和监管框架。其中,政府监管机构应与学术界、工业界和民间社会合作,确保制定的法律和政策既符合技术发展,又能够保护边缘群体权利,同时定期审查和更新法律和监管框架,

以适应快速变化的 AI 领域。通过强有力的法律和监管,可以为生成式 AI 支持下的社会科学知识生产共同体的公正性的建立提供法律保障。其次,学术界应积极履行监督职能,在高校伦理审查委员会要引入对抗性评估机制,使用第四范式提供的 BiasBench 测评工具定期检测已发表成果的算法歧视指数。最后,企业在 AI 系统开发和部署之前,需要进行系统性的评估,包括对训练数据、算法选择、模型输出和决策过程的全面分析。

3　数据幻觉与共同体的“客观性”

数据幻觉是指在模型训练、数据分析或技术解读过程中,由于算法缺陷、技术手段的局限性或数据来源可靠性存疑,导致系统错误地生成、强化或曲解原本不存在的信息、模式或因果关系,最终产生看似合理实则偏离客观真实的研究结论或输出结果的现象,这种幻觉既可能源自技术系统的内在局限(如模型架构导致的归纳偏差),也可能反映人类研究者对技术中介数据的过度信任,本质上揭示了数据驱动决策中“技术可信度”与“客观真实性”之间的潜在裂隙[26]。在社会科学中,数据幻觉可以细分为理解幻觉(illusions of understanding)和客观幻觉(illusion of objectivity)。理解幻觉是指社会科学研究者在使用 AI 工具进行研究时,误以为自己对模型的工作原理和预测结果有着深刻的理解,但实际上他们可能只是依赖于 AI 的预测能力,而对背后数据生成的过程知之甚少。客观幻觉是指社会科学研究者错误地相信 AI 工具能够提供完全客观的分析和结果,不受任何偏见或立场的影响,但 AI 工具实际上是由人类设计和训练的,它们内嵌了训练数据和开发者的视角和偏见。因此,当社会科学知识生产研究者认为 AI 工具没有立场或能够代表所有可能的立场时,他们可能正经历一种客观性的幻觉[27]。

社会科学知识生产共同体的客观性德性是指共同体应该在研究中采取科学的方法,使研究结果能够被其他研究者验证和重复,并在研究中保持价值中立,避免将个人的价值观和偏见带入研究中。生成式 AI 支持下的社会科学知识生产正在打破传统科学知识生产的壁垒,随之而来的数据幻觉也对社会科学知识生产共同体的客观性产生影响。首先,数据幻觉降低科学社会知识生产者判断的客观性。AI 的运行过程存在算法黑箱,使得 AI 生成的内容难以追溯和验证,社会科学知识生产共同体如果过分相信 AI 工具能够提供完全客观、准确的分析结果,则可能导致基于 AI 提供的错误信息进行研究,逐步销蚀社会科学知识生产者对正确信息的敏感度以及客观性,降低社会科学知识生产的准确性。例如, ChatGPT 在某些情况下难以避免地会写出似是而非的错误或者荒谬的答案,生成缺乏科学内容的科学论文[28]。其次,数据幻觉消解社会科学知识生产内容的客观性。社会科学知识生产共同体的权威性建立在研究的严谨性和客观性基础之上。生成式 AI 基于收集到的海量数据可以生成任何文本,但这些用于分析的数据可能本身就带有价值偏见,由此生成的内容也不可避免地带有一定的价值观倾向,如果过分依赖生成式 AI 的分析结果,而不加以综合判断和审视,则可能会导致研究结果的非客观性。

社会科学追求的是事实和证据的知识生产,社会科学知识生产共同体的客观性德性是其研究的重要原则之一。只有经过专业训练,掌握特定的知识,遵循一定的规范,达到特定资格的人,才能进行科学研究。提高社会科学知识生产共同体的客观性,消解数据幻觉的弊端,可以从以下三个方面着手:

