doi: 10.15936/j.cnki.1008-3758.2019.02.002

算法的伦理问题及其解决进路

刘 培, 池忠军
(中国矿业大学 马克思主义学院, 江苏 徐州 221116)

摘 要: 算法作为强大的参与者介入甚至主导人类社会各领域。然而,日益复杂且自主的算法引发了透明性、公平性、算法歧视、自主性、隐私安全以及可责性等诸多伦理问题。从解决算法伦理问题的内部进路看,至少有算法的伦理设计和应用范围限定两种方式;从解决算法伦理问题的外部进路看,应该注重设计者的伦理责任与多领域合作、鼓励公众参与算法设计以及对算法进行监管。只有采取切实有效的措施解决算法的伦理问题,才能走向公平、透明、负责任的算法。

关 键 词: 算法歧视; 自主性; 可责性; 透明性; 伦理责任
中图分类号: N 031 **文献标志码**: A **文章编号**: 1008-3758(2019)02-0118-08

Ethical Issues of Algorithms and Their Solutions

LIU Pei, CHI Zhong-jun
(School of Marxism, China University of Mining and Technology, Xuzhou 221116, China)

Abstract: As a powerful actor, algorithms intervene and even dominate in a wide variety of domains. However, increasingly complex and autonomous algorithms pose such ethical issues as transparency, fairness, algorithm discrimination, autonomy, privacy security and accountability. From the internalist approach, there are at least two solutions: algorithm ethical design and application limits. From the externalist approach, designers' ethical responsibilities and multi-disciplinary cooperation should be emphasized, and the public should be encouraged to participate in algorithm design and algorithm supervision. Only by taking effective measures to solve the ethical issues of algorithms can fair, transparent and accountable algorithms be obtained.

Key words: algorithm discrimination; autonomy; accountability; transparency; ethical responsibity

大数据与人工智能的结合使算法"润物细无声"地渗透到我们的生活世界。日常生活中越来越多的方面,诸如游玩、工作、家务、通信等活动是由数字设备和计算系统进行干预、调节和管理。这些数字设备和计算系统都是以大数据为驱动、以算法为核心的。更准确地说,"大数据时代干预生活世界的算法能够解决大量复杂的任务,包括执行搜索、安全加密、优先推荐、模式识别、预测、分析、仿真和优化等,算法已然成为社会新的权力中间人"[1]。基于此,耶鲁大学学者杰克·M.巴尔金(Jack M. Balkin)认为算法社会正在到来[2],每个人都主动或被动地沉浸在算法生活之中。我国也是如此:一方面,大数据与人工智能的发展已上升到国家战略层面。国务院于2015年发布《促进大数据发展行动纲要》,2017年发布《新一代人工智能发展规划》。另一方面,我国已成为大数据与人工智能发展大国。我国在数据产生与收集、互联网公司的技术创新以及消费领域

收稿日期: 2018-05-19
基金项目: 国家社会科学基金资助项目(14BKS039)。
作者简介: 刘 培(1986-),女,河南孟州人,中国矿业大学博士研究生,主要从事大数据与公共治理研究; 池忠军(1963-),男,内蒙古赤峰人,中国矿业大学教授,博士生导师,主要从事公共哲学与公共治理研究。

大数据的创新应用方面拥有不俗表现，且人工智能的企业数量与人才拥有量位居全球第二，投资融资规模占全球 60%[3]。在此境况下，我们需要追问：对算法社会中每个人来说，算法生活带来的仅仅是好处吗？全球权威调查机构皮尤研究中心邀请技术专家、学者、政府官员等各界人士深入探讨算法可能产生的各种影响，并最终于 2017 年 2 月发布了名为《依赖代码：算法时代的利与弊》的报告。该报告指出：一方面，算法能帮助我们理解大量数据，这将会带来诸多可见的与不可见的益处，如激发科学突破、提升人们的创造力和自我表达能力、创造新的便捷生活方式等；另一方面，算法的这些益处也伴随着诸多挑战，如过于追求数据和建模而消弱人类的判断、算法偏见的产生、社会分歧的加深等各种不良后果[4]。然而，认识到可能存在哪些负面影响还远远不够，算法歧视、人类判断的消弱等后果实质上更深地涉及到社会伦理问题，因而对算法伦理问题的讨论也成为当前算法技术发展进程的一部分。例如算法系统作为一种黑箱缺少透明性，公众未获得关于算法权力运作的充分知情权；算法驱动的决策系统可能导致潜在的社会歧视从而强化不平等。面对日益强大的算法，厘清其社会伦理问题，并提出可能的解决策略才能让算法真正符合善法。

