DOI:10.13530/j.cnki.jlis.2022051

算法偏见概念、哲理基础与后果的系统回顾*

贾诗威　闫　慧

摘　要　算法偏见是社会偏见在算法信息社会的延伸,映射信息不平等。本文采用系统综述方法,对国内外算法偏见类型的实证研究进行梳理综合,在界定算法偏见的内涵和外延基础上,整合算法偏见的二元主体概念框架,并归纳算法偏见的三条形成链,包括人类智能系统内的偏见链、人工智能系统内的偏见循环圈、人类智能系统到人工智能系统的偏见作用路径。然后从功能主义、冲突论、还原论和马克思主义角度反思算法偏见研究的哲理基础。最后从信息科学视角构建技术、信息、用户之间的三元交互影响模型,讨论算法偏见在技术、信息、用户三要素互动过程中的存在形式,并发现算法偏见导致的三类不平等现象——信息呈现不平等、信息分布不均衡和新型数字不平等。研究结果深刻揭示了算法时代的信息社会问题,并赋予信息领域的传统不平等话语新的内涵与外延,为信息职业人员应对算法时代新挑战提供思路和参考。图3。表2。参考文献92。

关键词　算法偏见　信息不平等　数字不平等　信息呈现　信息分布

分类号　G252

A Systematic Review of Concept, Philosophy Foundation and Impacts of Algorithmic Bias

JIA Shiwei & YAN Hui

ABSTRACT

More and more algorithms are being used to replace or assist humans in distributing important social goods and making automated decision-making, which has contributed greatly to social development. But algorithms are not entirely fair, and public concern about algorithmic bias has become a social issue as algorithmic bias is discovered and reported. This paper adopts a systematic review method to sort out and synthesize empirical research on algorithmic bias, aiming to answer three questions: What is algorithmic bias? What are the philosophy foundations of algorithmic bias? And what are the impacts of algorithmic bias in the perspective of information science?

Therefore, on the basis of defining the connotation and extension of algorithmic bias, the dual subject conceptual framework of algorithmic bias—human intelligence system and artificial intelligence system—is summarized. It reveals the bias chain within the human intelligence system, the bias loop within the artificial intelligence system, and the bias action path from the human intelligence system to the artificial intelligence system. Algorithmic bias is essentially a philosophical topic, and reflection on its

* 本文系国家社会科学基金项目“数字中国背景下数字贫困消减行动研究”(编号:18BTQ007)的研究成果之一。(This article is an outcome of the project “Action Research on Digital Poverty Alleviation in Context of Digital China” (No. 18BTQ007) supported by the National Social Science Foundation of China.)

通信作者:闫慧,Email:hyanpku@ruc.edu.cn,ORCID:0000-0002-3649-1601 (Correspondence should be addressed to YAN Hui, Email:hyanpku@ruc.edu.cn;ORCID:0000-0002-3649-1601)

philosophical foundations can help people better understand the formation mechanism of algorithmic bias. The philosophical perspectives adopted by algorithmic bias reflection mainly include Functionalism (including Structural Functionalism and Machine Functionalism), Conflict Theories, Reductionism, and Marxism. Different perspectives have different interpretations of the forming mechanism of algorithmic bias. At last, a three-dimensional interaction model of technology, information, and users is constructed to discuss the impacts of algorithmic bias from the perspective of information science. Three major impacts of algorithm bias on information inequality and digital inequality are described: inequality of information presentation, unbalanced distribution of information, and a new form of digital inequality known as the algorithms divide.

By deeply revealing the algorithmic social problems in the era of artificial intelligence, this research endows the traditional discourse of inequality in the information field with new connotation and extension, proposing possible references for information professionals to face new challenges. 3 figs. 2 tabs. 92 refs.

KEY WORDS

Algorithmic bias. Information inequality. Digital inequality. Information presentation. Information distribution.

0 引言

算法被广泛应用于社会治理、行业变革等领域,是推动社会发展的新动力。在带来便利的同时,算法在医疗保健[1]、刑事司法[2]等各领域存在偏见的事例被相继报道,引发公众对算法偏见的担忧。对于算法偏见,以图灵奖得主杨立昆(Yann LeCun)为代表的部分学者持算法中立态度[3],即算法本身不存在任何偏见。与此相反,另一部分学者则认为算法并非是完全客观和价值中立的技术,并从法律、伦理、技术与权利互构论等角度对算法偏见本质进行深入探讨,认为算法被内在或外在地赋予政治性,算法偏见的本质是社会偏见在算法技术中的映射,是对公民的平等权、隐私权、数据安全的侵犯[3,4],如检索系统偏爱具有某些特定特征的文档[5],图像搜索结果显示男性比女性更能代表"person",性别陈规定型观念从社会延伸至网络空间[6]。

算法偏见问题已经成为信息科学、社会学、伦理学、传播学等学科的关注焦点,并纳入图书情报学[7]、计算机科学[8]专业的课程教育之中。作为新兴概念,学界对算法偏见的定义并不清晰,不同学科对算法偏见的概念界定、形成来源各有不同,不利于跨学科话语体系构建。例如,算法偏见既可指无法完全代表被采样对象整体的数据属性[9],也可指算法技术不公平地偏袒或不偏袒特定群体或个人[10]。算法偏见产生于不同环节并表现出多种形式[11],形成原因包括显性因素和隐性因素。本文采用系统综述方法对国内外算法偏见研究的概念、哲理基础、后果进行系统梳理,探索该领域未来的研究方向。具体问题包括:什么是算法偏见,算法偏见的基础理论来源有哪些,信息科学视角下算法偏见的后果有哪些。

1 研究方法

本研究采用循证社会科学领域广泛使用的PRISMA(Preferred Reporting Items for Systematic Reviews and Meta-analyses)声明对算法偏见研究进行文献分析和综合。PRISMA声明由条目清单和流程图组成,其中条目清单从标题、摘要、引言、方法、结果、讨论、其他信息等七方面对综

述过程进行严格规范,流程图描述系统综述各阶段的文献纳入/排除的数量和原因[12]。PRISMA 声明可以最大限度地减少发表偏倚,呈现可靠的证据和可信的结论,是国内外期刊广泛采用的系统综述流程规范。

1.1　检索策略

研究遵循 PRISMA 声明,选取中国知网、Scopus、Web of Science 核心合集、ProQuest、IEEE Xplore Digital Library、ACM Digital Library 作为检索源,以“算法偏见”OR“数据偏见”,“algorithm * bias”OR “data bias”作为检索词,语种限定为中英文,文献类型限定为期刊论文和会议论文。检索时间截至 2021 年 12 月 31 日,共获得文献 2 000 篇。随后以同题名、同作者文献按照最新发布时间、期刊论文优先的原则,合并多源检索结果并删除重复记录,获得文献 1 332 篇。

1.2　文献纳入/排除标准

根据研究问题和目的,制定本研究文献纳入/排除标准如下:①纳入以算法偏见或数据偏见为主要研究对象的论文,讨论范围涵盖算法偏见或数据偏见的类型、来源、模型验证、表现等方面,排除解决对策研究(属算法公平范畴);②纳入有具体研究设计和研究结果的实证论文,排除纯理论思辨类和文献综述类研究;③排除论文列表中摘要、海报、更正等无关内容;④纳入 CCAT 批判性评估达 60%及以上的研究。

需要注意的是,“bias” 可译为偏见或偏差,前者描述结果与受保护属性的不公平依赖关系[13]或对利益相关方合法权益的不公平损害[8],如性别偏见、种族偏见等;后者描述结果偏离标准实例的差异[14],如统计偏差。本文取“偏见”之意,在文献筛选时需人工排除非算法偏见的文献。依照 PRISMA 流程图,对文献检索和纳入/排除后,最终确定纳入 42 篇文献,均为英文文献(见图 1)。

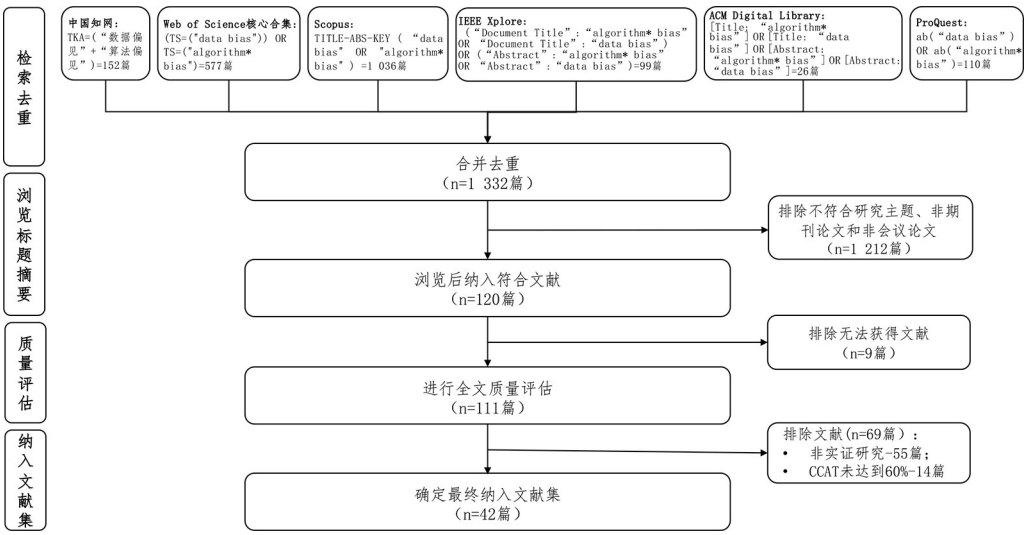

图 1　PRISMA 研究选择流程

1.3　质量评估

为保证纳入文献的质量,本文采用 Crowe 批判性评估工具(CCAT 1.4 版)进行科学规范的评估。CCAT 评估表由预备知识、引言、设计、抽样、数据收集、道德事项、结果、讨论八部分组成,各部分以 0—5 分进行评分和计算百分

