

“向善”且“为善”：人工智能时代的 算法治理与社会科学的源头参与

周旅军 吕 鹏

(中华女子学院, 北京 100101; 中国社会科学院, 北京 100732)

摘 要 在人工智能成为新基建重要内容的时代, 面对与算法如影随行的社会风险问题, 如何规训算法使其“向善”, 又如何驾驭算法使能“为善”, 日益成为政府、社会和学术界所关注的重要议题。在讨论算法治理现有进展和现实困境的基础上, 需要立足于中国本土的具体情况, 从技术架构和社会环境的角度分析算法生态供给侧的社会化生产路径及其治理在人工智能时代的特殊性。在中国人工智能头部企业的行动研究案例说明, 社会科学在将计算的社会化生产和社会风险作为重要的研究内容时, 以源头参与、多方合作的方式推进算法治理能够降低人工智能可能带来的社会风险, 创造出更多的社会价值。

关键词 人工智能; 算法治理; 计算社会科学; 搜索引擎; 社会性别

DOI:10.16059/j.cnki.cn43-1008/c.2022.01.015

一、人工智能何以成为社会问题

人工智能作为算法中的一组技术, 赋予计算机感知、学习、推理及协助决策的能力, 使计算机能够以与人类相似的方式来解决问题^①。作为“后互联网时代”的技术代表, 其已成为新型基础设施建设的重要内容。然而, 并非所有群体都能享受到算法红利, 算法也不总是产生积极后果。近年来出现的平台企业和政府服务问题引发社会对人工智能的关注和争议。人工智能时代的算法本质上是在社会化生产中形成的权力体现, 正影响甚至型塑着新的社会结构和社会心态, 相应的算法治理 (Algorithm Governance) 对于推进社会治理体系和治理能力现代化而言是必然而紧迫的要求。但目前人工智能相关的政策、法规和倡导更多的是聚焦在如何促进快速发展技术本身或者重点讨论伦理原则。本研究讨论算法治理的现有进展和面临的实践困境, 进而分析算法生态供给侧的社会化生产路径及其治理在人工智能时代的特殊性。在中国人工智能头部企业的行动研究案例说明, 社会科学以源头参与、多方合作的方式推进算法治理, 有助于打造“向善”且“为善”的算法生态, 降低

作者简介: 周旅军, 男, 博士, 中华女子学院社会工作学院讲师; 吕鹏, 男, 中国社会科学院社会学研究所研究员、博士生导师。

基金项目: 国家社科基金项目“人工智能促进社会性别平等研究”(项目编号: 20BSH040)。

^①Microsoft, “The Future Computed: Artificial Intelligence and its role in society”, <https://blogs.microsoft.com/blog/2018/01/17/future-computed-artificial-intelligence-role-society/>, 最后访问时间: 2021 年 10 月 29 日。

人工智能作为基础社会设施可能带来的社会风险,创造出更多的社会价值。

二、算法治理的实践困境

从国内外研究来看,算法审计(Algorithm Auditing)是计算社会科学(Computational Social Sciences)的重要组成部分,它是算法治理的具体方式之一。算法审计的研究对象大多涉及平台企业,范围覆盖搜索引擎、电子商务、即时出行、在线地图、就业市场等。在研究设计上主要是基于计算的方法,通过自动化的数据收集手段,对平台算法可能存在的社会影响进行实证检验。由此发现的问题包括:在线租赁市场 Airbnb 中针对房东的种族歧视^①;搜索引擎中网页检索结果与政治党派性质间的关联^②;广告交易平台之间的用户信息共享行为^③;Uber 价格算法中的公平性和透明性问题^④;电子商务网站中的价格操纵和歧视^⑤;检索系统结果排名会造成对受保护群体的不利情形^⑥;在线招聘平台中的偏见^⑦等。