一、算法的界定与厘清

从词源角度而言，算法（algorithm）一词源于拉丁语“algorism”，意为“花拉子米提出的运算法则”。此后，算法概念进一步扩展为执行基本算术的分步法。直到 20 世纪中叶，随着科学计算和高级编程语言的发展，算法被定义为解决问题的相应清晰步骤，即只要按照此步骤输入指令或数据就会产生预期的结果。智能时代算法悄然介入甚至主导人类事务，算法从计算机领域的一个专业术语转变为智能时代文化与社会的关键概念，对此概念的理解和界定也日益多元化。按照学者塔尔顿·吉莱斯佩（Tarleton Gillespie）的观点，可以从以下几个层面界定算法，主要包括作为技术解决方案的算法、作为社会技术集合的算法以及作为权力的算法[5]。

首先，算法是技术性的解决方案。在计算机领域，算法是纯技术性的，其经典界定为：算法＝逻辑＋控制。逻辑条件是指与所要解决问题相关的知识，控制是指解决问题的策略以及在不同情境下处理逻辑的指示。其中，有关算法应用的情境、所产生的后果等内容被严格限定在算法的概念框架之外。换言之，作为纯粹推理形式的算法只关注严格的理性推理，即算法的功效，是数学确定性与技术客观性的结合体[1]。然而，这里存在一个关键的问题，即如果仅将算法等同于算法功效，那么算法“能做的”就是其“应做的”吗？如算法在没有种族、性别等敏感数据输入的基础上也能基于其他数据之间的细微关联而推断出人的性别与种族，进而基于此进行评分并作出决策，但这种行为显然是不符合伦理与法律的要求。

其次，算法是社会技术集合的简称。这是对算法的借代性使用，即以算法来指代与其开发、实施相关的一切人和物，包括众多算法设计者、算法逻辑、训练数据集合、实现算法的硬件、算法使用者、算法应用的情境及其所产生的社会影响与后果、相关的法律与标准等。在此意义上，算法不是独立的技术工具，而是庞大的社会技术网络[5]。进一步而言，可将算法视为“物的集合”。正如迈克·安妮（Mike Ananny）借用拉图尔行动者网络中的“集合”概念，将算法界定为“由规定的计算机编码、人类实践和规范逻辑构成的物的集合。其中，规范逻辑通过最低限度的、可观察的、半自主的行动，创造、维持、表征人和数据的关系。尽管编码、人类实践和规范可以在其他背景中被单独观察到，但算法全部的意义和力量只能在与集合中其他行动者的关系中被理解。”[6] 如此一来，算法就不再被置于人的对立面，算法所引发的伦理问题也超越了纯粹的技术性维度，而是关涉到由编码、规范、人类实践等算法的每个部分所构成的整个集合如何行动。进一步而言，作为社会技术集合的算法只有在其手段、目的和美德以某种方式组合时才是符合伦理的[6]，如算法代码透明、设计者的良善意图、操作制度的规范健全。这既为思考算法伦理问题何以可能提供了视角，也为如何有效治理算法的伦理问题指明了方向。

再次，作为权力的算法。此视角认为算法所指涉的并不是其技术性和物质性特征，而是当算法在人类社会实践中“实现”之时所具有的干涉甚至主导人类社会事务的能力，即算法权力（algorithmic power）。诸多学者都对算法权力的强大进行了描述：“越来越多的权力存在于算法之中，它不仅塑造社会与文化，直接影响与控制个人

生活,甚至获得了真理的地位。"[7]与此同时,如此强大的算法权力却是不透明的、不可理解的:我们无法知道它是如何对个体进行排序与归类,也不知道它如何从既定的数据输入得出某种输出结果,甚至伴随着其自主性的增强,即使知道输入与输出也无法理解它的运行原理。这也就造成了即使算法可能会带来某些负面影响,我们也无力对算法权力进行质疑与规制。

综上所述,以上三种理解方式的敞现为了解不同语境下的算法提供了基础。尽管算法概念存在争议和模糊性,却是一个值得深思的问题。对此,美国康奈尔大学科学与技术研究系学者马尔特·泽维茨(Malte Ziewitz)认为:可将算法概念视为一种敏化概念(sensitizing concept),即虽然缺乏精准的定义,却有助于重新思考我们对算法的认知,并为理解和解决算法潜在的透明性、自主性、规范性等社会伦理问题提供了线索和启示[8]。