比[15]。相关研究表明,CCAT 质量得分在 60%及以上时,可认为被评估文献属于高质量研究[16]。纳入文献大多采用建模与仿真实验、统计分析等定量研究方法,另有 8 篇采用内容分析、情感分析、德尔菲法、案例分析等定性研究方法,两篇采用混合研究方法。对纳入文献进行分析后,采用批判性解释综合方法对算法偏见的概念、哲理基础与后果进行数据综合,并发展出新的概念和理论。表 1 列举了部分高质量文献的数据特征。

表 1 高质量文献的数据特征(部分)

作者及年份	研究主题	研究方法	算法偏见定义	偏见类型	哲理基础	算法偏见后果	CCAT(%)
Pandey 等[17],2021	价格歧视中的 AI 偏见	建模与仿真实验	与乘客社群人口统计特征相关的票价定价算法结果差异	人口统计特征偏见	马克思主义	不同人口统计特征的用户或社群面临不同的车费定价	63
Peralta 等[18],2021	验证具有算法偏见的社交网络意见形成	建模与仿真实验	个性化过滤算法控制用户看到和发送的信息,人们倾向于选择与本人一致的观点,导致社交网络的意见形成	迭代算法偏见	还原论	偏见不平衡会对最终的意见状态和总体动态产生重要影响,可能决定最终的全球观点和人口的动态行为,催生过滤泡、回声室等现象	63
Abul-Fottouh 等[19],2020	YouTube 疫苗视频推荐中的算法偏见	视频情绪编码分析、社交网络分析	/	交互偏见	/	YouTube 推荐算法降低了反疫苗视频的可见性,疫苗相关视频中存在过滤泡和同质化效应	75
Hargittai[20],2020	大数据中的数据偏见	统计分析	/	数据偏见、代表性偏见	冲突论	社交平台偏向于教育程度更高、互联网使用更熟练的用户,用户意见和行为仅代表部分人口	73
Srinivas 等[21],2019	人脸识别中的算法偏见问题	建模与仿真实验	/	年龄偏见、紧急偏见、数据偏见	机器功能主义	当前儿童面部识别远落后于成人面部识别,每种算法都对儿童面部识别存在偏见现象	85
Flexer 等[22],2018	音乐推荐中高维数据的算法偏见	建模与仿真实验	在计算机实验中产生不公平结果的系统性和可重复错误,即为某些用户或某些数据生成一个结果,为其他用户或数据生成另一个结果	技术算法偏见	还原论	Hubness 是一种技术算法偏见,依据歌曲靠近或远离数据中心的特性,导致推荐系统对某些音乐的不公平对待,要么频繁推荐,要么从不推荐	68
Wilkie 等[5],2018	领域检索算法性能受算法偏见影响	实验法	在领域检索时,根据字段的权重方式以及字段是否填充,检索算法可能会不适当地偏袒某些文档	可检索性偏见	/	当字段应用不正确时,字段可能会引入算法偏见,从而损害检索性能	70

2　算法偏见的内涵与外延

2.1　算法偏见的内涵

在42篇文献中,有20篇文献明确定义了算法偏见。经批判性解释综合,将算法偏见内涵的组成要素归纳为偏见来源、偏见对象、偏见后果、偏见特征四类,各要素具体解释见表2。

算法偏见内涵由四个关键要素组成:①偏见来源:产生于业务理解、数据收集、算法模型、交互反馈等环节的计算机实验全流程;②偏见对象:对应算法流程,可分为数据级、内容级和用户级三种算法偏见类型;③偏见后果:体现为不公平地支持或反对某些个人或群体、某些信息内容等,这种不公平结果不一定表现为歧视,但可能会扩大社会不平等;④偏见特性:偏见具有系统性、可重复性,即算法偏见不仅仅发生在孤立案例中。

表2　算法偏见内涵要素

内涵要素	具体解释	文献来源
偏见来源	业务理解	Akter 等[23],2021
	数据收集	Akter 等[23],2021;Mansoury 等[24],2020
	算法模型	Akter 等[23],2021;Mansoury 等[24],2020;Flexer 等[22],2018;Morstatter 等[25],2017
	交互反馈	Akter 等[23],2021;Mansoury 等[24],2020;Sun 等[26],2020;Lin 等[27],2019;Sun 等[28],2018
偏见对象	数据级别	Mansoury 等[24],2020
	内容级别	Abdollahpouri 等[29],2020;Wilkie 等[5],2018
	用户级别	Metaxa 等[30],2021;Pandey 等[17],2021;Wang 等[31],2021;Lin 等[27],2019
偏见后果	对内容和用户呈现不公平结果	Bonezzi 等[32],2021;Heuer 等[33],2021;Peralta 等[18],2021;Peralta 等[34],2021;Abdollahpouri 等[29],2020;Aysolmaz 等[35],2020;Sîrbu 等[36],2019;Díaz 等[37],2018;Flexer 等[22],2018;Wilkie 等[5],2018;Otterbacher 等[6],2017
	扩大社会不平等	Pandey 等[17],2021;Metaxa 等[30],2021
偏见特性	系统性、可重复性	Otterbacher 等[6],2017;Flexer 等[22],2018;Aysolmaz 等[35],2020

由此,可将算法偏见定义为:在计算机实验全流程中,对某些个人或群体、信息内容等产生不公平结果的系统性和可重复性错误。算法偏见的划分类型多样,有40篇文献涉及各类算法偏见的描述,存在同一类型不同命名的情况。本文以关键要素作为统一算法偏见类型的依据,拓展算法偏见概念外延。

2.2　外延标准一:算法偏见来源

从业务理解到实践反馈,算法偏见包括理解偏见(understanding bias)、数据集偏见(dataset bias)、技术偏见(technical bias)和实践偏见(practical bias)。

业务理解是偏见进入系统的首要环节[35],设计者需在此环节中明确算法用于达成什么目标,决定收集哪些数据,由谁收集,如何处理这些数据,选用哪种算法模型等内容。当算法设计者对问题规范和目标变量的理解与预期目标不匹配时,理解偏见就此产生[38]。例如,311平台是美国使用最为广泛的电子政务系统之一,用于居民反馈非紧急城市服务问题,以实现高效公平的公共服务交付目标。但实际上,设计

者并未考虑到文化、经济、地理等因素对平台决策的影响,从设计之初就将低收入、少数族裔聚居的社区排除在城市治理之外,最终导致不公平的城市治理决策结果[39]。除此之外,理解偏见同样产生于算法设计者采用不匹配的算法模型的情况,例如将健康成人面部识别模型用于识别儿童[21]、老年痴呆症患者[40],这种理解偏见也被称为紧急偏见(emergency bias)。

正如谚语"bias in,bias out"(偏见进则偏见出)所言,当数据集本身呈现出偏见时,由此衍生出的结果一定存在某种偏见。数据集偏见的形成根植于社会制度、实践和采集者态度,因此也被称为预存偏见(pre-existing bias)。针对数据集构建流程,已有文献将数据集偏见划分为历史偏见(historic bias)、代表性偏见(representativeness bias)和标签偏见(label bias)。其中,历史偏见源自训练算法的历史数据,受到先前抽样不当、标签错误等影响[41],导致算法模型延续或加剧历史数据中的偏见;代表性偏见的产生是因为数据集规模不足或数据集无法代表目标群体的整体数据属性[20,23],以此数据集训练出的算法会倾向性地利于某类群体;数据标签是帮助算法达成目标的关键决定因素,标签偏见产生于人类选择标注和删除信息的过程,这一过程最容易受到人类认知偏见和社会偏见的影响,反映出社会结构的不平等[42]。

受制于算法模型、计算能力、系统约束等条件,算法技术本身可能会存在某种技术偏见,产生不公平的算法结果。根据技术偏见形成的影响因素,已有文献将技术偏见进一步划分为关联偏见(association bias)和确认偏见(confirmation bias)。关联偏见与带有偏见的数据集、陈规定型观念密切相关,表现为算法模型受偏见数据训练后强化并放大了训练集中潜藏的文化偏见,如机器翻译软件放大了语言中的性别偏见[43];确认偏见常在于个性化推荐系统中,体现为过度简化的个性化推荐对群体或个人做出有偏见的假设,例如电影推荐系统暗藏性别陈规定型观念,依据性别特征偏见性地呈现推荐结果[27]。

实践过程是算法偏见形成的最后环节,体现为交互偏见(interaction bias)。交互偏见来自算法选择给人类呈现信息子集的过程,是由于人机交互而产生的算法偏见,体现了算法偏见的动态性和迭代特征,又称迭代算法偏见(iterated algorithmic bias)[26]。迭代算法偏见增加了预测相关性的不平等性,包括迭代过滤、主动学习和随机选择三种存在形式[28],在算法推荐[19,29]、社交网络[34,36]领域被广泛研究。