随着社会环境的变化、技术内涵的发展和法律条文的完善,算法审计遭遇多重挑战。算法透明性是进行代码审计的前提条件,但在现实中未必可行。尤其对于平台企业而言,对代码的公开可能是在为垃圾信息滥发者或黑客们大开方便之门,反而会影响正常功能的运转。而在人工智能时代,即使算法的代码可得,也存在过于复杂、难以理解的问题。而且,像推荐系统这样强烈依赖数据的算法,即使在某个时刻表现正常,也不等于说在与其他数据配合时不会出问题。特别是,在大数据时代,算法中公式体系所涵盖的变量可以在成千上万的量级中进行搜集整理。并且,它们不仅来自服务器的数据集,还受制于用户属性、行为历史数据等,数据类型也可以是结构化或非结构化的数据,根本无法只靠阅读静态的代码指令来判断其是否存在问题。

通过远程数据抓取的方式来获取平台服务的返回结果,比如得到搜索引擎对特定关键词的网页排名情况,是对平台企业进行算法审计的数据基础。但这就很有可能违背平台企业在用户协议中的规定,甚至是触犯国家在爬虫软件使用方面的法律条款。尤其是在研究与数据隐私相关的算法时,很可能会侵犯用户自身的数据隐私,造成学术伦理问题。研究者单靠一己之力进行外部审计会越来越遇到技术上、法律上和伦理上的困难。

算法治理关心的议题并不限于之前所列举的算法审计的研究主题,随着新技术的出现和发展,

①B. G. Edelman, M. Luca, "Digital Discrimination: The Case of Airbnb. com," *Harvard Business School NOM Unit Working Paper*, vol. 14(2014), p. 54.

②R. E. Robertson et al., "Auditing the Personalization and Composition of Politically - Related Search Engine Results Pages," *WWW'18: Proceedings of the 2018 World Wide Web Conference*, 2018, pp. 955 - 965.

③M. A. Bashir, C. Wilson, "Diffusion of User Tracking Data in the Online Advertising Ecosystem," *Proceedings on Privacy Enhancing Technologies*, vol. 2018, no. 4 (2018), pp. 85 - 103.

④L. Chen et al., "Peeking Beneath the Hood of Uber," *IMC'15: Proceedings of the 2015 Internet Measurement Conference*, 2015, pp. 495 - 508.

⑤A. Hannak et al., "Measuring Price Discrimination and Steering on E - commerce Web Sites," *IMC'14: Proceedings of the 2014 Conference on Internet Measurement Conference*, 2014, pp. 305 - 318.

⑥Piotr Sapiezynski et al., "Quantifying the Impact of User Attention on Fair Group Representation in Ranked Lists," *WWW'19: Companion Proceedings of The 2019 World Wide Web Conference*, 2019, pp. 553 - 562.

⑦T. Sühr et al., "Does Fair Ranking Improve Minority Outcomes? Understanding the Interplay of Human and Algorithmic Biases in Online Hiring," *AIES'21: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021, pp. 989 - 999.

更多的问题可能涌现出来。其中也有不少恶意算法能够轻松地为普通人所掌握和使用,这些都使算法治理变得更为困难。就算法治理的目标而言,它不仅要清除恶意算法,还要发现和解决算法生态中所有的潜在风险,包括数据操纵、社会偏误、删截数据、社会歧视、侵犯隐私和产权、滥用市场能力和影响认知能力等方面^①。但科技“向善”的表现不能只限于除恶、避恶/不作恶,还应包括“为善”,要在关注社会问题的基础上开发和利用算法来促进社会问题的解决。然而,目前却存在着研究与现实之间缺乏互动的状况。以国内讨论较多的算法中社会性别不平等问题为例,学者基于消费、传播、就业、金融、政务等场景,对短视频、招聘、借贷等工具运用过程中的算法性别歧视进行具体的考察和分析并提出治理思路。但是,鲜见有真正产生现实影响的研究,甚至只有在社会负面后果难以遮掩时才能看到企业的回应。

在算法治理中强调伦理原则,并不意味着算法治理的目标止步于考察平台企业能否宣称或认同“科技向善”的理念和原则,也不仅仅是拦截恶意算法或对算法中的公平性进行评估。就人工智能的算法治理而言,如何看待算法生态供给侧的社会性质和潜在的社会风险?当前阶段的算法治理存在哪些特殊性?如何让算法治理真正运转起来?回答这些问题,需要我们从技术架构和社会环境的角度来做算法生态供给侧的整体分析。尤其是有必要立足中国本土情况对成功实现行动干预的个案进行充分分析,以更好地理解算法治理有更多可行的方向。不然,只是止步于理论批判/想像的抽象,与算法生态供给侧之间缺乏沟通和理解的基础,无异于自说自话。