二、算法伦理何以可能

首先,作为技术手段的算法内在地关涉伦理问题。在技术中介论视角下,算法不只是被动的工具,"它本身蕴含着积极的、能动的、独立的'意向性',或者说它也是'行动者网络'中的'行动者',具有一定的能动性"[9]。算法的能动性体现在算法的优先排序、标准分类、关联标记与过滤。具体而言,"第一,算法内在涉及优先排序。每一种寻求优先排序的算法都嵌入了度量标准,这些标准都前置性地植入了一系列价值选择与主张;第二,算法内在的分类过程能够决定特定实体的归属类别,分类过程则是通过设置固定阈值或复杂的聚类分析而进行界定,因此分类算法可能具有偏差导致出错,进而引起潜在的社会危害;第三,算法的关联标记功能涉及标记实体之间的关系,但是算法关联同样是建立在关联的标准之上,在此过程中需要引入许多相似度函数,从而根据给定的关联定义对比不同事物的匹配性,当达到特定阈值之际,事物之间就被认为存在关联;第四,过滤则是依据各种规则、标准包含或排除信息。事实上,过滤通常会根据上述优先排序、分类及关联决定显示并排除哪些信息"[10]。

算法内在地负载着设计者的价值,即算法的能动性或意向性是通过算法设计者的基本价值判断而实现的。在设计算法时,技术专业人员虽然尽可能地保持严谨的客观性,但他们永远无法逃避所置身的社会文化、道德准则与知识背景,即算法本身在创建时就蕴含了大量的专业知识、价值判断、道德选择和约束条件[1]。例如,算法会给出诸多可能的结果并显示最优选择与决策,这不仅需要技术专家代替作为使用者的公众设定某个阈值以达到使用者的满意,而且为了有效地实现预期目标,技术专家会依据自身的判断或过去用户的惯例选择数据以训练算法,以便它可以"学习"将查询和结果配对并找到最优结果[5]。换言之,这些训练数据的选择本身就体现着设计者的价值观及其所处的文化、知识与社会背景的影响。对此,荷兰技术哲学学者菲利斯塔斯·克雷默(Felicitas Kraemer)认为,"算法本身就包含着基本的价值判断,因而在本质上是负载价值的从而具有基本的伦理维度"[11]。

其次,算法应用后果的不确定性导致伦理问题。算法应用后果的不确定性来自于算法本体的不确定性——算法本身内在地具有不能被规避的不确定性。"不同于普通算法按照指令从数据输入到既定结果输出这一过程,机器学习算法则颠倒这一过程,它们被编程去学习如何解决问题。因而,对于机器学习算法,即使我们可以观察它的输入和输出,却不能解释其原理。"[12]更为重要的是学习能力赋予算法某种程度的自主权,这种自主性在某种程度上必然使算法输出结果难以预测与解释。大多数算法的指令相对简单,结果也相对确定。如果出现问题,程序员可以返回程序的指令,找出错误发生的原因并纠正错误。但具有自主性的算法却由于系统过于复杂,增加了预测与解释其输出的难度,甚至即使我们能够充分描述它们的工作原理,它们实施解决方案的实际机制可能仍然不透明,因而仍无法预测与解释其应用后果。同时,由于人类与机器学习算法之间并不存在天然的、"由此及彼"的理解能力,所以人类在预测与解释机器学习算法时天然地处于劣势[12]。当算法被应用于自动驾驶汽车、协助医学诊断与手术、社会治理等场景时,一旦算法失败或出错,将会给个人和社会带来重大风险,从而引发社会伦理问题与争议。