2.3 外延标准二:算法偏见对象

算法偏见是一种体现人与算法之间的偏见映射,偏见对象涉及数据、算法生成内容和终端用户,分别对应数据级、内容级、用户级三种算法偏见类型。

其中,数据级算法偏见又称数据偏见,主要描述数据集中隐含的偏见情况,详细分类参考外延标准一中的数据集偏见。

内容级算法偏见,又称算法推荐偏见(algorithmic recommendation bias),是从算法生成内容的角度考虑的算法偏见[31],侧重关注算法本身导致的偏见,即在A、B内容合理共存情况下,算法为什么向用户呈现A内容而非B内容。算法推荐偏见常现于各类推荐系统、社交媒体、检索系统中,在传播学、信息科学领域被广泛研究。根据已有文献,算法推荐偏见可细分为暴露偏见(exposure bias)、流行性偏见(popularity bias)和可检索性偏见(retrievability bias)。暴露偏见是指算法模型只允许用户接触到特定项目的一部分,例如Twitter的API过滤机制[25]和时间线管理算法[44]均扭曲了用户看到推文内容的平等性。流行性偏见是指流行项目比不流行项目更频繁地推荐给特定群体或所有用户,例如高分视频会被推荐给更多用户[33,45],在某种程度上是另一种形式的暴露偏见;流行性偏见会增强长尾现象,可能会对产品提供者、个性化追求用户产生不公平结果。可检索性偏见是用于描述检索系统性能的概念,即文档在检索系统中的

可访问性[5];不同检索算法对文档的选择倾向性不同,系统的可检索性又决定了相关性文档的呈现机会[46],最终表现出检索系统的可检索性偏见,可能会对医疗保健、新闻舆论等造成恶劣影响。

用户级算法偏见,又称算法用户偏见(algorithmic user bias),是从目标用户的角度考虑的算法偏见,主要考虑算法在不同用户组之间的性能差异(如准确性)[31]。用户组的划分标准依赖于人口统计学特征,包括性别、年龄、种族、宗教、地区、经济水平、教育水平等。此时,算法偏见可被描述为对不同人口统计学特征的歧视,是社会偏见在虚拟世界的映射,又称人口偏见。按照用户特征的组合情况,包括两种类型。①单一特征偏见。关注单一用户特征对算法结果的影响,包括算法性别偏见[47]、算法年龄偏见[37]、算法种族偏见[42]、算法地理偏见[48]等。其中,算法性别偏见隶属于性别研究范畴,不仅关注性别特征在算法应用中的影响作用,更重要的是对分类系统本身(如男人/女人)以及分类系统如何映射或再现社会不平等的现象进行批判性反思[49];算法年龄偏见是基于年龄对个人或群体存在刻板印象,通常用于表达对老年人/年轻人、成人/儿童的偏见态度;算法种族偏见特指算法开发者在权力结构、社会期望、信念和价值观的引导下,对原始算法进行修改或再训练,最终体现出对某类种族群体显性或隐性的种族偏见[50];地理特征通常是种族、经济因素等社会敏感特征的代理属性(proxy attributes)[51],算法地理偏见体现地理因素在算法实践中的偏见影响。②多特征交叉偏见。偏见通常不是由单一特征造成的,是多种特征因素的排列组合对算法结果产生不同影响。在12篇多特征交叉偏见研究文献中,性别与种族特征的交叉影响最受关注,涉及5篇文献[30,52-55]。

2.4　算法偏见的概念框架

综合算法偏见的内涵与外延研究成果,确定算法偏见的二元主体框架——人类智能系统与人工智能系统,由此构建算法偏见的整体概念框架(见图2)。其中,人类智能系统由设计者、数据采集者、打标员、程序员和用户五类群体组成,前四者又可统称为算法项目人员或人工智能系统项目人员;人工智能系统由数据、算法模型(以下简称算法)和界面等主要部分构成。算法偏见的概念框架由三部分组成。

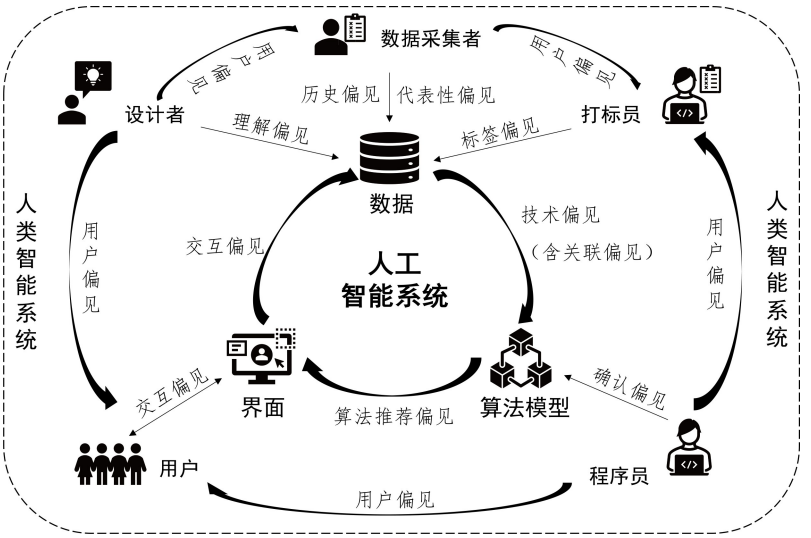

图2　算法偏见的概念框架

(1)人类智能系统内的偏见链

人类智能系统内的偏见链条是算法有关的人类主体在参与人工智能系统开发、设计与运营等事务前具有的相互偏见关系,其隐式或显式偏见关系影响到算法的设计和性能[56],是各类算法偏见产生的根源。人工智能系统的设计者对特定性别、年龄、种族和地理的用户存在的偏见会通过系统设计环节传递给数据采集者。在社交媒体中,数据采集者与用户两个身份高度融合,在设计者带有偏见的系统设计思想指引下,数据采集者在上传数据内容的过程中融入了可能存在于自身的偏见。打标员包括专业人士和普通用户,分别对应专业生产内容(Professional Generated Content)和用户生成内容(User Generated Content)两种系统场景[57]。打标员为融合了设计者偏见和数据采集者偏见的数据内容赋予标签,同时又不可避免地将自己的社会和认知偏见体现在选择标注词与删除相关数据的过程中。程序员在开发人工智能系统过程中带着其对用户和打标员群体可能的偏见,设计和改进供打标员使用的标注平台。人工智能系统用户在人类智能系统的偏见链条中常常处于最底端,无意识或者有意识地被人工智能系统的其他四个人类主体置于偏见的恶性循环之中。用户、数据采集者和打标员的身份经常融合到同一个自然人之中,成为多重偏见的映射集合。

(2)人工智能系统内的偏见循环圈

人工智能系统内的偏见循环圈由数据到算法、算法到界面、界面到数据三个环节构成。数据到算法阶段的算法偏见表现为技术偏见,数据集融合了来自人工智能系统项目人员如设计人员、数据采集员、打标员等的社会和认知偏见,在运用体现了程序员确认偏见的算法对数据集进行训练、验证和测试过程中呈现出关联偏见,训练集中的文化偏见容易被放大,算法模型将数据集中算法用户偏见进一步延续和强化;算法到界面阶段的算法推荐偏见常产生于拥有个性化推荐机制的人工智能系统,比如社交媒体、流媒体、检索系统等,推荐算法将相关人员的偏见融合到一起,形成了可能的暴露偏见、流行性偏见和可检索性偏见;界面到数据阶段的偏见是用户与界面之间交互偏见的延续,用户与被算法推荐偏见影响的界面交互过程中,与用户自身可能存在的对其他用户的偏见融合到一起,生成用户日志数据,反馈到数据中,从而进一步叠加到数据集偏见中。

(3)人类智能系统到人工智能系统的偏见作用路径

在人类主体与人工智能系统进行交互时,人类智能系统内的偏见链条会以多条路径进入并影响人工智能系统[56]。从人工智能系统项目人员到数据的作用路径中,设计者个人对业务的理解状况很大程度决定了人工智能系统项目的成功与否。当业务理解出现偏差或因个人认知隐式地带入社会偏见时,由此产生的理解偏见将通过数据采集者对后续数据集的构建产生影响,促使数据集中产生可能的历史偏见和代表性偏见。同时,打标员在可能存在偏见的数据集中进行带有标签偏见属性的打标行为,在处理数据时从其自身社会偏见出发对部分属性或数据做出无意识的保留、舍弃或标注行为。

从程序员到算法的作用路径中,以推荐智能系统为例,程序员从对特定群体或个人有文化、认知与社会偏见的假设出发,过度简化个性化推荐算法,或者采用适合于其他人群但不适合于被偏见人群的个性化推荐算法,导致为被偏见人群推荐的内容主题要么过于狭窄,要么过于流行而不适用,要么无法便利地检索到真正需要的内容。

从用户到界面的作用路径中,界面成为算法及人工智能系统与用户之间沟通的中介要素。大多数情况下,有限的人类认知和精力无法接收所有的算法结果,这就要求算法模型通过界面选择性地呈现信息。选择呈现信息的类型受到算法推荐偏见的影响,也受到界面个性化程度及界面设计者所持的文化、认知与社会偏见状况的影响,用户会根据自身需求和个人

认知对算法结果做出不同的操作行为,通过个人获取及使用信息行为特征和习惯为界面及其后台反馈了有价值的日志数据,数据中带着交互偏见补充到数据集中,对预测用户后续行为提供有价值依据,同时触发进入新一轮的偏见循环。

