政策规划

# 中国算法公平的研究热点与拓展空间

## ——基于CiteSpace知识图谱的分析

周晨阳[1]，曹翰阳[1,2]

（1. 扬州大学马克思主义学院，扬州 225009；2. 扬州大学政府治理与公共政策研究中心，扬州 225127）

**摘要**：算法公平对社会发展起到关键的推动作用。以CNKI（中国知网）1997—2025年的文献为数据来源，利用可视化工具CiteSpace发现：中国算法公平研究自1997年以来呈现显著波动，后续研究将持续增长；研究热点涵盖算法偏见、算法歧视等“去歧视化”维度。纵观研究脉络，相关研究经历了探索阶段、调整阶段和复苏阶段，前沿热点与研究拓展具有鲜明的问题导向特征，体现了多元、聚焦的发展趋势。展望未来，要建立跨学科合作机制、构建科学共同体、更新社会治理体系。

**关键词**：算法公平；知识图谱；研究热点；前沿趋势

**中图分类号**：C93-05；TP18 **文献标志码**：A **文章编号**：1671－1807(2026)01－0284－12

算法作为数字社会的关键基石与核心驱动，正深刻重塑着社会运行方式。其影响力源于算法本身的内在特性。作为一组机械执行指令，算法能够针对符合规范的输入，在有限时间内生成确定的输出[1]。这本质上是将复杂问题的求解过程形式化、自动化，使其具有确定性、高效性和结果可验证性等特点。正是凭借这些特点，算法得以广泛应用在社会治理、技术赋能、决策支持等领域。

然而，算法在提升社会效率的同时，其核心弊端也日益凸显。由于算法本质上是人类思维的外化[2]，社会固有的歧视性因素可能被编码其中，导致算法难以实现“价值中立的平等”[3]，进而引发算法公平性问题。尽管中国政府出台了《互联网信息服务算法推荐管理规定》[4]等政策以规范算法应用并对其潜在歧视风险施加约束，但面对技术快速迭代所催生的新型歧视范式，现有制度供给存在显著滞后。这种制度响应与技术发展之间的动态失衡，进一步强调了深化算法公平研究的紧迫性。为系统把握该领域的研究脉络、前沿热点与发展方向，运用CiteSpace软件，旨在揭示中国算法公平研究的研究热点与拓展空间，以期为促进数字时代下的社会公平提供学理支持。

## 1 数据来源和研究方法

### 1.1 数据来源

基于CNKI（中国知网）数据库1997—2025年的文献进行检索。检索式为“主题＝（‘算法公平’）OR（‘算法公正’）OR（‘算法平等’）and 年＝1997—2025”。为了确保研究可信度并最大限度减少无关文献干扰，首先运用CiteSpace进行除重。随后，通过人工审阅标题、摘要等剔除与算法公平研究主题明显不相关的文献。最后，为进一步完善样本、避免遗漏重要文献，通过引文追溯、核心期刊审查对检索结果进行补充。截至2025年6月，共获得437篇有效文献样本。

### 1.2 研究方法

运用文献计量软件CiteSpace，结合关键词共现分析、聚类分析及突现分析等方法对样本文献中的共现、时序关系进行系统梳理。CiteSpace的核心功能在于研究领域发展图谱的绘制，同时能够直观展现相关研究全景并识别关键文献、研究热点和前沿方向[5]，具有操作简单、功能强大等优点。该工具在

**收稿日期**：2025-06-26

**基金项目**：国家社会科学基金重点项目（22AZZ008）

**作者简介**：周晨阳（2002—），男，江苏扬州人，硕士研究生，研究方向为数字治理和马克思科技思想；曹翰阳（1997—），男，江苏扬州人，博士研究生，研究方向为协商民主、公民参与。

探寻学科演化路径、分析研究潜在动力、预测未来发展趋势等方面具有重要价值[6],为研究目标提供了有效的方法论支撑。

## 2 算法公平研究现状

通过考察特定时间段内的发文量及变化趋势,可以准确把握研究领域的发展动态,为了解其研究现状提供关键线索。根据算法公平的发文量及变化情况(图 1)可知,1997—2009 年相关研究发文呈由慢到快的稳步增长态势,并于 2009 年达到峰值(35 篇);而在 2010—2021 年,尽管期间经历了 3 次短期回弹,但整体发文量还是呈下滑趋势,反映出研究可能已趋于成熟[7];2022 年至今,相关研究显现复苏迹象,发文量重拾升势。

### 2.1 算法公平研究的作者共现和核心作者测定

作为学术研究的主体,作者的研究方向集中了领域内的热点问题、核心主题等关键要素,为掌握该领域的研究动态提供了关键视角。通过将节点类型设置为“作者”,时间切片跨度设置为 1 年,得到作者合作共现网络图(图 2)。

在作者合作共现图谱中,节点代表作者,其大小与发文量与其贡献大小正相关,即节点越大,发文量越多,贡献也就越大。图 2 显示,算法公平领域的高产作者数量有限,仅有崔维嘉、巴斌(聚焦电信技术维度的公平分配算法设计)、张震(探索互联网