三、算法的伦理问题

虽然算法可以提升决策效率和公平性,但也

会导致诸如算法歧视、主体的自主性、隐私以及可责性等伦理问题。

1. 算法歧视

虽然算法可以每秒执行数百万次操作进而解决大量复杂任务，且最大限度地减少人为错误与偏见，但并不意味着算法决策就是公平的。其不公平集中体现在算法歧视。算法歧视是指数据驱动的算法决策会导致歧视的做法和后果，换言之，算法决策可以再现已有的社会歧视模式，继承先前决策者的偏见从而加深流行的不平等[13]。例如惠普公司曾经设计一个基于特征识别的人脸定位算法。但是此算法是一种依赖于白色皮肤等特征的识别模式从而无法鉴别黑色皮肤。进一步而言，其用于定位人脸的算法代表了不同的权衡。甚至我们可以说，算法的选择代表了定义“什么是脸”的能力，从而内在地再现种族主义的观念。按照莱普利·普鲁诺(Lepri Bruno)的观点，算法歧视可以分为三种类别[13]。首先，预先存在的偏见所导致的算法歧视。此类歧视通常在创建算法系统之前就已存在，算法只是将其反馈出来并转化为行为即歧视。换言之，先前存在的偏见可以通过个人或机构有意识地或无意识地进入算法系统：算法设计者和参与者可以通过设计而将个人偏见嵌入算法中；即使算法设计者没有个人偏见，“喂养”算法的数据也并非客观公正的，它们从现实世界中抽取因而必然携带着社会、文化和价值已有的偏见，从而导致算法的结果也将体现甚至放大不合理的歧视。即使算法开发人员没有歧视意图，且抑制所输入数据的敏感属性，即有关种族、政治倾向、宗教信仰等的数据，但经过良好训练的机器学习算法能从大量的数据集群中发现不太明显的相关性，而且由于数据集中且数量巨大，详尽地识别和排除与“敏感数据”相关的数据特征几乎不太可能[14]。其次，使用算法进行决策导致的算法歧视。在算法系统中，通过算法划分类别、优先化排序、关联性选择和过滤性排除被认为是一种直接的歧视，其涉及差别性地对待从而导致不公正[13]。进一步而言，算法系统会赋予现实主体一种新的身份——“算法身份”(algorithmic identity)[15]。不同于现实身份，算法身份是算法基于数据主体的数据足迹而推断出来的分类，其自动确定了个体的性别、阶级、种族等身份特征。再次，在算法决策中各类数据所设置的权重不同，从而有可能导致间接的歧视行为。例如，在警务预测中，所用算法中过分强调邮政编码的权重可能导致低收入的美国黑人社区与犯罪的关联更大。

2. 削弱主体的自主性

自主性即自我管理，是主体构建自我目标和价值且自由地作出决定付诸行动的能力[16]。然而，算法系统越来越多地代替我们作出决定。“在大数据智能时代，越来越多的重要决定由算法裁定。从搜索引擎、在线评论系统到教育评估，市场运行、政治运动如何展开，甚至社会服务、公共安全管理等生活领域和公共政治领域都是由算法进行决策与管理。”[17]在此过程中，算法的自主性与主体的自主性呈现出此消彼长的态势。也就是说，随着算法自主性的增强，作为主体的人的自主性减弱，甚至出现主体隐匿。

具体而言，一方面，算法的个性化推荐、营销、推送等功能能够很好地迎合个体需求，但同时也会削弱和损害主体的自主性。信息多样性是主体自主决策的前提条件，而个性化所造成的信息茧房则会排除主体不感兴趣的信息而只推送相关信息，进而阻碍主体进行自主决策[18]。进一步而言，随着深度学习算法在个性化推荐、营销、推送等领域的发展，个性化不再只是依据主体行为数据的历史展开，而是可以超越基于数据的推荐而基于“内容本身”进行推荐，能够解决无历史行为数据用户的冷启动问题。这些无疑会使得算法的个性化功能更强大，但同时也可能将人的主体性减弱并进一步推进为控制甚至遮蔽主体。人可以轻松便捷地安享算法主动诊断问题并给出相应解决方案的能力，与此同时，为保证算法的这种自我决定的权力，我们不得不放弃部分“自主”的权力。另一方面，除主体自主性损害之外，算法的个性化推荐、营销、推送也涉及到隐含的经济与价值导向，从而导致对主体自主性的不尊重。进而言之，算法程序的个性化推荐不完全是依据主体行为数据将信息与主体的兴趣进行匹配，也包含着第三方的利益和价值[18]。例如，对在线消费者进行商品、服务和信息推送的算法系统不仅包含着满足消费者自身的偏好，也涉及到经济利益，即维护和扩大市场的动机，它决定谁的商品、服务或信息能够被优先排序从而推送给公众以供选择[19]。

3. 侵犯个人隐私与集体隐私

在计算机伦理学领域，个体隐私可以大致分为“数据隐私”和“身体—空间隐私”两类。“前者

是关于数据主体的数据信息,即通过数据信息能够识别主体的现实身份,是对主体自我及相关现象的表征;后者则指与公共相对应的私有空间,是对介入身体和空间限制的权利,如不受干扰或独处的权利。"[16]据此,算法对个人隐私的侵犯可以分为两种:一是在数据主体不知情的情况下,算法系统挖掘与收集个人数据。具体而言,在算法系统中,算法只有与数据结合起来才能运行。然而,在数据主体不知情的情况下,算法系统对其网上行为进行跟踪、收集和计算以更多地了解其用户,甚至获取用户所有的"数据轨迹与记忆"。为此,数据平台必须跟踪他们的用户,并通过相关的技术和服务鼓励用户持续地、频繁地使用其平台。在此过程中,用户可能不知道他们的网络行为正在被跟踪,即使他们知道,也没有能力挑战这种安排,除非用户能够回避所有的数据收集[20]。二是算法系统通过不透明的计算、剖析、解释与预测而介入或侵犯个人隐私。例如,"著名的脸书公司(Facebook)通过算法系统分析用户的数据信息,能够准确地预测男性用户的性取向、种族、宗教信仰、政治偏好以及对酒精、毒品与香烟的使用情况;而源自推特(Twitter)的算法系统通过数据分析能够识别那些在出现抑郁临床症状之前就可能陷入抑郁的人"[21]。