一是提高数据质量,扩宽社会科学知识研究的视野。数据是生成式 AI 的燃料,也是社会科学研究的基础,提高数据质量是确保社会科学知识生产共同体的客观性的关键。首先,对训练数据进行严格的筛选和清洗,建立动态的知识更新机制。如高校社科数据中心可以建立动态知识更新机制,每季度对中国知网收录的上万篇社科论文进行知识图谱重构,并通过阿里云提供的自然语言处理(NLP)工具自动识别过时理论范式,以对现有知识库的定期审查和更新,排除错误信息和偏见,减少数据源的误导性。其次,增强数据的全面性。加强政企协作,建立包含至少 38 种民族语言的口述史录音及转写文本,并通过区块链存证确保文化阐释的原真性,增加与异质性的交流,增加保持认知和人口统计多样性的优势,增强数据的代表性和全面性,以弥补生成式 AI 在社会科学知识生产中的局限性。

二是加强人机协作,建立人机融合的公共知识系统。公共知识包括社会科学、自然科学、艺术、文学、音乐等等,而在现代社会中,科学成为其公共知识体系中最重要的组成部分。通过建立人机融合的公共知识系统,可以帮助生成式 AI 系统的设计和应用符合社会科学研究的标准和伦理要求。首先,通

过学界把关制定 AI 生成内容的三重验证,包括传统统计检验、领域专家小组盲审、企业科技开发的生成溯源系统核查。其次,政府加大基建投入,部署社会知识融合平台,利用盘古大模型和公共知识资源,实现多模态数据的关联推理。同时,借鉴白惠仁提出的"基本善"理论[29],设立"人机协同创新"专项,要求受资助项目必须包含人类研究者与 AI 系统的分工协议。

三是增强算法透明度,构建包容性的解释资源。缺乏透明度可能导致算法偏见的隐藏和累积,从而影响社会科学知识生产共同体的客观性。通过提高算法透明度,开发者和研究者应提供算法决策过程的详细文档,并使用可解释的 AI 模型,使研究者能够追踪和理解 AI 生成内容的逻辑和依据,提高知识生产过程的客观性。同时,采用主流研究范式与边缘研究范式相结合,验证生成式 AI 输出结果的准确性,提高研究者对 AI 工具的批判性思维以及识别和纠正数据幻觉的能力,以保持社会科学知识生产共同体的客观性。

4 数据傲慢与共同体的"批判性"

数据傲慢是指在数据分析和解释过程中,过分自信和依赖数据,忽视了数据的局限性和可能的偏差。这种现象在社会科学研究领域表现为对数据无条件的信任,以及对非数据驱动的见解和方法的轻视。数据傲慢的产生是一个复杂的过程,随着计算机技术的进步,特别是云计算和移动互联网的出现,使人们拥有了前所未有的数据存储和处理能力,从而产生了大数据的概念。在大数据成为主要认知资源的背景下,传统的科学研究范式逐渐被抛弃,社会科学知识生产共同体过度依赖大数据进行研究的趋向明显[30]。在这种背景下,技术被认为是中立和客观的,社会科学知识生产研究者开始相信,只要有足够好的技术和足够多的数据,计算机就能比人更了解人类自己。但是这种对数据的盲目信任,使专业主义者们将自己拥有的知识毫不抵抗地让渡给了技术乌托邦者们,放弃对专业领域的常识判断和分析,好像社会科学研究只有在数据的包装下才能显得更专业,这种做法不仅导致了数据傲慢的产生,而且深化了社会对数据傲慢的容忍和传播。数据傲慢的存在导致社会科学知识生产共同体忽视数据收集和处理过程中的偏见,抑制社会科学共同体在知识生产过程中的批判性。