三、算法生态的社会化生产与治理特殊性

人工智能的发展从技术角度通常可分为三个阶段,即计算智能、感知智能和认知智能,当前正处于从感知智能向认知智能演进的阶段,而促使人工智能技术发挥作用的要素是数据/知识、算法和算力。其中,算法框架属于人工智能的根技术,是生态建立的基础和关键。但是,算法治理不能简单理解为只是对算法本身的治理。为完整理解人工智能生态供给侧的社会化生产过程和社会风险后果,给社会科学介入人工智能根技术的生产和发展明确方向,需要考察算法生态的从业者(包括开发者、数据标注人员等)在不同阶段与各要素之间的关联。我们从以下四个方面着手分析人工智能时代中算法治理的特殊性:以开放平台为中心的开发者群体;开源文化影响下的知识协作社区;依附性质的数据标注从业者群体;社会文化影响下的算法应用场景。

(一)开发者的社会来源因头部企业的人工智能开放平台的算法集中供给而更为多元,但在技术规范教育上存在缺失

人工智能头部企业都有提供人工智能开放平台,有兴趣从事人工智能应用开发的人士可以通过多种简便的方式来无偿或廉价使用平台提供的算法和算力。仅以百度 AI 开放平台提供的数据为例,它目前已开放 335 项技术能力,360 万开发者正在使用这些平台开放的功能^②。我们从中不难想像到开发者规模之大,人工智能应用场景之丰富。从算法治理角度来看,借助海量数据、大规模知识图谱、超强算力和高精度算法带来的优势,人工智能平台的集中供给行为体现出治理方面的特殊性。这在当前存在两方面问题。

一是由于开发者只处于算法使用的位置,无法去了解和接触技术底层细节,因此也难以作出有效的反馈和修改,平台企业处于智识上的垄断地位。举例来说,在文本审核部分,尽管提供识别色情、暴恐涉政、恶意推广、辱骂、违反广告法、涉及违禁品等文本内容的功能,但对性别歧视等社会问

^①Danilo Doneda, Virgilio Almeida, “What Is Algorithm Governance?,” *IEEE Internet Computing*, vol. 20, no. 4 (2016), pp. 60–63.

^②百度:《百度 AI 开放平台》, <https://ai.baidu.com/>, 最后访问时间:2021 年 10 月 29 日。

题的识别却难如人意。经测试即使是最直接的歧视女性的话语也能通过当前的文本审核。这意味着平台在集中供给算法时也在依据企业自身的理解定义社会秩序的维护,在缺乏“守门人”的情境中,对社会后果的影响可能是一损俱损。

二是人工智能头部企业出于占领未来市场的目的为 AI 普及教育投入巨大的成本,但在给庞大的开发者(企业或个人)群体技术赋能的同时缺乏相应的社会规范教育。在技术文档、操作界面、教学专栏中并未给出与算法伦理和算法公平相关的教程和工具。比如,在人脸识别方面,多家厂商推出口罩检测算法助力疫情防控。然而,在人脸识别应用的伦理问题及其技术应对上却缺乏充分认识和行业规范。

(二) 开源共享社区对算法生态建设的贡献正转换为平台企业的发展红利,是算法、数据和知识社会化生产的重要途径,但治理上存在复杂特殊性

目前,主流的算法框架包括 Google 的 TensorFlow、Caffe、PyTorch、Keras 等,国内则有百度的 PaddlePaddle。它们的共同点不仅在于免费,还在于开源(open source)。这意味着任何愿意了解底层细节甚至想提交改变的开发者都可以参与到算法的社会化生产过程。对于平台企业而言,开源项目能够无门槛地展现企业的优势力量,吸引到众多的优秀人才无偿贡献力量,并能从大规模的开发者群体中免费获得 bug 反馈,激发新设想,最终巩固算法的行业地位,形成平台企业的发展红利。

开源文化正成为人工智能领域中的发展趋势。从算法治理的角度来看,开源算法具有代码上的透明性,有利于评估其中是否有损害社会利益的部分存在。不过,遗憾的是,尽管算法框架中的工具组件、基础模型库和开发套件在不断迭代升级,但专门审计算法中各种社会问题的工具并未成为受人关注的部分。