随着算法技术的发展,算法系统对隐私的影响从个人隐私逐渐转向了集体隐私(group privacy)。传统的集体概念通常是基于集体意识且由其成员和局外人所感知到的共性而形成;而在算法背景中,集体隐私指涉的集体则是由符合算法系统预先设定的目的、具有某一显著特性的成员聚集而成。集体既不是被发现的也不是被发明的,而是在抽象层次(level of abstraction)上被设计的,其逻辑顺序为目的—算法分组—集体[22]。因此,集体隐私不是"他们的"隐私的所有集合,而是基于集体成员一个或多个共同属性的"它"的隐私(from "their" to "its" privacy with regard to the group)[23]。具体而言,集体隐私不再是构成该群体的所有个体隐私的总和,而是构成集体的所有个体的某一种或几种共同的属性。即使在现有数据匿名化、限制交叉引用数据集等数据保护技术有效保护个人数据的情况下,集体隐私依然受到危害。正如荷兰蒂尔堡大学法律、技术与社会研究所学者林内特·泰勒(Linnet Taylor)指出:"算法分组的目的大多不是访问或识别个人,而是识别一个群体的偏好或特征,从而进行有效地干预。在此过程中包括算法设计者在内的每个人都可能不知道个体数据是否正在被误用或滥用,而现有的数据保护也并没有涉及到集体隐私保护问题。"[24]

4. 基于透明性的可责性

算法透明性是指算法相关信息的可访问性和可解释性,是问责制与责任分配的基础[13]。

由于算法日益复杂且具有自主性,导致算法系统成为不可观察、不可解释的技术黑箱,从而引发了算法可责性的伦理问题。在伦理学中,"责任"一词包含两层含义:一是主体或行动者有义务给出一个解释;二是客体处于可以被说明或解释的状态[25]。就算法责任而言,如果算法是确定且完全公开的,那么只需运行算法并检查输入与输出的匹配,主体就能对此作出解释。然而,由于算法所输入的数据集越来越大且运行也日益复杂,从而具有不易理解和自主性特征,由此出现了"责任盲区"。

在算法责任归属中,算法系统的日益复杂性和不透明性导致算法责任的分配与承担出现困境。对此,安德鲁·塔特(Andrew Tutt)从三个层面分析了算法责任困境[12]:第一,算法责任难以测量。例如,在功利主义算法的指引下自动驾驶汽车可能在某些情况下故意杀人(行人或乘客),以减少总体伤害。在此情境下,我们虽然已经对如何判断和解释传统交通事故中个人需要承担的责任形成了一致观点,但对于此类算法行为的责任测量却没有类似的共识。第二,算法危害难以追踪。即使算法在编程时严格按照法律与伦理规范,但很难预测算法是否会在任何给定的情形下按照既定的设计规范而行动。以算法歧视为例,如果招聘公司利用算法来筛选合适的求职者,那么算法最终可能会根据种族、性别等进行区分,进而造成算法歧视。然而追溯算法歧视的确定性来源则很困难,可能是程序员程序设计的失误、训练数据的不完整、数据中潜在的偏见,甚至根本就不存在算法歧视,其发生只是偶然性的低概率事件。第三,人的责任难以分配。算法以多种方式被"碎片化"。例如,算法由一个或多个公司的众多设计者共同完成,该算法可以被复制、修改、出售,甚至应用于其设计者从未想象过的各种情景中。因而,当算法产生危害时,很难明确其设计者、出售者、应用者、数据提供者等各个利益相关