首先,数据傲慢阻碍社会科学研究中的批判性思维和创造力的发展。一是在生成式 AI 的热潮中,社会科学知识生产共同体对数据的盲目崇拜使人们往往忽视了数据的局限性和可能的偏差,以及人的主观能动性和社会正义性,从而使数据傲慢不断蔓延。这种现象"可能会导致学术失范,出现'机器越来越聪明,人类越来越懒、越来越笨'的局面,学术研究的意义和价值将会被消解"[31]。当前,生成式 AI 在人类反馈强化学习(RLHF)训练中具有承认错误和修复答案的功能,强化了学习和思考能力,使得科学知识生产者更加信任生成式 AI 的自修功能。但在这一过程中,研究者因惯性而接收的信息往往不加以"咀嚼消化",也弱化了作为社会科学知识生产能动主体的批判能力与行动意识,甚至形成一种惰性精神,削弱其继续深度探索的动力。二是生成式 AI 的输出结果是基于现有语库而来的,特点是在技术加持下的一种物化特征,无法复现人类最基础的活动,消解了社会科学的人文特质。这一特点导致生成式 AI 指导下的社会科学研究成果构成缺乏真实的情感体验和价值输出,无法解决复杂问题,可能还会产出某些偏见和陈旧观点,在一定程度上固化了研究者的思维,导致研究者接受数据所呈现的表面结论,阻碍社会科学研究者批判性思维和创造力的发展。

其次,数据傲慢导致社会科学研究中"人 - 机"主导权的角色调换。生成式 AI 对海量文献归纳总结的能力,使科研人员为节省时间选择 AI 替读。这一做法可能使人失去阅读原文的能力,从"替"读进一步走向弃读,从此只会接受由机器归纳的现成要点,导致学术研究人员阅读和分析科研文献能力的降低[32]。替读还可能引发误读,如当今的生成式大模型普遍存在"幻觉"或"杜撰"事实的缺陷,其中包括引用了不存在的科学研究[33]。在这种意义上,社会科学知识生产前期的资料收集、归纳和整理阶段,生成式 AI 的结果也是极不可靠的。如果过度依赖生成式 AI,导致人为机器所"奴役"而缺乏个人思考和人为矫正,只会为技术所反噬,逐步丧失某些领域的技能退化和知识缺位,继而引发"人 - 机"主导权偏移的角色调换,即由"训练"生成式 AI 变为被生成式 AI"训练"[34]。生成式 AI 所具有的深度合成功能使其可以产出以假乱真的虚假信息[35],在《错觉:AI 如何通过数据挖掘误导我们》一书中,加里·

史密斯(Gary Smith)指出,计算机在数据处理和记忆方面的优势是人类无法匹敌的,但这并不意味着计算机的数据分析结果就完全可靠[36],我们应该对计算机的数据分析结果保持批判性思维,而不是盲目信任。

批判性思维是人工智能无法取代的技能[37]。为了提高社会科学知识生产共同体的批判性德性,使其充分意识到自身在科学知识生产中的主体地位,在 AI 赋能社会科学知识生产的基础之上,保留并锻造社会科学知识生产共同体的批判性,可以采取以下策略:首先,加强数据素养教育。数据素养是社会科学研究者必须具备的能力,它包括对数据的敏感性、收集能力、分析处理能力以及利用数据进行决策的能力,提高这一能力能够有效避免数据傲慢对研究的干扰,提高社会科学研究者的批判性思维。为提高社会科学知识生产共同体的数据素养,学校可以增设“数据批判性分析”核心课程,并配套包含数据殖民主义、算法权力等前沿模块教材,供社会科学知识生产共同体学习。学界可举办“数据素养工作坊”,邀请专业工程师演示数据清洗中的认知偏差修正技术。通过教育和培训,可以提高社会科学知识生产共同体的数据素养,使其能够有效地获取、分析、处理、利用和展现数据,批判性地接受数据布道者思维,避免陷入“伊卡洛斯陷阱”。

其次,鼓励社会科学知识生产共同体对数据和算法保持质疑态度。社会科学知识生产共同体需要具备对数据的批判性思维能力,这涉及对数据来源、质量和潜在偏见的识别与评估。通过培养对数据的批判性思维,社会科学知识生产共同体可以提高成员的独立思考能力,打破“信息茧房”,避免陷入被数据集体同化的危机,提高对科学深入探索的能力。对此,政府可以部署数据素养实训平台,企业可以提供对抗样本生成工具供研究者模拟识别数据操纵陷阱,以提高共同体对数据的批判性思维能力。