由此可见,在以二元主体为支撑的算法偏见概念框架中,不同角色的人类主体构成了算法偏见的根源,用户偏见在人类智能系统中出现多种形态与相互关系,并发生类似于食物链的传递,算法用户偏见与个人认知、社会文化有关,是社会偏见在算法世界的延续。数据—算法—界面(中介)构成的人工智能系统中的偏见,以不同形态在人类智能系统代表性主体的偏见作用之下形成循环,在循环中,算法偏见不断地自我加强和放大[24],最终影响用户未来的决策和行为。

3　算法偏见的哲理反思

近年来,社会各界掀起了对算法技术自身、技术参与所带来的风险挑战等问题的哲理反思热潮。算法偏见是其中的一个哲学主题,已经成为人工智能伦理道德领域的关注重点。当前算法偏见反思研究采用的哲学思想主要有功能主义、冲突论、还原论和马克思主义。

3.1　功能主义与算法偏见

功能主义(Functionalism)在不同学科语境下指代的思想内涵稍有不同,譬如社会学中有结构功能主义(Structural functionalism),心灵哲学中有机器功能主义(Machine functionalism)。

结构功能主义者将社会类比为生物有机体,认为社会是一个由众多部件组成的功能综合体,旨在根据社会现象与某种系统的关系来理解社会现象[58]268-269。这一假设决定功能主义者追求一致、稳定的社会关系,即使默顿引入反功能的概念来描述变迁和冲突,也难以动摇人们对稳定系统的注意力,最终产生保守主义的偏见[58]288-289。算法偏见作为一种破坏社会稳定的现象,其形成机理可从功能主义视角阐释。在算法应用中,算法功能实现被视为首要目标,但这一过程往往忽视了对算法道德和伦理的强调,这将对受道德规范保护的目标群体发挥反功能,由此产生的算法偏见将导致越轨行为的出现。功能主义对算法偏见的影响主要体现在分类系统的设计缺陷。例如,社会角色理论将男性定义为功能性角色、女性是表达性角色,强调互补性劳动分工对家庭、社会稳定的重要性[59]。这一设计理念限制了性别的多元性,剥夺了残疾、性少数群体[60]等非常规个体或家庭的参与权和发言权,在算法社会中再现现实社会对这类群体的不公平现象。

机器功能主义者借助图灵机模型来理解心灵,认为心灵状态是一种在因果网络(感官输入、行为输出、其他心灵状态)中具有特定因果作用的中间环节,即功能状态。机器功能主义的核心概念是“多重可实现”(multiple realization)[61],这正是人工智能的哲学根源所在。近年来,强人工智能以神经科学为基,发展出多种神经网络算法来解决复杂图像识别、行为预判等问题,并在计算机视觉领域取得进展。但心灵之于大脑是否等价程序之于硬件?就当前人工智能发展程度来看,答案是否定的,这也导致了算法偏见的形成。例如,现有的面部识别算法常因设计训练时被隐式植入对肤色(或种族)、年龄、健康状态的偏见,最终导致深肤色群体[2]、老年/儿童[21]、患病群体[40]的识别错误率远高于其他群体,在结果上落后于人工识别的准确率,对部分群体造成不可逆转的伤害。

3.2　冲突论与算法偏见

冲突论(Conflict theories)起源于20世纪50年代,是在马克思、韦伯、齐美尔社会冲突思想基础上,对结构功能主义进行批判和修正的西方社会学流派。不同于结构功能主义,冲突论首要关切的是社会不平等问题[58]292,重点探讨社会冲突的本质和根源。

从冲突论视角理解,算法偏见是社会结构性压迫在算法应用中的体现,反映出权力阶层之间的冲突。计算机科学家和活动家乔伊·博拉维尼(Joy Buolamwini)提出“权力阴影”(power shadows)的概念,以此描述算法应用所反映的世界结构性不平等现象[60]。冲突论在算法偏见研究中主要体现在性别社会分层理论中,表现为男女两性对资源占有问题展开统治权争夺[59]。在父权制意识的控制下,男性通过构建男强女弱的性别话语来巩固男性统治地位,并通过社会媒体等途径构建网络空间的性别秩序和性别—能力规范,即男性比女性更能代表“person”[6],男性具备高能力而女性具备高温暖。当出现违反性别—能力规范的情况(如高能力的女性)时,这种违规行为将受到整体环境的抵制和惩罚[6]。同样的情况还出现在种族[30]、政治阶层[62]之间,体现了算法偏见背后的冲突对抗和权力不对等关系。

3.3 还原论与算法偏见

还原论(Reductionism)是一种将复杂现象或理论还原为基础现象或理论的哲学思想,目前主流的还原论说法认为所有学科都可以还原为物理学的表达形式[63]。

算法偏见是一种复杂的社会心理学现象在算法世界的体现,受到多种因素的影响。如何将现实世界的偏见现象转换为计算机理解的语言,还原论作为中介桥梁发挥了重要作用。例如,计算机无法直观理解什么是“美”,但引入还原论的理论方法,可将“美”这一文化概念分解成多个物理特征的组合,将身体吸引力还原为人类产生生殖冲动的函数[55]。种族、年龄、性别、国家作为函数构成要素,影响算法判断结果从而产生算法偏见[56]。不同主题领域对算法偏见的还原论解释不同,社交媒体网络中算法偏见可还原为社交网络中节点间的偏置强度[18],推荐算法中算法偏见可还原为高维空间中测量数据间距离的问题[22]。采用还原论的理论视角,可将算法偏见现象还原为物理现象,其影响因素和形成机理还原为数理逻辑集合,实现算法世界的偏见转移和再现。

3.4 马克思主义与算法偏见

马克思主义(Marxism)是一种从历史唯物主义、辩证法和对资本主义的批判中发展而来的经济、政治和社会世界观。其中,马克思主义哲学、马克思主义政治经济学、科学社会主义是组成马克思主义的三大基石。下文将用马克思主义政治经济学的经济人假设来解释算法偏见的形成原因。

继承古典经济学经济人假设的合理部分,马克思主义政治经济学以社会生产关系作为研究对象,假定经济人的自利行为不具备增进社会福利和自动调节经济均衡功能,反而会造成社会经济的波动和冲突[64,65],这与古典经济学的经济人假定相反。基于此,广告、医疗保健、游戏、网约车等领域中的算法偏见现象便可以解释为:算法决策者或供应商作为资本方,在设计算法应用过程中必然遵循经济人假设,力图从信息资源中获得最大利益。这一理念驱使资本方做出关注有利可图的信息市场,同时规避无利可图的信息市场的决策,由此设计或发行的算法应用必将受到经济因素的扭曲,可能会不公平地偏袒或歧视特定群体。例如,算法结果增加用户获得同等广告服务[66]、医疗保健服务[42]的难度,割裂不同经济水平地区用户的游戏体验[48],或增加用户遭遇网约车计费的算法收割可能[17]。

4 算法偏见的后果

正如本文构建的算法偏见概念框架所示,算法偏见的后果涉及技术要素、信息要素和用户要素。其中,技术是算法偏见的形成手段,信息是算法偏见的呈现载体,用户是算法偏见的作用对象或催化剂。只有厘清技术、信息与用户之间的关系,才能明晰算法偏见在各要素互动过程中的表现形式和后果,为消除算法偏见

提供针对性参考。由此构建技术、信息、用户(特指内容消费者而非内容生产者)之间的三元交互影响模型(见图3),归纳出当前研究中存在的三类算法偏见后果,包括信息呈现不平等、信息分布不均衡和新型数字不平等。

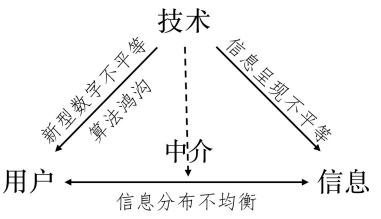

图3　算法偏见的后果框架

4.1　技术与信息:信息呈现不平等

在技术与信息这对关系中,将侧重描述技术如何影响信息呈现的不平等,淡化用户在其中的影响,重点关注两类信息呈现不平等形式:信息呈现机会不平等和信息话语权不平等。

(1)信息呈现机会不平等

信息呈现机会描述的是多元化信息在特定平台得到展示的时机和条件,侧重点是映射到信息中的来自不同特征人群的形象、声音与利益诉求,以信息为表象,以群体特征为实质。信息呈现机会不平等建立在信息内容平等的假设基础上,仅考虑特定的过滤算法技术如何影响信息呈现机会,关心信息呈现的有或无、真实或扭曲。

过滤算法成为应对海量信息的信息生态系统的必要组成部分[6]。越来越多的网络应用将算法作为新的信息守门人[67],通过限制信息访问来压制不相关信息的呈现机会[26],以此帮助人们应对信息过载问题。虽然这种方法有效减少了用户的信息接触量,但并不是所有人都了解信息是基于什么原理过滤的[25],以及信息的过滤或选中呈现过程是否遵循公平原则。研究指出,过滤算法可能会通过增加喜欢和不喜欢项目[28]、中心和边缘项目[22]之间的距离,来限制人类发现相关数据的能力,增加了信息呈现机会的不平等性,这在推荐系统中尤为常见。

除了推荐系统外,信息检索系统也是信息过滤实现的主要载体[26],由此产生的信息不平等给信息检索带来诸多挑战。检索查全率和查准率一直是信息检索研究的重点,但在过滤算法作用下,算法偏见可能会通过奖励或惩罚特定用户或文档组的形式,对系统性能产生不利影响[5]。例如,谷歌广告系统为男性推荐的高收入工作比女性更多;谷歌对CEO检索词的图像结果反馈中女性占比(11%)小于真实世界中女性CEO的占比(27%);另一个平台关于逮捕记录的查询更倾向于与非洲裔美国人名字相关联[68];TF-IDF算法以偏爱较长文档而闻名,PageRank算法更偏向发布时间更早、链接更多的页面[46]。为了描述和衡量信息检索系统的信息呈现机会不平等现象[46],阿兹帕迪(Azzopardi)等引入了可检索性的概念[5],用来描述文档是否容易被信息检索系统检索到的程度。检索性能越佳的系统,其相关文档的呈现机会越大。这意味着该系统的算法偏见现象越强,那么其他主题集受到算法惩罚的情况也就越严重[46],对多元化信息的公平创造和均等传播有负面影响。