此外,开源共享社区不仅供应算法,还提供数据和知识。开展相应治理的复杂特殊性在于,缺乏数据和知识的算法无法训练出可用的人工智能模型,因此只针对算法开展的算法治理其实并不能解决系统问题,而对数据和知识的治理则要着眼于两个维度。

一是数据和知识既可以来源于人工智能企业的自行生产,也可以采集自互联网上众多的贡献者。数据资源和知识图谱中存在的问题,由于其数量庞大、内容动态和缺乏问题意识,难以得到有效发现和纠正。尤其是作为知识图谱基础的百科类知识资源,更能体现现实社会中权力的不平等秩序。比如,2019年3月8日,联合国教科文组织发起#Wiki4Women 倡议活动,旨在通过创建、编辑和翻译维基百科上的女性简介和列表,促进在数字领域更好地代表有奉献精神的知识女性。这是对维基百科上不到五分之一的传记是关于女性,体现出数字空间中性别差距的反应。另外,围绕“PX”开展的百度百科“词条保卫战”也体现出知识建构中的社会博弈过程^①。

除此之外,算法治理的复杂性还在于,“知识”的形态并非总是明文可审查,它可以是从海量数据中提炼出的知识模式。比如,在对基于大规模 Google News 文本训练生成的 word2vec 模型进行分析后发现,如果我们询问“医生:男性::x:女性”,系统给出 x = 护士。如果输入“程序员:男性::x:女性”,回答是 x = 主妇^②。这意味着当用人单位在搜索“程序员”时,返回结果会优先显示男性。因为“程序员”与男性的关系,比跟女性的关系更强。显然,如果不对此类“知识”加以后期纠偏,而直接用于搜索引擎、知识问答、机器翻译等人工智能应用中,只会重现并加剧社会中的不平等。但要对此类问题进行治理并不是靠技术对算法本身的监管可以实现的。

^① 百度百科:《词条保卫战》, <https://baike.baidu.com/item/词条保卫战>, 最后访问时间:2021年10月29日。

^② Tolga Bolukbasi et al., “Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings,” *NIPS’16: Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2016, pp. 4356–4364.

二是数据和知识即使来自互联网使用者,但在用于算法模型的过程中实际仍归平台企业控制。以百科知识为例,百度百科在用户协议中规定,百度可使用百科用户发表的内容,但其他用户只有在获得百度百科及其他相关权利人的授权之后才能使用百度百科站内的信息内容。也就是说,百度人工智能开放平台中提供的知识图谱相关能力,是百度公司根据用户协议合法地建基于众多开源共享社区参与者的贡献。如果其他开发者要使用大规模百科知识训练自己的算法,需要百度公司和其他所有相关用户的授权。不难想像,这在实践中较难实行。因此,建设和运营百科知识的平台企业在为社会生产和共享知识提供方便的同时,也可能最终实现了对社会知识的使用垄断权,相应地建立起认知智能方面算法的技术垄断地位。

(三)数据标注是保障深度学习模型质量的重要基础,但从业人员的社会来源与工作处境不利于实现算法治理

将算法按功能分类,可以相对简化地归纳为:视觉智能(如人脸识别、文字识别、视觉生产等)、语音交互(如语音识别、语音合成等)、自然语言处理(如情感倾向分析、评论观点抽取、机器翻译等)、内容安全(如图像审核、文本审核、视频内容审核等)、AR与VR(如虚拟主播等)、知识图谱(如实体标注、事件属性抽取等)、数据智能(如推荐算法等)。

其中,绝大部分功能属于感知智能(如视觉智能、语音交互等),少部分属于认知智能(如知识图谱、自然语言处理等)。如前文所述,知识图谱相关的数据和知识通常能够为公众所编辑和审校,加之有参考资料作为基础,其质量不能说一定有保证,但至少能公开回溯。但在感知智能相关的数据标注方面,情况却不太一样。“有多少人工,就有多少智能”这句话里的“人工”其实就是指算法训练数据进行建模之前的数据标注过程。该过程通常由人工完成,属于人工智能领域“入门级”的工种。单从工作流程看,其技术含量较低,吸引了众多文化水平较低、社会阅历单一的群体。在我国河南、河北、贵州、山西等地的城市还出现了一些特色的“数据标注村”。不止中国,向劳动力更充足、成本更低的地方迁移也是全球数据标注行业的发展趋势。