者具体应承担多少责任。

四、解决算法伦理问题的两种进路

算法的伦理问题主要是由算法本身或人为因素造成，因而其解决进路也大致可分为内部和外部两种进路。

1. 内部进路

从内部进路看，主要有以下两种解决途径：

第一，算法的伦理设计。鉴于算法所引发的诸多伦理问题，基本的解决进路之一就是伦理设计，其实践探索主要包括两个路径：依据某种伦理理论编写算法和算法的价值敏感设计。具体而言：其一，依据某种伦理理论编写与设计算法。此种方法侧重于技术伦理的内在主义进路，即将伦理规范与价值判断前置性嵌入算法之中。例如，为了解决无人驾驶的伦理困境，德里克·里本(Derek Leben)依据罗尔斯正义理论中的“无知之幕”来设计自动驾驶汽车的算法。具体而言，里本通过编程使算法处于“无知之幕”，“无知之幕”之下的算法会隐藏事故中人作为乘客抑或行人的身份，从而使得算法能够选择最优解决方案，以最大程度地减少最坏情况的决策[26]。总之，算法的伦理设计是一种自上而下的进路，即把具体的伦理原则作为算法系统的准则。其二，价值敏感设计。相较于算法的伦理设计，价值敏感设计更加强调人与算法的互动关系，即将算法的使用情景与设计结合起来。根据弗里德曼的价值敏感设计理论，算法设计的第一步是概念审慎，即对算法所涉及的隐私、透明、公平、责任等概念进行厘定；第二步是经验上的调研，设计者主要思考如何处理自身持有的价值以及权衡各种不同价值的矛盾；最后是技术研究阶段，依据已确定的相关价值进行具体的算法设计[27]。由此，我们就可以把算法的伦理设计建立在具体情景分析的基础上，而不是提前进行主观决策以决定算法应嵌入哪些价值。

第二，应用范围的限定。由于算法的复杂性、不透明性，从而导致在算法系统模拟中未出现的情况可能会出现在具体应用中，因而需要对算法的应用进行限定。具体而言，一方面在某些领域禁止使用“黑盒”算法。对此，2018 年 4 月发布的《英国人工智能发展的计划、能力与志向》明确指出：避免在特定重大领域采用“黑盒”算法，即如果在特定重大领域采用该算法，则它必须能够为其决策提供令人满意的解释，如若不能，则在找到替代性方案之前禁止将其应用于特定重大领域。另一方面，对算法能否应用以及如何应用进行具体的限定。以自动驾驶汽车算法的应用为例，在自动驾驶汽车算法进入市场应用之前，需要生产商依据国家公路安全管理局所发布的交通安全统计数据向审查机构证明其安全性能，当其安全性能符合标准时，才能进入市场且只允许在公路上行驶[12]。

2. 外部进路

从外部进路看，主要有以下三种解决途径：

第一，设计者的伦理责任与多领域合作。算法是非常专业化的科技知识，因此作为专业人员的算法设计者是化解上述算法伦理问题的重要主体。一方面，应明确行业伦理规范与准则，包括公平性、安全性、透明性、可责性、包容性、预防性等。另一方面，通过科技伦理课程教育培养负责任的科技人才。2018 年初，美国高校纷纷开设科技伦理课程。如德克萨斯州大学为计算机专业的学生开设名为“计算机科学的伦理基础”的课程；康奈尔大学也开设了数据伦理课程，聚焦数据的伦理问题；斯坦福大学也正在开发“伦理、公共政策和计算机科学”的计算机伦理课程。同时，解决算法的伦理问题不仅仅是设计者的责任，更需要不同领域研究者的密切合作。例如公平算法的设计已经暗含了此问题的重要性。由于公平本身不是一个统计概念而是社会与伦理概念，因而不同的人会有不同的理解。因此，仅仅强调算法具有公平性是不充分的，更重要的是需要来自政治哲学、伦理学、计算机、法律等领域的学者深入分析算法所针对的具体问题，确定哪一种伦理准则是最恰当的[13]。

第二，公众参与。作为受算法影响的最广泛群体，公众参与对解决上述算法伦理问题日益重要。“公众参与的重要性在于算法系统的特定功能与公众对特定功能的解释之间会发生错位。例如，在谷歌商店搜索同性恋者的约会程序，其算法会推荐相关的程序以确定用户是否居住在性犯罪者附近。从纯粹的设计角度，即在科学环境或实验室中，它被解释为一种异常；然而，当算法面向公众并与公众交互时，则会被公众解读为歧视。因此，将公众对算法结果的理解融入算法设计是解决算法伦理问题的一种方法。”[28]进一步而言，“算法决策涉及社会共识这一伦理问题，而社会共

识仅依靠机器学习并不能解决,因而需要将人的判断融入到算法决策中。更准确地说,鉴于算法的广泛影响,需将公众对一些社会问题的共识融入到算法中,确定算法必须遵守的基本准则,从而使算法决策与社会判断相结合以促使公众与算法之间也达成契约,进而实现众机回圈(society-in-the-loop)"[29]。

第三,监管算法。监管算法主要包括依据算法的设计、复杂性、潜在危害等对其进行分类,制定性能标准、设计标准、责任标准以及上市前审计等。具体而言,主要包括以下几个方面[12]。