(2)信息话语权不平等

信息话语权不平等建立在信息内容本身不平等的假设上,认为信息内容的呈现受到社会分层的影响,表现为网络空间的信息话语权不平等,以发声权利为表象,以阶层关系为实质。网络世界会优先或主要呈现社会特权阶层或优势阶层的声音,展现了社会分层结构在信息不平等呈现中的映射过程。

话语权是社会各阶层巩固阶层存在和维护阶层利益的工具,各阶层的话语权存在强弱之分[69]。网络空间也是如此。越来越多的算法偏见研究关注到了网络话语权不平等的现象,开始反思信息内容的分布结构代表了哪些阶层的利益,以及哪些要素影响算法对话语权分配的公平性。从流程上看,信息呈现不平等源自数据集偏见。相较于特权较低者,特权较高者的观点和行为更容易被数据集捕获[20],由此训练

得到的算法结果主要体现特权阶层的话语权。从本质上看,信息呈现不平等源自社会偏见,是权力阶层抢夺统治权的结果。

根据已有文献,这类信息不平等通常借助性别多样性(女性或性少数群体)、职业多样性(职业性别隔离)、种族多样性(特定的种族主义)等特征得以体现。①性别多样性导致的信息不平等表现为男性对其他性别的话语控制。通过构建网络空间中男强女弱的性别秩序和性别—能力规范[6],控制女性和性少数群体[60]的话语权,以此巩固男性统治地位。②职业多样性导致的信息不平等表现为职业性别隔离,通过单词嵌入模型将男性和女性集中在不同的职业之中。例如国王、医生、程序员、工程师等通常与男性单词相关联[70,71],而缝纫、保姆、护士等通常与女性单词相关联[47,71]。即使是中性职业(如警察、空中乘务员等),一旦嵌入了特定性别后,便开始积累性别偏见并排斥性别秩序之外的群体[71]。③种族多样性导致的信息不平等表现为白人对非白人[30,55,60]、不同种族[52]之间的权利争夺,导致信息内容的构成呈现对种族特权群体的积极话语,或对特权较低群体的消极话语,或剥夺特权较低群体的话语权利。由此可见,信息呈现不平等的结果将为某一社会群体创造优势。

本文从算法过滤的公平性和信息内容表现的公平性出发,将信息呈现不平等细分为信息呈现机会不平等和信息话语权不平等。前者侧重描述算法偏见对信息可见性的影响,后者侧重描述算法偏见对信息内容构成的影响。现如今,越来越多的企业和政府利用大数据提供产品服务和社会治理,如果在决策制定过程中不考虑以上两类信息呈现不平等形式的存在可能,那么由此产生的行动便有可能加剧社会不平等[20]。

4.2 信息与用户:信息分布不均衡

信息分布描述的是信息通过科学交流系统、大众交流系统、邮政邮电系统等社会交流系统传递、传播到不同特征人群的结果[72],是传播学话语中信息分发的结果表现。随着算法逐渐取代人类成为信息分布结果实现的守门人,信息分发模式从人工编辑主导的媒体型分发、社交链传播形成的关系型分发演进为智能算法决定的信息和人匹配的算法型分发模式[73]。因此,在信息与用户这对关系中,侧重描述用户线上信息资源的分布不均衡状况,以及信息分布不均衡会造成哪些不平等的社会效应。此时,算法仅作为信息分布结构实现的技术手段,不考虑技术可能带来的其他影响。

个性化算法型分发的工作原理是通过个性化推荐算法实现信息与人的精准匹配和推送[74],优先推送用户感兴趣的信息,同时过滤掉价值较低的信息[6]。在实际操作中,个性化推荐需要基于用户隐私数据的监控、分析和整合[75],当推荐系统无法以一致性能的算法公平地满足不同用户群体的信息偏好时[29],便可认为信息分布产生了不均衡现象。研究表明,流行性偏见是造成信息分布不均衡的重要因素[29],主要体现在以下两方面:①随着时间的推移,用户在转向更流行项目过程中会出现不同用户群体的同质化[24,29],从而导致异质信息的交流传播难度加大,符合各类用户群体差异性偏好的信息在其间的分布更加不均衡;②体现优势群体对非优势群体的偏好控制,导致用户群体间获取信息的差距缩小并造成特定信息领域的同质化[33,62],比如 YouTube 对德国政治信息的推荐与传播[33]、俄罗斯搜索引擎 Yandex 在特定政治事件上的推荐[62]等均呈现出用户获得的信息被同质化的特征。除此之外,当推荐系统过度满足用户信息需求时,也会产生其他类型的信息分布不均衡现象——过滤气泡(filter bubbles)与回音室(echo chambers)。

过滤气泡和回音室都是描述信息在算法个性化推荐后给用户造成的社会效应的学术概念,与此相似的概念还有信息茧房(information cocoons)。上述概念经常发生混淆[74,76]。其中,信息茧房强调用户对信息的主动选择,过滤气

泡强调用户对信息的被动接收,回音室强调同质化信息对用户意见加强和群体意见极化的作用[77]。由于信息分布是从中立的角度对信息交流、传播效果的客观评价,达到特定信息分布结构的信息分发强调信息推荐和算法技术的主动性,因此本文使用过滤气泡和回音室效应来描述这类信息分布不均衡现象。过滤气泡概念最早由伊莱·帕里泽(Eli Pariser)提出,他将新一代网络过滤器通过算法推荐生成用户画像,并将通过算法技术为每个人打造的独特信息世界定义为过滤气泡[78]。在过滤气泡中,算法平台驱使着人们选择或靠近观点一致的信息[33],导致过滤气泡内群体的观点趋于同质化(homophily)[36]。也就是说,过滤气泡会限制用户对多元信息、观点的接触和交流[27,45],造成信息窄化。对此,凯斯·R. 桑斯坦(Cass R. Sunstein)认为过度的个性化可能会导致用户的"在线隔离"(online segregation)[79],即用户会被具有相似特征或意识形态的人所包围,这种现象便是回音室[80]。在回音室内部,用户频繁地接触与自身观点相似的同质化信息,促使回音室内形成一致的群体意见[18],并不断扩大与回音室外观点群体的距离[28],表现出观点的两极分化[34,81]。随着时间的推移,回音室内部观点被不断巩固,回音室内群体对其他观点变得异常排斥,最终导致群体观点的极化[33,36]。过滤气泡与回音室的启示在于,具有一定相同或相似特征的人群被算法依据信息及其包含观点的偏好圈在气泡或者回音室中,从而导致这些人群被狭隘的信息内容与极端的观点所控制,不同过滤气泡的信息之间、不同回音室之间存在信息流动的壁垒,同质信息过于集中在特定的内部结构中,无法实现不同类型信息的均衡分布。

4.3　技术与用户:新型数字不平等

算法偏见的后果主要表现在某些特征的用户在算法控制的平台得到不公平的对待,尤其是平台各利益方对这些用户形成的观点、态度和行为中带有偏见。此类偏见映射到技术与用户的关系中形成了同类技术面前不同用户有差异的自我呈现、话语权和诉求表达效果,并直接作用于用户的线下资源获取、生产与生活水平。算法偏见不仅加深了传统数字不平等,还催生了新型的数字不平等。

在传统数字不平等研究中,通常使用数字鸿沟这一概念来描述数字时代的技术两极分化现象。尽管经历了从接入沟[82,83]到使用沟[84]和知识沟[85,86]的学术焦点演进历程,数字鸿沟研究一直持续聚焦于那些无法接入和使用 ICT 资源的人群,关心这类群体与拥有 ICT 技术接入和利用机会人群之间的差距。算法偏见研究也覆盖了这类传统数字鸿沟中处在"错误一端"的人群,例如纽约 311 平台实现城市服务分配[39]的算法在停车计时器、人行道、街道、坑洼和交通信号灯等公共交通基础设施维修请求业务中,遗漏了低收入人群、老年人、女性、非洲裔美国人、西班牙裔美国人和有孩子的家庭等人群对上述问题的报告,在现实生活中沿袭了数字鸿沟中这些人群的劣势地位;波士顿市道路坑洞检测手机软件 StreetBump[87]的众包工作方式与算法对缺乏 ICT 接入和使用机会的群体不友好,没有移动手机的低收入社区居民也没有机会和条件利用移动设备报告社区内外道路的坑洼,道路数据遗漏了经济弱势群体的诉求,在数字世界中的弱势地位被复制到道路基础设施代表的物理世界中。算法作为传统数字鸿沟的"帮凶",进一步弱化了数字贫困者的社会地位和物质生活水平。

算法催生的新型数字不平等集中表现在扩大了数字贫困人群的范畴和数字鸿沟的维度,形成算法鸿沟(algorithms divide)[88]61。数字贫困者不仅仅是缺乏 ICT 技术接入与使用的人群,而且还演化出了更多受到算法压迫的弱势群体。算法曾被寄予厚望——能够破除存在已久的社会不平等现象,有望形成新的正义——算法正义。这样的期待也顺理成章地促使社会实现公平正义的权力逐步让渡于算法[89],越来越多的算法被用于代替或辅佐人类完成重要社会