一方面,人工智能应用的场景丰富和技术发展对数据标注的质量要求不断提高,如自动驾驶、运动图像等领域的数据标注难度很高,但更大的挑战是在面对抽象的社会概念如颜值、善恶时,数据标注员应如何操作的问题。也正是在这一环节中,对社会不平等秩序和规范的内化会反映在标注结果之中。比如,对性别的判断,如果在标注阶段由于标注员对女性持有传统刻板印象,那么很可能将打扮偏中性的女性标注为男性,造成后期算法识别结果的错误;另一方面,由于数据标注众包平台的涌现,作为“幽灵工人”(ghost-workers)的零工在平台算法的剥削下无谈判的砝码和地位,基本权益难以受到保障^①,为使人工智能整个产业链能够可持续健康发展,算法治理也需要对此情况加以重视和应对。

人工智能算法的审计与以往世代的“白盒”代码测试过程不同。它往往是处于“黑箱”状态,加上深度学习方法中对数据标注和神经网络调参带有复杂的人为主观因素,使得最终训练出的人工智能模型与包括数据提供者、数据标注员、算法生产者和程序开发工程师在内的各方参与主体的主观偏见无法脱离。这也使得对算法治理的讨论必然要延展到对算法生态治理的讨论。考虑到数据标注的重要作用,算法治理的对象不仅包括管物的算法,也要包括管人的算法。

(四)目前人工智能的应用场景多集中于经济价值高的产业,少有社会问题方面如发现性别不平等、促进性别平等的应对方案

显然,在关切的问题域和服务对象上的区别,也能反映出算法生态供给侧的社会性质。“向善”

^①M. Graham et al., “Digital labour and development: impacts of global digital labour platforms and the gig economy on worker livelihoods,” *Transfer* (2017), pp. 135 – 162.

的科技不应该只是保证算法的代码不出技术上的 bug,或者只是规避恶意的动机,或者只是作为企业对外建立公关形象的临时道具,而应该积极地将应用场景在社会问题领域的扩展作为 AI 伦理的落地方式。有研究在系统评估 22 份人工智能伦理准则后发现,几乎没有准则在照顾、培养、帮助、福利、社会责任或生态网络的背景下谈论人工智能。技术产品主要被视为孤立的实体,可以由专家进行优化,以便为技术问题找到技术解决方案。与女性为主要领导的组织相比,对人工智能算法的认识通常缺乏的是对更广泛的背景和技术系统所嵌入的全方位关系网络的考虑^①。

正如艾斯勒在其“权—性—钱”三部曲中所指出的那样,“在当代,当高科技被征服、剥削和统治等价值观所引导而威胁到我们的生存时,我们需要由关怀的精神所推动的经济发明。我们需要一场关怀革命”^②。如前所述,作为新型基础设施的人工智能,包括数据/知识、算法和算力,无不反映着社会秩序和价值观。要从根本上解决算法生态中存在的诸多问题,需要社会文化的转型,尤其是从统治关系走向一种伙伴关系的未来^③。对于统治关系与伙伴关系,艾斯勒从两性关系、暴力、社会结构、性特征、灵性、快乐与痛苦、力量与爱等七个维度阐释二者之间互相消长又互相印证的基本差别^④。而当前社会中存在的算法盲视、滥用、偏见与歧视等社会风险与它们有着千丝万缕的关系。显然,这远不是依靠科技企业在技术架构中就能解决的问题。

四、社会科学的源头参与

社会文化的影响、技术环境的日益复杂、开发者来源的多元化、数据与知识的众包性质、平台治理的力量不足,给人工智能时代的算法治理带来与以往世代不同的困境。当面对技术可能带来和正在产生的社会风险,计算社会科学不能只是以“批判”的立场和态度“置身事外”,还要能够通过“建设”的源头参与来规训和驾驭算法,使其向善,使能为善。