最后,促进跨学科合作。生成式 AI 的发展为社会科学研究提供了新的工具和方法,但同时也带来了新的挑战。跨学科合作可以使研究者从不同角度审视数据和算法,从而提高研究的客观性和准确性。例如,计算机科学家、社会学家和伦理学家可以共同工作,以确保 AI 工具的开发和应用符合社会科学的伦理和方法论要求。通过对社会科学知识生产共同体德性的塑造,提高共同体人机共存的能力,使生成式 AI 科学合理地赋能于社会科学知识生产中,发挥 1 +1 >2 的智慧合力。

5　结语

从数据模式的机器识别到跨模态内容的自主生成,生成式 AI 的演进轨迹正在重构社会科学知识生产的底层逻辑。这种技术革命不仅催生了“数据驱动 - 算法生成 - 人类研判”的新型研究范式,更在深层次上重塑着社会科学知识生产共同体的认知维度:研究者既可通过算法洞察超越人类经验的数据关联,亦能在人机对话中突破固有理论框架的束缚。但技术赋能的狂欢背后,知识生产的公共性衰减、算法黑箱导致的公正性危机、数据殖民引发的客观性质疑等系统性风险已悄然显现。

在此双重境遇下,构建“政府 - 企业 - 学界”三元协同治理框架势在必行。政府层面需主导建立动态适应的制度体系,重点完善数据主权立法、构建算法影响评估机制,通过《生成式人工智能服务管理暂行办法》等政策工具平衡创新激励与风险防控;企业维度应践行技术向善准则,在模型开发中嵌入可解释性模块,开放训练数据集的元数据描述,并建立 AI 伦理影响声明制度,使商业逻辑与学术伦理形成价值共振;学术共同体则须坚守德性内核,既要运用批判性思维构建 AI 辅助研究的验证标准,又需通过跨学科合作重新定义人机分工边界。这种三元架构的协同效应,本质上是将默顿范式中的“有组织怀疑”升级为“有组织制衡”。当政府的数据治理框架、企业的算法透明机制、学界的批判性验证体系形成制度耦合时,生成式 AI 才能真正成为拓展人类认知边疆的“增强型智识伙伴”。而新一代社会科学研究者的培养,恰需在此交互场域中展开,使其既精通人机协作的算法工具,又深谙三元主体的责任边界,更有对知识社会效应的伦理自觉。唯有通过这种立体化、多向度的制度设计,社会科学知识生产共同体方能在技术极速发展中守住“安全、可靠、可控”的价值底线,使生成式 AI 驱动的知识革命,最终指向人类文明的创造性跃升而非解构性危机。这既是波拉尼“科学自治”理想在数字时代的范式转型,更是库恩式科学革命在认知伦理维度的必然延伸。

参考文献:

[1] Polanyi M. The autonomy of science[J]. The Scientific

Monthly,1945,60(2):141 – 150.

[2] 亚里士多德. 尼各马可伦理学[M]. 北京:商务印书馆,2003:110. Aristotle. The Nicomachean Ethics[M]. BeiJing:The Commercial Press,2003:110.

[3] R. 默顿,林聚任. 科学的规范结构[J]. 哲学译丛,2000(3):56 – 60. Merton R,Lin J R. The normative structure of science[J]. World Philosophy,2000(3):56 – 60.

[4] 中华人民共和国国家互联网信息办公室. 生成式人工智能服务管理暂行办法[EB/OL]. https://www.cac.gov.cn/2023 – 07/13/c_1690898327029107.htm. 2023 – 07 – 13/2025 – 03 – 20. Cyberspace Administration of China. Interim Measures for the Administration of Generative Artificial Intelligence Services[EB/OL]. https://www.cac.gov.cn/2023 – 07/13/c_1690898327029107.htm. 2023 – 07 – 13/2025 – 03 – 20.

[5] 杭敏,张亦晨. 从数据主权到数据确权:数字时代传媒发展的创新路径[J]. 中国编辑,2022(12):10 – 14. Hang M,Zhang Y C. From data sovereignty to data rights confirmation:An innovative path for media development in the Digital Age[J]. Chinese Editors Journal,2022(12):10 – 14.

[6] 周枏. 罗马法原论(上册)[M]. 北京:商务印书馆,1994:321. Zhou N. The Theory of Roman Law(Volume I)[M]. BeiJing:The Commercial Press,1994:321.