物品分配和自动化决策,如医疗保健服务[42]、就业机会[66]、城市公共服务[39]等,出现了各类针对人类识别[40]、分类[42]和管理[23]的算法应用。与人们期待相反的现实是,越来越多的证据显示了算法偏见的存在,以及它对社会资源分配的影响,使得算法偏见的负面影响波及了更为广泛的群体。除了传统数字鸿沟所关心的ICT接入和使用困难的群体之外,还包括没有足够技能逃避算法建议的群体,以及因个人或社会特征而受到算法惩罚的群体[88]41。显然算法偏见对用户的影响已经超过了传统数字不平等中弱势人群的研究范围,有必要将以上三类人群全部纳入研究视野,这些群体共同拥有一个新的身份——数字底层阶级(digital underclass),类似于笔者曾经提出并系统验证过的数字化贫困群体[90],但在测量维度上拓展了此前数字化贫困的研究,算法成为与数字化设备、服务等类似的可以决定数字化位置的标准。

算法偏见对传统与新型数字不平等中弱势群体的影响会进一步传导到这些群体所处的社会结构和现实世界中,进一步强化已有的传统不平等问题。具体地说,算法偏见结果可能会通过以下社会不平等要素得以体现:①性别差异导致用户获得不对等的就业机会和广告服务机会[66],削弱了女性获得权力提升的可能,对女性的算法帮助(如增加补偿因子改善对算法性别偏见)似乎加深了女性面临的性别不平等[91];②种族差异不仅会导致用户获得有差异的医疗保健服务和护理机会[42],也会导致用户获得不对等的刑事司法对待[2],伤害了公民获得同等健康服务以及其他的基本人身权利;③地理差异导致用户获得不同的APP服务[17,37],拉大了城乡间或高、低收入社区间群体的数字鸿沟;④代际差异导致未成年人或老年人获得有别于普通成年人的算法决策结果,从而影响对特殊人群的可用产品服务和救助救治效果[21,37,53]。性别、种族、地理和代际特征层面的算法偏见导致原先处于社会弱势地位的群体更有可能被数字排斥,算法偏见的结果继承并强化了原有的社会、文化、个人、经济和政治的不平等。

由此推断,数字底层阶级与传统的社会弱势群体是密切相关的,算法偏见的负面影响实际上就是社会偏见和权力结构下的社会弱势群体通过数字化呈现映射为数字底层阶级,算法是实现这种映射的工具。通过算法技术的隐性继承和强化,社会不平等现象从线下转移至线上,再借由算法的不公平分配实现线下资源的再分配,加剧线下资源的马太效应。在这个过程中,社会弱势群体在算法身份重构中不断重复其社会固有地位,从而限制了个体或资源在结构框架外的流动[92],最终完成线下资源的进一步掠夺和社会结构不平等的强化,体现出算法对社会分层和社会分配不平等的催化和控制。

4.4 算法偏见对信息领域不平等话语的贡献

算法偏见赋予了前人工智能时代产生的传统学术话语信息不平等与数字不平等新的内涵与外延。传统信息不平等更多关注社会主体在有效查询与获取外部信息过程中形成的非均衡分布状态,甚至是信息两极分化状况,其所对应的外延大多圈定在信息服务机构在与用户实现信息传递与交流的过程中的各类制度设计、主客体关系、行为规律和影响后果。如图书馆等传统公益信息服务机构作为信息源对处于不同时空范围、多元化智识水平和差别化动机下用户的分布不均匀与服务不均等;数据库商和互联网服务商等逐利的资本力量在过度市场化、政商联盟化的信息社会发展趋势中形成市场垄断,侵蚀公益信息空间,从而导致普通公众的基本信息权益受到影响。算法偏见将信息社会中有意或无意创造信息不平等的主体,从公益信息服务机构与商业信息机构等肉眼可见的自然人与法人概念等,扩展到了人类智能与人工智能混合驱动下的各类公共治理机制与信息商业力量,算法偏见甚至将盲目追随商业热潮的公益信息机构也卷入到信息不平等的洪流中,扩

大了人工智能条件催生的信息弱势群体范畴。算法偏见视野中,信息不平等解释的核心活动范围从有效查询与有效获取扩展到了信息自主创造与呈现环节,后者越来越显著地左右着社会主体查询与获取信息的边界与方向。

传统数字不平等话语致力于探究智能化程度处于相对低级阶段的电脑等信息与通信技术复制传统社会结构到新兴数字化社会中的机理、过程、结果与根源。从认知的整体性和深刻度来说,相较于数字鸿沟过度关注技术应用于人群与社会中两极分化的表象,数字不平等更多地将数字化社会的资源分配不均问题归根于宏观的社会结构问题及其背后的阶层关系,人类社会进化中优胜劣汰规则通过信息与通信技术进一步催生了对个人利益的过度追逐及更大程度的社会不平等。算法作为拓展自然人能力极限的人工智能典型,将数字化社会中的阶层关系固化得更加不对等,阶层间的偏见被指数级地嵌入到数字化资源分配技术中,其自动化程度更是前所未有,数字化的"离心力"也日趋显著,原本在数字化社会中层与底层边缘地带徘徊的信息中产偏弱群体被进一步底层化,原处于底层的信息弱势群体在算法偏见的排他性作用下无限逼近绝对的数字贫困境地。

在算法盛行的人工智能时代,信息职业人员面临新的、更富挑战的信息不平等与数字不平等现象,长期致力于信息公平与数字公平价值观的图书馆职业人员是与可能偏航甚至是遇险的数字化航船抗争的主流力量,其所推崇的信息自主创造和多元呈现、信息的有效查询与有效获取核心技能,可以重新分配社会主体在数字化航船上的分布,让更多的自然人和人工智能人处在船头、船舱和船尾相对均衡的状态,让更多的数字化信息流动到有利于数字化航船前进的方向上,将自然人之间的偏见禁锢在人类理智之内。

5　总结

本文采用系统综述方法分类汇总算法偏见实证研究中的概念、理论基础和后果,并对各主题进行批判式解释综合。结果显示,算法偏见是指在计算机实验全流程中,对某些个人或群体、信息内容等产生不公平结果的系统性和可重复性错误,其内涵可分解为偏见来源、偏见对象、偏见后果、偏见特性四要素。根据内涵要素,从偏见来源和偏见对象两条路径扩展算法偏见的外延内容,构建算法偏见的概念框架。进一步梳理算法偏见问题反思的哲理基础,探讨功能主义、冲突论、还原论、马克思主义与算法偏见之间的关系。最后本文试图从信息科学视角阐述算法偏见的后果,讨论算法偏见在技术、信息、用户三要素互动过程中的存在形式,发现三类不平等后果——信息呈现不平等、信息分布不均衡和新型数字不平等,并提炼了算法偏见对信息不平等与数字不平等的可能贡献和突破。

本文尚存在不足之处:①在检索策略方面,由于学界对算法偏见的定义范围尚未统一,使得检索过程可能遗漏了部分检索词,如 AI 偏见、分析偏见等;②在研究结论方面,本文对算法偏见的哲理反思认识较为浅薄,对于所构建的算法偏见概念框架的合理性和完备性,有待未来实证检验。

参考文献

[1] THOMASIAN N M,EICKHOFF C,ADASHI E Y. Advancing health equity with artificial intelligence[J]. Journal of Public Health Policy,2021,42(4):1-10.

[2] SILVA JÚNIOR J J,SOARES A S. Improving face recognition accuracy for Brazilian faces in a criminal investigation department[C]//9th Brazilian Conference on Intelligent Systems. Rio Grande,RS,Brazil,2020:287-301.

[3] 黄晓伟,李育慧. 算法偏见问题的技术——权力互构论解析[J]. 理论与现代化,2021(1):39-48.

(HUANG X W, LI Y H. An analysis of the problem of algorithmic bias based on the co-production theory of technology and power[J]. Theory and Modernization, 2021(1): 39-48.)

[4] 岳平,苗越. 社会治理:人工智能时代算法偏见的问题与规制[J]. 上海大学学报(社会科学版),2021,38(6):1-11. (YUE P, MIAO Y. Social governance: problems and regulation of algorithm bias in the age of AI[J]. Journal of Shanghai University (Social Sciences Edition), 2021, 38(6): 1-11.)

[5] WILKIE C, AZZOPARDI L. The impact of fielding on retrieval performance and bias[C]//Proceedings of the Association for Information Science and Technology. New Jersey: John Wiley & Sons, 2018, 55(1): 564-572.

[6] OTTERBACHER J, BATES J, CLOUGH P. Competent men and warm women: gender stereotypes and backlash in image search results[C]//Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. New York, USA, 2017: 6620-6631.

[7] GARDNER C C. Teaching algorithmic bias in a credit-bearing course[J]. International Information & Library Review, 2019, 51(4): 321-327.

[8] DOMAN M, GARRISON C. Introducing algorithmic bias considerations in an introductory CS course[J]. Journal of Computing Sciences in Colleges, 2021, 37(5): 31-42.

[9] OLTEANU A, CASTILLO C, DIAZ F, et al. Social data: biases, methodological pitfalls, and ethical boundaries[J]. Frontiers in Big Data, 2019(2): 13.

[10] FRIEDMAN B, NISSENBAUM H. Bias in computer systems[J]. ACM Transactions on Information Systems (TOIS), 1996, 14(3): 330-347.