计算社会科学既要在外部作为独立的社会力量评估算法的社会后果,揭示算法的社会风险,又要在内部成为算法生态的重要主体,凭借系统的社会科学专业知识,以提供咨询、合作开发、多方博弈等方式来引导算法厂商。这样可以使缺乏可持续发展理念的厂商在意识上能够选择符合社会期待的发展战略,实现价值观上的“向善”;使仅有“向善”价值观却不知如何在自身体系中落地的厂商能够有能力来主动地评估当前的算法生态。

在此以一项审计搜索引擎算法中社会性别歧视的行动研究为例,阐述社会科学在算法治理中的源头参与对于人工智能平台治理的影响和作用。根据本研究对国内各家搜索引擎(百度、搜狗、360 搜索等)的考察^⑤,充斥于网络话语中的性别偏见和女性权利并未受到人工智能驱动下的搜索引擎平台服务商的充分关注,改进“为善”更无从谈起。具体而言,搜索引擎作为人工智能广泛而影响

① Thilo Hagendorff, “The Ethics of AI Ethics: An Evaluation of Guidelines,” *Minds and Machines*, vol. 30, no. 1 (2020), pp. 99–120; Inga Strümke et al., “The social dilemma in AI development and why we have to solve it,” *arXiv preprint arXiv:2107.12977* (2021).

② [美]理安·艾斯勒:《国家的真正财富:创建关怀经济学》,社会科学文献出版社 2009 年版,第 19 页。

③ [美]理安·艾斯勒:《圣杯与剑:我们的历史,我们的未来》,社会科学文献出版社 2009 年版,第 234 页。

④ [美]理安·艾斯勒:《神圣的欢爱:性、神话与女性肉体的政治学》,社会科学文献出版社 2009 年版,第 410—411 页。

⑤ 根据 Statcounter 发布的搜索引擎中国市场统计数据,在 2021 年 7 月,国内厂商中的百度占 73.65% 的份额,搜狗占 19.13%,360 搜索占 1.11%。这与我们在 2019 年进行算法审计时的排名情况相同,比例差异不大。

深远的应用,普遍存在与社会性别相关的三个方面问题^①。

一是在搜索框中输入“女”“女人”“女性”等关键词时,查询到的文字和图片存在污名化女性的现象、消费女性的结果,体现出显性歧视。

二是在搜索“医生”“科学家”等词时,出现的图片结果只是在反映现实中的职业隔离,体现出隐性歧视。比如,当搜索“程序员”时,不仅首页图片中鲜有女性程序员,而且在“相关搜索”中也无“女程序员”的搜索联想词提示。

三是在查询“家暴”“虐待妇女”等检索词或相关语句时,厂商未能充分运用人工智能的技术手段为可能的遭受家暴或者其他不公平对待的幸存者提供必要的帮助和支持资源,只是将它们作为普通的查询字词对待,无视其所代表的社会问题和社会意义,也就无所作为,甚至不加甄别地推送戏谑化或者有悖公序良俗的搜索结果。

当厂商借技术的“客观”“中立”镜像社会现实而忽略社会问题、仅仅是重演甚而加深现实社会中的不平等时,人工智能的算法就不会是“为善”的技术,其社会后果必会有悖厂商自身倡导的“AI伦理”。如果要改变上述状况,首先就要进行人工智能算法中的性别平等审计,有针对性地开展干预,避免因技术的误用导致人们在生活、工作各个方面变得越来越不平等。

为此,我们曾致信百度公司反映相关问题,在得到积极回应后与相关部门合作对搜索引擎的部分算法实施干预。在初步的合作中实现对搜索结果中体现性别歧视部分的削弱、屏蔽和消除,同时优化和增强为女性维权提供支持的功能渠道。从长期效果来看,干预的作用明显。时隔两年,对比干预后的百度和未干预的其他搜索引擎,仍能发现差异显著:根据百度的搜索大数据,发现“被老公打了”“爸爸打妈妈”“遇到家暴怎么办”等词条是用户检索的高频词。为了尽可能地帮助被家暴的女性群体,百度设置了包括“家暴”在内的11个关键词条,并与“反家暴,你要大声说出来!”的专题卡片相关联。只要用户搜索相关关键词,就能够得到家暴判定、咨询、法律帮助等信息。