[7] Jürgen F,Baur,Stürner R. Lehrbuch Sachenrecht[M]. München:C. H. Beck,1999:773.

[8] 中国政府网. 国务院关于印发"十四五"数字经济发展规划的通知[EB/OL]. https://www.gov.cn/zhengce/content/2022 – 01/12/content_5667817.htm. 2022 – 01 – 12/2025 – 04 – 22. Chinese Government Website. Notice of The State Council on Printing and Distributing the "14th Five – Year Plan" for the Development of the Digital Economy[EB/OL]. https://www.gov.cn/zhengce/content/2022 – 01/12/content_5667817.htm. 2022 – 01 – 12/2025 – 04 – 22.

[9] 唐要家. 数据产权的经济分析[J]. 社会科学辑刊,2021(1):98 – 106,209. Tang Y J. Economic analysis of data property rights[J]. Social Science Journal,2021(1):98 – 106,209.

[10] 中国政府网. 中共中央国务院关于构建数据基础制度更好发挥数据要素作用的意见[EB/OL]. https://www.gov.cn/zhengce/2022 – 12/19/content_5732695.htm. 2022 – 12 – 19/2025 – 04 – 22. Chinese Government Website. Opinions of the Central Committee of The State Council of the People's Republic of China on Building a Data Foundation System to Better Leverage the Role of Data Elements[EB/OL]. https://www.gov.cn/zhengce/2022 – 12/19/content_5732695.htm. 2022 – 12 – 19/2025 – 04 – 22.

[11] 王东方. 生成式人工智能对个人信息权益的侵害风险及其法律规制[J]. 征信,2024,42(2):31 – 37. Wang D F. The risk of infringement of personal information rights by Generative Artificial Intelligence and its legal regulation[J]. Credit Reference,2024,42(2):31 – 37.

[12] 陈潭,刘璇. 智能政务 ChatGPT 化的前景与隐忧[J]. 电子政务,2023(4):36 – 44. Chen T,Liu X. The prospects and hidden concerns of ChatGPT – ization in intelligent government affairs[J]. E – Government,2023(4):36 – 44.

[13] 中国青年网. AI 可以更聪明,也可以更善良[EB/OL]. http://news.youth.cn/hotnews_41880/202501/t20250116_15775608.htm. 2025 – 01 – 16/2025 – 04 – 27. China Youth Network. AI can be smarter and kinder[EB/OL]. http://news.youth.cn/hotnews_41880/202501/t20250116_15775608.htm. 2025 – 01 – 16/2025 – 04 – 27.

[14] National Academies. How Al Is Shaping Scientifc Discovery[EB/OL]. https://www.nationalacademies.org/news/2023/11/how – ai – is – shaping – scientifc – discovery. 2023 – 11 – 06/2025 – 05 – 02.

[15] 赵晓伟,王小雨,王艺蓉,等. 人工智能赋能高校科学研究范式创新:价值、风险与进路[J]. 重庆高教研究,2025,13(1):9 – 20. Zhao X W,Wang X Y,Wang Y R,et al. Artificial intelligence empowering the innovation of scientific research paradigms in Universities:Value,risk and approach[J]. Chongqing Higher Education Research,2025,13(1):9 – 20.

[16] Zack T,Lehmen E,Suzgun M,et al. Assessing the potential of GPT – 4 to perpetuate racial and gender biases in health care:A model evaluation study[J]. The Lancet Digital Health,2024,6(1):12 – 22.

[17] Center for American Progress. Systematic Inequality and Economic Opportunity[EB/OL]. https://www.americanprogress.org/article/systematic – inequality – economic – opportunity/. 2019 – 08 – 07/2025 – 05 – 06.

[18] Nature. Al Will Transform Science – now Researchers must Tame it[EB/OL]. https://www.nature.com/articles/d41586 – 023 – 02988 – 6. 2023 – 09 – 27/2025 – 05 – 15.

[19] Meyer J G. ChatGPT and Large Language Models in Academia:Opportunities and challenges[J]. BioData Min,2023,16(1):20.