其一,制定各类标准。首先,依据算法的复杂程度可将其分类,主要包括:白箱(算法是完全确定的)、灰箱(算法非确定但易预测与解释)、黑箱(算法难以预测与解释)、已达到或超过人类智能的有感知能力的算法以及能够自我改进的奇点。监管机构可据此五类算法而进行不同程度的监管。其次,建立算法的性能标准。再次,制定算法的设计标准,例如,尽可能设计更加透明的算法、将解释能力融入到算法设计中等。最后,制定责任标准。监管机构可将算法的不同利益相关者聚集起来,在不影响创新的情况下制定灵活的责任标准。其二,制定最低限制程度的软触法规(soft-tough regulations)。例如,就算法的透明性而言,灵活规定算法是否需要公开,公开的程度、范围、时间等。其三,上市前对算法进行审核与批准,尤其是自动驾驶汽车、无人飞机等对公共安全有重大影响的算法。监管机构可要求公司证实其算法的安全性能,并依据上述相关标准,决定是否可以推向市场以及限定其上市推广后的使用范围。

五、结语与启示

面对算法所带来的潜在社会伦理问题与风险,促进算法的负责任创新与发展已经成为学者们的共识。因而,在中国大数据技术与人工智能高速发展的背景下,如何将算法伦理融入算法的发展与创新之中是一个重要的问题。

首先,在制度与规范层面上,算法已成为国外政策与法规的监管重点。例如,2018年1月,美国计算机协会下属的公共政策委员会发布了《关于算法透明性与可责性的声明》,该声明旨在确保算法透明与可责性,并明确了算法的七条原则:意识、获取和补救、问责制、解释、数据来源、可审计性以及严格的验证与测试[30]。同样以算法问责为目的,2018年5月25日已正式生效的欧盟《一般数据保护条例》表达了对算法复杂性与不透明性的担忧,并提出了算法可解释性的要求。另外,鉴于算法在人工智能、社会公益等领域的广泛应用,各领域具体层面的规范也相应产生。如前面提到的《英国人工智能发展的计划、能力与志向》对算法具体应用的限定。相比之下,我国在算法伦理的相关制度与规范层面就稍显薄弱,对算法的监管也仅是零星地存在于相关政策法规之中。如《新一代人工智能发展规划》中指出需加强人工智能的相关法律、伦理和社会研究,建立确保其健康发展的法律法规与伦理道德框架。因而,借鉴国外的相关做法,立足于中国特定的社会文化背景,从制度与规范的层面加强对算法的监管,确保算法技术在完善的伦理框架中发展。

其次,进一步加强算法领域的科技人员的伦理意识。如上所述,算法的伦理设计是应对算法诸多伦理问题的基本路径之一。因而,算法科技人员的伦理意识也就变得至关重要。国外已有一些大学对计算机专业的学生开设了相关课程,如哈佛大学与麻省理工学院共同开设了以人工智能的伦理道德与管理规范为核心内容的新课程,主要包括算法风险评分的普及、如何保证数据无偏倚、是否应当用机器来评判人类等问题。因而,借鉴国外相关课程,提升科技人员的伦理意识,促使科技人员在算法设计、应用中自觉遵守伦理规则,实现自律和他律的结合。

再次,加强人文学者、科技工作者与公众的交流,以确保算法决策的公平、透明与可责。相比较克隆技术、纳米技术等技术的伦理反思,算法的伦理讨论显示出更强的多方参与性。算法伦理研究者既有来自计算机科学技术、人工智能领域的科学家,也有来自于法律、社会学、伦理学、科技哲学等领域的人文学者。进而言之,只有不同领域的学者、政策制定者、公众的共同协作,才能让算法符合"善法"。因而,应当促使各领域学者和公众共同参与到算法伦理问题的研究与讨论之中。

参考文献:

[1] Kitchin R. Thinking Critically About and Researching Algorithms[J]. Information Communication & Society, 2017,20(1):14-29.

[2] Balkin J M. The Three Laws of Robotics in the Age of Big Data[J]. Ohio State Law Journal, 2017,78(5):1217

-1241.

[3] 清华大学中国科技政策研究中心. 中国人工智能发展报告 2018[EB/OL]. (2018-08-02)[2018-10-22]. https://www.sohu.com/a/244758046_673573.

[4] Rainie L, Anderson J. Code-dependent: Pros and Cons of the Algorithm Age[R]. Washington, D. C.: Pew Research Center, 2017.

[5] Gillespie T. Algorithm[draft][#digitalkeyword][EB/OL]. [2018-10-22]. http://culturedigitally.org/2014/06/algorithm-draft-digitalkeyword/.

[6] Ananny M. Toward an Ethics of Algorithms: Convening, Observation, Probability, and Timeliness[J]. Science, Technology & Human Values, 2016,41(1):93-117.