[11] CRAMER H, GARCIA-GATHRIGHT J, REDDY S, et al. Translation, tracks & data: an algorithmic bias effort in practice[C]//Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems. Glasgow, Scotland, UK, 2019: 1-8.

[12] PAGE M J, MCKENZIE J E, BOSSUYT P M, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews[J]. Systematic Reviews, 2021, 10(1): 1-11.

[13] AZZALINI F, CRISCUOLO C, TANCA L. FAIR-DB: function AI dependencies to discover data bias[C/OL]// Proceedings of the Workshops of the EDBT/ICDT 2021 Joint Conference, 2021[2022-02-18]. http://ceur-ws.org/Vol-2841/PIE+Q_4.pdf.

[14] DANKS D, LONDON A J. Algorithmic bias in autonomous systems[C]//26th International Joint Conference on Artificial Intelligence. Melbourne, Australia, 2017: 4691-4697.

[15] CROWE M. Crowe Critical Appraisal Tool (CCAT) user guide[EB/OL]. [2022-07-15]. https://conchra.com.au/wp-content/uploads/2015/12/CCAT-user-guide-v1.4.pdf.

[16] PANEERCHELVAM P T, MARUTHAVEERAN S, MAULAN S, et al. The use and associated constraints of urban greenway from a socioecological perspective: a systematic review[J]. Urban Forestry & Urban Greening, 2020, 47: 126508.

[17] PANDEY A, CALISKAN A. Disparate impact of artificial intelligence bias in ride-hailing economy's price discrimination algorithms[C]//Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. USA, 2021: 822-833.

[18] PERALTA A F, KERTÉSZ J, IÑIGUEZ G. Opinion formation on social networks with algorithmic bias: dynamics and bias imbalance[J]. Journal of Physics: Complexity, 2021, 2(4): 045009.

[19] ABUL-FOTTOUH D, SONG M Y, GRUZD A. Examining algorithmic biases in YouTube's recommendations of vaccine videos[J]. International Journal of Medical Informatics, 2020, 140: 1-6.

[20] HARGITTAI E. Potential biases in big data: omitted voices on social media[J]. Social Science Computer Review, 2020, 38(1): 10-24.

[21] SRINIVAS N, RICANEK K, MICHALSKI D, et al. Face recognition algorithm bias: performance differences on images of children and adults[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Long Beach, USA, 2019: 2269-2277.

[22] FLEXER A,DÖRFLER M,SCHLÜTER J,et al. Hubness as a case of technical algorithmic bias in music recommendation[C]//2018 IEEE International Conference on Data Mining Workshops(ICDMW). Singapore,2018:1062-1069.

[23] AKTER S,MCCARTHY G,SAJIB S,et al. Algorithmic bias in data-driven innovation in the age of AI[J]. International Journal of Information Management,2021,60:1-13.

[24] MANSOURY M,ABDOLLAHPOURI H,PECHENIZKIY M,et al. Feedback loop and bias amplification in recommender systems[C]//Proceedings of the 29th ACM International Conference on Information & Knowledge Management. Ireland,2020:2145-2148.

[25] MORSTATTER F,LIU H. Discovering,assessing,and mitigating data bias in social media[J]. Online Social Networks and Media,2017(1):1-13.

[26] SUN W,NASRAOUI O,SHAFTO P. Evolution and impact of bias in human and machine learning algorithm interaction[J]. PloS One,2020,15(8):e0235502.

[27] LIN K,SONBOLI N,MOBASHER B,et al. Crank up the volume:preference bias amplification in collaborative recommendation[C/OL]// RMSE Workshop in 13th ACM Conference on Recommender Systems(RecSys). Copenhagen,Denmark,2019.(2019-09-13)[2022-03-16]. https://doi.org/10.48550/arXiv.1909.06362.

[28] SUN W,NASRAOUI O,SHAFTO P. Iterated algorithmic bias in the interactive machine learning process of information filtering[C]//10th International Conference on Knowledge Discovery and Information Retrieval. Seville,Spain,2018:108-116.

[29] ABDOLLAHPOURI H,MANSOURY M,BURKE R,et al. The connection between popularity bias,calibration,and fairness in recommendation[C]// Proceedings of the 14th ACM Conference on Recommender Systems. Brazil,2020:726-731.

[30] METAXA D,GAN M A,GOH S,et al. An image of society:gender and racial representation and impact in image search results for occupations[J]. Proceedings of the ACM on Human-Computer Interaction,2021,5(CSCW1):1-23.

[31] WANG N,CHEN L. User bias in beyond-accuracy measurement of recommendation algorithms[C]//Proceedings of the 15th ACM Conference on Recommender Systems. Amsterdam,Netherlands,2021:133-142.

[32] BONEZZI A,OSTINELLI M. Can algorithms legitimize discrimination?[J]. Journal of Experimental Psychology:Applied,2021,27(2):447-459.

[33] HEUER H,HOCH H,BREITER A,et al. Auditing the biases enacted by YouTube for political topics in Germany[C]//Proceedings of Mensch und Computer 2021. Ingolstadt,Germany,2021:456-468.

[34] PERALTA A F,NERI M,KERTÉSZ J,et al. Effect of algorithmic bias and network structure on coexistence,consensus,and polarization of opinions[J]. Physical Review E,2021,104(4):044312.

[35] AYSOLMAZ B,IREN D,DAU N. Preventing algorithmic bias in the development of algorithmic decision-making systems:a Delphi study[C]//Proceedings of the 53rd Hawaii International Conference on System Sciences. Hawaii,USA,2020:5267-5276.

[36] SÎRBU A,PEDRESCHI D,GIANNOTTI F,et al. Algorithmic bias amplifies opinion fragmentation and polarization:a bounded confidence model[J]. PLoS One,2019,14(3):e0213246.

[37] DÍAZ M,JOHNSON I,LAZAR A,et al. Addressing age-related bias in sentiment analysis[C]//Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. Montreal,Canada,2018:1-14.

[38] FAZELPOUR S,DANKS D. Algorithmic bias:senses,sources,solutions[J]. Philosophy Compass,2021,16(8):e12760.

[39] KONTOKOSTA C E,HONG B. Bias in smart city governance:how socio-spatial disparities in 311 complaint behavior impact the fairness of data-driven decisions[J]. Sustainable Cities and Society,2021,64:1-10.

[40] TAATI B,ZHAO S,ASHRAF A B,et al. Algorithmic bias in clinical populations—evaluating and improving facial analysis technology in older adults with dementia[J]. IEEE Access,2019(7):25527-25534.

[41] CUNNINGHAM P, DELANY S J. Underestimation bias and underfitting in machine learning[C]//International Workshop on the Foundations of Trustworthy AI Integrating Learning, Optimization and Reasoning. Santiago de Compestela, Spain, 2020: 20-31.

[42] OBERMEYER Z, POWERS B, VOGELI C, et al. Dissecting racial bias in an algorithm used to manage the health of populations[J]. Science, 2019, 366(6464): 447-453.

[43] VANMASSENHOVE E, SHTERIONOV D, GWILLIAM M. Machine translationese: effects of algorithmic bias on linguistic complexity in machine translation[C]// Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, 2021: 2203-2213.

[44] BARTLEY N, ABELIUK A, FERRARA E, et al. Auditing algorithmic bias on Twitter[C]//13th ACM Web Science Conference. UK, 2021: 65-73.

[45] BORATTO L, FENU G, MARRAS M. The effect of algorithmic bias on recommender systems for massive open online courses[C]//European Conference on Information Retrieval. Cham: Springer, 2019: 457-472.

[46] WILKIE C, AZZOPARDI L. Algorithmic bias: do good systems make relevant documents more retrievable?[C]// Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. Singapore, 2017: 2375-2378.

[47] LEAVY S, MEANEY G, WADE K, et al. Mitigating gender bias in machine learning data sets[C]//International Workshop on Algorithmic Bias in Search and Recommendation. Lisbon, Portugal, 2020: 12-26.

[48] COLLEY A, THEBAULT-SPIEKER J, LIN A Y, et al. The geography of Pokémon GO: beneficial and problematic effects on places and movement[C]//Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. New York, USA, 2017: 1179-1192.

[49] DRAUDE C, KLUMBYTE G, LÜCKING P, et al. Situated algorithms: a sociotechnical systemic approach to bias [J]. Online Information Review, 2020, 44(2): 325-342.

[50] LEE N T. Detecting racial bias in algorithms and machine learning[J]. Journal of Information, Communication and Ethics in Society, 2018, 16(3): 252-260.

[51] YARGER L, PAYTON F C, NEUPANE B. Algorithmic equity in the hiring of underrepresented IT job candidates [J]. Online Information Review, 2020, 44(2): 383-395.

[52] GUO W, CALISKAN A. Detecting emergent intersectional biases: contextualized word embeddings contain a distribution of human-like biases[C]//Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. USA, 2021: 122-133.

[53] BALAKRISHNAN G, XIONG Y, XIA W, et al. Towards causal benchmarking of bias in face analysis algorithms [C]// European Conference on Computer Vision. Cham: Springer, 2020: 547-563.

[54] MATSANGIDOU M, OTTERBACHER J. What is beautiful continues to be good: people images and algorithmic inferences on physical attractiveness[C]//Human-Computer Interaction-INTERACT 2019. Paphos, Cyprus, 2019: 243-264.

[55] ARAÚJO C S, MEIRA W, ALMEIDA V. Identifying stereotypes in the online perception of physical attractiveness [C]//International Conference on Social Informatics. Bellevue, WA, USA, 2016: 419-437.