[20] Laura B. A Critique of hermeneutical injustice[J]. Proceedings of the Aristotelian Society,2011,111(3):479 – 486.

[21] Dotson K. A Cautionary tale: On limiting epistemic oppression[J]. Frontiers: A Journal of Women Studies, 2012,33(1):24 – 47.

[22] Kurtulmus F. The epistemic basic structure[J]. Journal of Applied Philosophy,2020,37(5):818 – 835.

[23] 中华人民共和国国家互联网信息办公室. 权威数读丨近 11 亿! 青少年和"银发族"成我国网民新增主力[EB/OL]. https://www. cac. gov. cn/2024 – 08/30/c_1726701400881428. htm. 2024 – 08 – 30/2025 – 05 – 15. Cyberspace Administration of China. Authoritative reading丨Nearly1. 1Billion! Teenagers and the "Silver Haired Generation" Have Become the New Main Force of Internet Users in China[EB/OL]. https://www. cac. gov. cn/2024 – 08/30/c_1726701400881428. htm. 2024 – 08 – 30/2025 – 05 – 15.

[24] 马茹. 人文社会科学评价体系构建的国际镜鉴与中国路径[J]. 扬州大学学报(人文社会科学版),2024,28(2):106 – 116. Ma R. International mirror and Chinese path to the construction of evaluation system for humanities and social sciences[J]. Journal of Yangzhou University (Humanities and Social Sciences Edition), 2024, 28 (2): 106 – 116.

[25] VDU. Turning FAIR into Reality: Final Report and Action Plan From the European Commission Expert Group on FAIR Data[EB/OL]. https://www. vdu. lt/cris/entities/publication/b0b91838 – e103 – 421f – b3fe – 274876ba51a7. 2018/2025 – 05 – 15.

[26] VDU. Turning FAIR into Reality: Final Report and Action Plan From the European Commission Expert Group on FAIR Data[EB/OL]. https://www. vdu. lt/cris/entities/publication/b0b91838 – e103 – 421f – b3fe – 274876ba51a7. 2018/2025 – 05 – 15.

[27] ARDC. Citation and Identifers[EB/OL]. https://ardc. edu. au/resource/citation – and – identifiers/. 2024 – 09 – 2/2025 – 05 – 15.

[28] 柯文涌. 模块副现象论与自由意志危机[J]. 自然辩证法研究,2017,33(4):22 – 28. Ke W Y. Module sub – phenomenon theory and the crisis of free will[J]. Studies in Dialectics of Nature,2017,33(4):22 – 28.

[29] Messeri L, Crockett M J. Artificial Intelligence and illusions of understanding in scientific research[J]. Nature, 2024,627(8002):49 – 58.

[30] Grimald G, Ehrler B, AI, et al. Machines are about to change scientific publishing forever[J]. ACS Energy Lett,2023,8(1):878 – 880.

[31] 白惠仁. 作为"基本善"的知识及其正义问题[J]. 哲学分析,2020,11(4):17 – 30,196. Bai H R. Knowledge as a "Fundamental Good" and its justice issues [J]. Philosophical Analysis, 2020, 11 (4): 17 – 30,196.

[32] Lazer D, Kennedy R, et al. The parable of google flu: Traps in Big Data analysis[J]. Science, 2014, 343 (6176):1203 – 1205.

[33] 李二斌. ChatGPT 对学术期刊影响如何[N]. 中国社会科学报,2023 – 02 – 28. Li E B. How ChatGPT affects academic journals [N]. China Social Sciences Journal, 2023 – 02 – 28.

[34] 张九庆. 生成式人工智能撬动科研范式重塑变革[J]. 今日科技,2023(5):3. Zhang J Q. Generative Artificial Intelligence drives the reshaping and transformation of scientific research paradigms [J]. Today's Science and Technology,2023(5):3.

[35] Mike P. Academic integrity considerations of AI large language models in the Post – Pandemic Era: ChatGPT and beyond [J]. Journal of University Teaching and Learning Practice,2023(2):7 – 24.