[7] Neyland D, Möllers N. Algorithmic IF... THEN Rules and the Conditions and Consequences of Power[J]. Information, Communication & Society, 2017,20(1):1-18.

[8] Ziewitz M. Governing Algorithms: Myth, Mess, and Methods[J]. Science, Technology & Human Values, 2015,41(1):3-16.

[9] 张卫. 技术伦理学何以可能?[J]. 伦理学研究, 2017(2):79-83.

[10] Nicholas D. Algorithmic-accountability: The Investigation of Black Boxes[EB/OL]. [2018-10-22]. https://www.cjr.org/tow_center_reports/algorithmic_accountability_on_the_investigation_of_black_boxes.php.

[11] Kraemer F, van Overveld K, Peterson M. Is There an Ethics of Algorithms?[J]. Ethics and Information Technology, 2011,13(3):251-260.

[12] Tutt A. An FDA for Algorithms[J]. Administrative Law Review, 2017,69(1):83-123.

[13] Bruno L, Oliver N, Letouzé E, et al. Fair, Transparent, and Accountable Algorithmic Decision-making Processes[J]. Philosophy & Technology, 2018,31(4):611-627.

[14] Goodman B, Flaxman S. European Union Regulations on Algorithmic Decision-making and a "Right to Explanation"[J]. Ai Magazine, 2016,38(3):1-9.

[15] Cheney-Lippold J. A New Algorithmic Identity: Soft Biopolitics and the Modulation of Control[J]. Theory, Culture & Society, 2011,28(6):164-181.

[16] Stahl B C, Timmermans J, Mittelstadt B D. The Ethics of Computing: A Survey of the Computing-oriented Literature[J]. Acm Computing Surveys(CSUR), 2016,48(4):1-38.

[17] Diakopoulos N. Algorithmic Accountability: Journalistic Investigation of Computational Power Structures[J]. Digital Journalism, 2015,3(3):398-415.

[18] Mittelstadt B D, Allo P, Taddeo M, et al. The Ethics of Algorithms: Mapping the Debate[J]. Social Science Electronic Publishing, 2016,3(2):1-21.

[19] Latzer M, Hollnbuchner K, Just N, et al. The Economics of Algorithmic Selection on the Internet[M]// Bauer J M, Latzer M. Handbook on the Economics of the Internet. Cheltenham: Edward Elgar, 2016:395-425.

[20] Gillespie T. TheRelevance of Algorithms[M]// Gillespie T, Boczkowski P, Foot K. Media Technologies: Essays on Communication, Materiality, and Society. Cambridge: MIT Press, 2014:1-32.

[21] Lepri B, Staiano J, Sangokoya D, et al. The Tyranny of Data? The Bright and Dark Sides of Data-driven Decision-making for Social Good[M]// Cerquitelli T, Quercia D, Pasquale F. Transparent Data Mining for Big and Small Data. Dordrecht: Springer, 2017:3-24.

[22] Floridi L. Group Privacy: A Defence and an Interpretation[M]// Taylor L, Floridi L, van der Sloot B. Group Privacy: New Challenges of Data Technologies. Dordrecht: Springer, 2017:83-100.

[23] Taylor L, Floridi L, van der Sloot B. Introduction: A New Perspective on Privacy[M]// Taylor L, Floridi L, van der Sloot B. Group Privacy: New Challenges of Data Technologies. Dordrecht: Springer, 2017:1-12.

[24] Taylor L. Safety in Numbers? Group Privacy and Big Data Analytics in the Developing World[M]// Taylor L, Floridi L, van der Sloot B. Group Privacy: New Challenges of Data Technologies. Dordrecht: Springer, 2017:13-36.

[25] Eriks S. Designing for Accountability[C]// Bertelsen O W, Bodker S. Proceedings of the Second Nordic Conference on Human-computer Interaction. New York: ACM, 2002.

[26] Leben D. A Rawlsian Algorithm for Autonomous Vehicles[J]. Ethics and Information Technology, 2017,19(2):107-115.

[27] 郭林生. 论算法伦理[J]. 华中科技大学学报(社会科学版), 2018,32(2):40-45.

[28] Baumer E P. TowardHuman-centered Algorithm Design[J]. Big Data & Society, 2017,4(2):1-12.

[29] Rahwan I. Society-in-the-loop: Programming the Algorithmic Social Contract[J]. Ethics and Information Technology, 2018,20(1):5-14.

[30] Statement on Algorithmic Transparency and Accountability[R]. Washington D. C.: ACM U. S. Public Policy Council, 2017.

(责任编辑：李新根)