[56] BARLAS P, KYRIAKOU K, GUEST O, et al. To "see" is to stereotype: image tagging algorithms, gender recognition, and the accuracy-fairness trade-off[J]. Proceedings of the ACM on Human-Computer Interaction, 2021, 4 (CSCW3): 1-31.

[57] 闫泽华. 内容算法: 把内容变成价值的效率系统[M]. 北京: 中信出版社, 2018: 28-30. (YAN Z H. Content-based algorithms: efficiency systems that turn content into value[M]. Beijing: CITIC Press, 2018: 28-30.)

[58] 于海. 西方社会思想史[M]. 上海: 复旦大学出版社, 2010. (YU H. A history of western sociological thought: from Plato to Bourdieu[M]. Shanghai: Fudan University Press, 2010.)

[59] 肖富群. 社会学视野中的社会性别理论[J]. 江淮论坛, 2010(1): 134-139. (XIAO F Q. Gender theory in the perspective of sociology[J]. Jianghuai Tribune, 2010(1): 134-139.)

[60] BOGERS L,NIEDERER S,BARDELLI F,et al. Confronting bias in the online representation of pregnancy[J]. Convergence,2020,26(5-6):1037-1059.

[61] 陈江进,郭琰. 心身问题解决的新尝试:机器功能主义[J]. 自然辩证法通讯,2003(4):28-33,110. (CHEN J J,GUO Y. The new attempt to solve mind-body problem:machine functionalism[J]. Journal of Dialectics of Nature,2003(4):28-33,110.)

[62] KRAVETS D,TOEPFL F. Gauging reference and source bias over time:how Russia's partially state-controlled search engine Yandex mediated an anti-regime protest event[J/OL]. Information,Communication & Society,2021:1-17[2022-07-15]. https://doi.org/10.1080/1369118X.2021.1933563.

[63] Internet encyclopedia of philosophy:reductionism[EB/OL].[2022-02-18]. https://iep.utm.edu/red-ism/.

[64] 孟捷. 经济人假设与马克思主义经济学[J]. 中国社会科学,2007(1):30-42,205. (MENG J. The hypothesis of economic man and Marxist economics[J]. Social Sciences in China,2007(1):30-42,205.)

[65] 程恩富. 现代马克思主义政治经济学的四大理论假设[J]. 中国社会科学,2007(1):16-29,205. (CHENG E F. Four theoretical hypotheses of the political economy of modern Marxism[J]. Social Sciences in China,2007(1):16-29,205.)

[66] LAMBRECHT A,TUCKER C. Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads[J]. Management Science,2019,65(7):2966-2981.

[67] 杨洸,佘佳玲. 算法新闻用户的数字鸿沟:表现及影响[J]. 现代传播(中国传媒大学学报),2020,42(4):145-154. (YANG G,SHE J L. The digital divide for algorithmic news users:performance and impact[J]. Modern Communication(Journal of Communication University of China),2020,42(4):145-154.)

[68] KIRKPATRICK K. Battling algorithmic bias:how do we ensure algorithms treat us fairly?[J]. Communications of the ACM,2016,59(10):16-17.

[69] 赵云泽,付冰清. 当下中国网络话语权的社会阶层结构分析[J]. 国际新闻界,2010,32(5):63-70. (ZHAO Y Z,FU B Q. An analysis of the social class structure of the network discourse power in China[J]. Chinese Journal of Journalism & Communication,2010,32(5):63-70.)

[70] ROZADO D. Wide range screening of algorithmic bias in word embedding models using large sentiment lexicons reveals underreported bias types[J]. PLoS One,2020,15(4):e0231189.

[71] JAIN S,DEY S,BAJAJ P,et al. A data-centric approach towards deducing bias in artificial intelligence systems for textual contexts[C]//CCF International Conference on Natural Language Processing and Chinese Computing. Qingdao,China,2021:323-333.

[72] 于良芝. 图书馆情报学概论[M]. 北京:国家图书馆出版社,2016:123-131. (YU L Z. Introduction to library and information science[M]. Beijing:National Library of China Publishing House,2016:123-131.)

[73] 喻国明,韩婷. 算法型信息分发:技术原理、机制创新与未来发展[J]. 新闻爱好者,2018(4):8-13. (YU G M,HAN T. Algorithm-based information distribution:technical principles,mechanism innovation and future development[J]. Journalism Lover,2018(4):8-13.)

[74] 姜婷婷,许艳闰. 国外过滤气泡研究:基础、脉络与展望[J]. 情报学报,2021,40(10):1108-1117. (JIANG T T,XU Y R. Filter bubbles—induced by personalized recommendation algorithms:a review of related research[J]. Journal of the China Society for Scientific and Technical Information,2021,40(10):1108-1117.)

[75] 郝喜. 数字化"圆形监狱":算法监控的规训与惩罚[J]. 昆明理工大学学报(社会科学版),2021,21(6):39-45. (HAO X. Digital "panopticon":discipline and punishment of algorithm monitoring[J]. Journal of Kunming University of Science and Technology(Social Sciences),2021,21(6):39-45.)

[76] 姜婷婷,许艳闰. 窄化的信息世界:国外信息茧房、选择性接触与回音室研究进展[J]. 图书情报知识,2021,38(5):134-144. (JIANG T T,XU Y R. Narrowed information universe:a review of research on information cocoons,selective exposure,and echo chambers[J]. Documentation,Information & Knowledge,2021,38(5):134-144.)

[77] 刘强,赵茜. 算法中选择的同化与异化——国外回音室效应研究20年述评与展望[J]. 新闻界,2021(6):

29-38. (LIU Q, ZHAO X. Assimilation and alienation in the algorithm of choice—review and prospect of the research on echo chamber effect abroad[J]. Journalism and Mass Communication, 2021(6): 29-38.)

[78] 伊莱·帕里泽. 过滤泡:互联网对我们的隐秘操纵[M]. 方师师,杨媛,译. 北京:中国人民大学出版社, 2020:8-9. (PARISER E. The filter bubble: what the Internet is hiding from you[M]. FANG S S, YANG Y, trans. Beijing: Renmin University of China Press, 2020: 8-9.)

[79] SUNSTEIN C R. #Republic: divided democracy in the age of social media[M]. New Jersey: Princeton University Press, 2017: 112-113.

[80] 凯斯·R. 桑斯坦. 信息乌托邦:众人如何生产知识[M]. 毕竞悦,译. 北京:法律出版社, 2008:6-7. (SUNSTEIN C R. Infotopia: how many minds produce knowledge[M]. BI J Y, trans. Beijing: Law Publishing House, 2008: 6-7.)

[81] BLEX C, YASSERI T. Positive algorithmic bias cannot stop fragmentation in homophilic networks[J]. The Journal of Mathematical Sociology, 2020, 46(1): 1-18.

[82] VAN DIJK J A G M. The deepening divide: inequality in the information society[M]. Thousand Oaks: Sage Publications, 2005: 27-94.

[83] VAN DIJK J A G M. Digital divide research, achievements and shortcomings[J]. Poetics, 2006, 34(4-5): 221-235.

[84] VAN DIJK J A G M, HACKER K. The digital divide as a complex and dynamic phenomenon[J]. The Information Society, 2003, 19(4): 315-326.

[85] VAN DEURSEN J A M, HELSPER E J. A nuanced understanding of Internet use and non-use among the elderly [J]. European Journal of Communication, 2015, 30(2): 171-187.

[86] 闫慧,孙立立. 1989年以来国内外数字鸿沟研究回顾:内涵、表现维度及影响因素综述[J]. 中国图书馆学报, 2012, 38(5): 82-94. (YAN H, SUN L L. Digital divides revisited: a review on definitions, dimensions and independent variables(1989-2012)[J]. Journal of Library Science in China, 2012, 38(5): 82-94.)

[87] PARK B, RAO D L, GUDIVADA V N. Dangers of bias in data-intensive information systems[C]//Next Generation Information Processing System, 2021: 259-271.

[88] RAGNEDDA M. Enhancing digital equity: connecting the digital underclass [M]. Switzerland: Springer Nature, 2020.

[89] 杰米·萨斯坎德. 算法的力量:人类如何共同生存?[M]. 李大白,译. 北京:北京日报出版社, 2022:231. (SUSSKIND J. Future politics: living together in a world transformed by tech?[M]. LI D B, trans. Beijing: Beijing Daily Press, 2022: 231.)

[90] 闫慧. 农民数字化贫困的结构性成因分析[J]. 中国图书馆学报, 2017, 43(2): 24-39. (YAN H. Structural origins of digital poverty in rural China[J]. Journal of Library Science in China, 2017, 43(2): 24-39.)

[91] COWGILL B, DELL'ACQUA F, DENG S, et al. Biased programmers? Or biased data? A field experiment in operationalizing AI ethics[C]//Proceedings of the 21st ACM Conference on Economics and Computation. Hungary, 2020: 679-681.

[92] 彭兰. 假象、算法囚徒与权利让渡:数据与算法时代的新风险[J]. 西北师大学报(社会科学版), 2018, 55(5): 20-29. (PENG L. Illusion, prisoner of algorithm, and transfer of rights: the new risks in the age of data and algorithm[J]. Journal of Northwest Normal University(Social Sciences), 2018, 55(5): 20-29.)

贾诗威 中国人民大学信息资源管理学院博士研究生。北京 100872。

闫 慧 中国人民大学信息资源管理学院教授,博士生导师。北京 100872。

(收稿日期:2022-06-03;修回日期:2022-09-06)