[36] 田贤鹏,肖智琦. 生成式 AI 赋能研究生科研写作的学术伦理与风险防控[J]. 现代教育技术,2024,34(8):23 – 32. Tian X P, Xiao Z Q. Academic ethics and risk prevention of Generative AI empowering graduate research writing [J]. Modern Educational Technology, 2024,34(8): 23 – 32.

[37] 张凌寒. 深度合成治理的逻辑更新与体系迭代——ChatGPT 等生成型人工智能治理的中国路径[J]. 法律科学(西北政法大学学报),2023(3):48. Zhang L H. Logic Update and system iteration of deep synthesis governance: The Chinese path of Generative Artificial Intelligence governance such as ChatGPT [J]. Legal Science (Journal of Northwest University of Political Science and Law),2023 (3):48.

[38] 加里·史密斯. 错觉:AI 如何通过数据挖掘误导我们[M]. 北京:中信出版集团,2019:7. Gary S. Illusion: How AI Misleads s Through Data Mining [M]. Beijing: CITIC Publishing Group,2019:7.

[39] 刘宝存,易学瑾. 英国人工智能人才培养政策变迁研究[J]. 比较教育学报,2023(6):12 – 23. Liu B C, Yi X J. Research on the change of UK AI talent training Policy[J]. Journal of Comparative Education,2023(6): 12 – 23.

The community virtue of social science knowledge production supported by Generative AI

ZHANG Rui, SI Meng – han

(College of Military and Political Basic Education, National University of Defense Technology, Changsha 410073, China)

Abstract: Generative artificial intelligence technology has been embedded in various aspects of social science knowledge production research, not only reshaping the basic process of social science knowledge production, but also leading the intelligent research paradigm to become one of the mainstream research paradigms in the field of social science knowledge production. However, generative AI also faces challenges such as unclear data ownership, data bias, data illusion, and data arrogance in processing complex data, which erode the virtues of community publicness, impartiality, objectivity, and criticality in social science knowledge production. To enhance the efficiency and quality of generative AI in social science knowledge production, and improve the community morality of social science knowledge production, it is necessary to construct a "government enterprise academia" tripartite collaborative governance framework to achieve the ideal collaborative vision of "knowledge co creation, human – machine symbiosis".

Key words: Generative Artificial Intelligence; data; social science knowledge; community; moral

---

(上接第 689 页)

Subsequently, the study explicates how these perceived identity threats motivate employees toward specific impression management strategies, including downplaying the contributions of AI tools, extensively modifying AI – generated outcomes, and concealing detailed information about tool usage. These impression management behaviors are strategic responses aimed at safeguarding professional status, competence perception, and autonomy within organizational social evaluations. Importantly, the research highlights that these behaviors do not merely stem from identity threats but also reinforce the very perceptions of threat that initiated them, resulting in a reinforcing negative feedback loop.

Further, the analysis elucidates how employees´impression management behaviors directly affect the construction of AI legitimacy within organizations. Through deliberate distortion and concealment of information, employees´behaviors adversely influence colleagues´accurate understanding and evaluation of AI tools´cognitive legitimacy (perceived necessity and value), pragmatic legitimacy (actual utility and effectiveness in practice), and normative legitimacy (alignment with organizational culture and ethical standards). Consequently, these behaviors significantly obstruct the institutionalization and effective integration of AI technologies.

The reciprocal dynamics between AI legitimacy construction and employees´identity threat perceptions are thoroughly explored, emphasizing that insufficient legitimacy enhances employees´perceptions of professional vulnerability, thus exacerbating further impression management behaviors and creating a persistent negative feedback cycle.

In conclusion, this research contributes significantly to theoretical development by providing an integrative, interactive theoretical model that bridges the micro – level psychological and behavioral responses of employees with macro – level organizational legitimacy processes. By explicitly articulating how identity threats, impression management behaviors, and legitimacy construction dynamically interact, the study advances identity theory, impression management theory, and legitimacy theory, offering a more nuanced understanding of technology integration within organizational contexts. Practically, this study provides actionable strategies for organizational leaders to effectively manage the intricate psychological and behavioral challenges accompanying AI adoption, facilitating smoother institutionalization and long – term acceptance of transformative technologies.

Key words: Artificial Intelligence; organizational legitimacy; identity threat; impression management behaviors