虽然百度早在2018年即已表示“AI的创新愿景是促进人类更平等地获取技术和能力”,但只有在社会科学的源头参与之后,以专家在社会性别领域的专业能力和创新思考,才能推动科技“向善”的意愿在特定社会问题领域中的实现,使人工智能的伦理倡议逐渐真正落地为科技“为善”的现实。同样地,正是在行动研究的过程中,社会科学才更能了解算法生态的社会化生产和运作过程,并在对源头参与、多方合作/博弈的反思中加深研究者对学科理论的认识,确定进一步研究和行动方向。

五、结论与建议

就算法治理而言,不管是国家层面,还是在市场和社会中,似乎在道义上存在着共识,即“科技向善”。然而近年来在人工智能领域中出现的种种乱象,包括伦理原则提出者在实践中的背道而行,都在不断提醒着我们,算法“向善”且“为善”之路漫漫其修远兮!通过理论上的分析和行动研究中的观察,我们认为在人工智能时代的算法治理中,社会科学的源头参与必不可少。

一是社会科学家有必要充分认识和揭示算法治理中的价值观立场。如何推动算法的服务目标和生产过程更为多元化,使其更有利于伙伴关系而非统治关系的建立,需要在理论和实践方面充分对话,纳入不同的声音。作为权力的算法在其社会化生产的过程中受到多方力量的影响,即使是最简单的分类法也反映着社会意义上的范畴争夺和身份认同。在社会科学的源头参与中,要警惕知

^①此处所描述问题的内容类别和严重程度多为2019年4月干预前的情况,但在未干预的搜索引擎中,有些情况即使在今天也未有所改善。为排除搜索引擎是基于研究者个人偏好或其他如IP地址等信息给予个性化的搜索结果,干预前后的观察都是在云端虚拟主机的Chrome浏览器的隐身模式(incognito)中未登录任何厂商账号时进行。

识精英、权力精英与经济精英的结盟,避免产生更大的社会风险。

二是在知识上多总结和引介国内外已有的算法问题和治理方案,督促相关企业或机构对其存在的的社会不平等问题予以纠正和改善,是社会科学源头参与的重要路径。尤其是推动人工智能的头部企业重视并参与建设行业级别的算法治理规范(而不仅仅是伦理原则的宣称),以确保社会计算不会对特定人群带来风险和伤害。最终借助平台的超大规模流量对开发者群体进行普及教育,从而在源头上保证未来人工智能应用的社会性和合规性。

三是社会科学家的作用更多体现在对社会问题的分析,为技术的落地提供知识基础而不仅仅满足于理论批判。比如,在社会性别问题领域方面,应当积累和标注“性别偏见”语料库,推动“性别偏见”成为用户画像的属性变量。这有利于针对性推送消弭性别偏见的信息,减弱“信息茧房”中的回音室效应,以正面的传播引导价值观的改变和型塑。但是,究竟如何定义和解析“性别偏见”,使其能在技术上具备相应的数据基础和正确的知识图谱,这更多是社会科学家的能力和责任所在。

四是算法生态的复杂性决定社会科学的源头参与只能是以多方合作的方式进行。在我们的行动研究中,尽管百度决定把“性别歧视”等女性相关信息的持续优化工作提上日常信息运维工作的日程,并且在思考向更广阔的社会议题扩展,但是,由于相关合作部门缺乏足够的力量协调和影响其他执行部门,因此,需要进一步的顶层设计和多方合作才能够更深入地推进算法治理。

五是社会科学家有必要加强对算法生态的理解,甚至参与到算法的社会化生产过程中去。在当前的人工智能时代,算法并非全是闭源的产品,国家和业界在倡导和生产并大规模使用开源的软件和硬件。计算社会科学家不仅可以更彻底地审查和批判现有算法中存在的社会风险,还可以参与到算法的迭代升级中,甚至以算法来人工智能地监测和评估算法的社会后果,并推动审计性质的算法组件成为产业应用的基础成分。也就是说,社会科学家通过参与算法的生产过程可以有效地影响人工智能这一基础社会设施,从而在未来数字社会中更好地履行社会科学的学科责任。

(责任编辑:何绍辉)