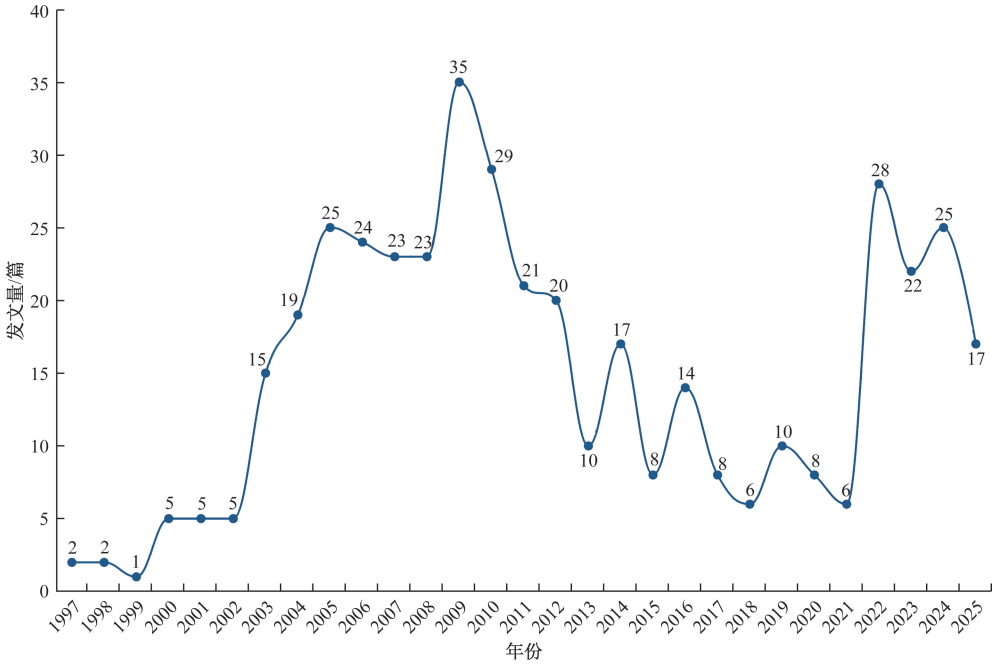



图 1 1997—2025 年算法公平发文量

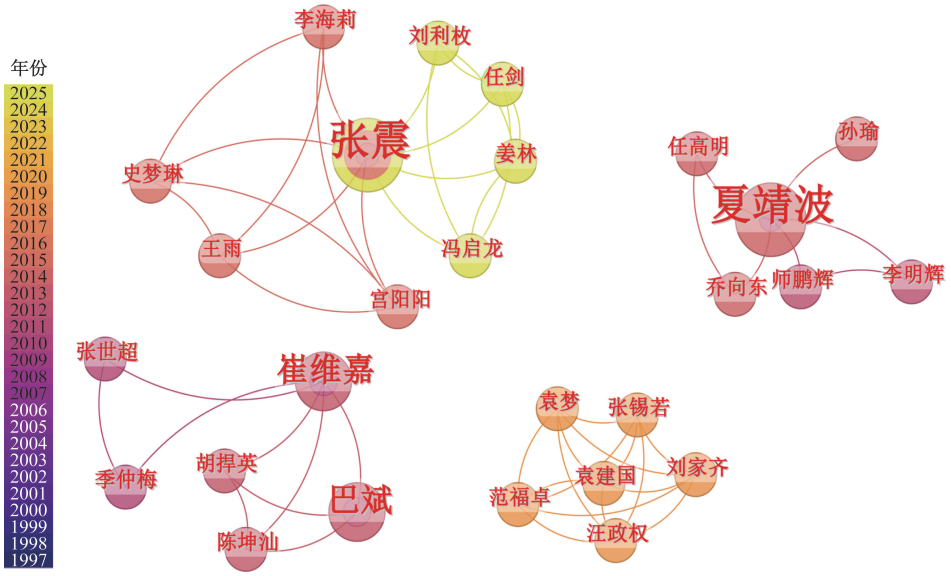



图 2 算法公平研究领域作者合作共现

技术维度的公平采样算法设计)以及夏靖波(致力于公平采样算法的自动化技术和软件支持)等学者。节点之间的连线则代表作者之间的合作关系,连线强度越大合作关系就越紧密。观察图2网络结构可见,各节点之间连线数量较少且强度普遍较弱,网络密度低,并没有形成复杂的社群结构。这表明算法公平研究领域内的作者虽然形成了一定数量的研究团队(如基于合作关系识别的张震团队或夏靖波团队),但整体上作者间合作关系稀疏、协同程度不高,可能对算法公平理论的深度整合与传播产生一定的制约。因此,为推进该领域研究的纵深发展,亟须加强学者间的交流与合作。

由于合作网络结构呈现分散性,识别并聚焦算法公平领域的核心作者就显得尤为重要。核心作者的研究包含了热点对象、主题等要素,在研究中起到了领头羊的作用[8]。因此运用普赖斯定律①对其进行测定:其中第一作者发文量最多的是黄家玮(4篇),计算得 $N\approx1.498$,取整为1,故发文量≥1的学者为核心候选。经测定指数②和影响力评估模型进一步筛选,$h$ 指数≥13的学者做出了显著学术贡献[9],结果见表1。

深入分析核心作者的研究成果可以看出,其研究主题高度集中于算法底层机制和工具、公平调度技术实现、算法社会影响和伦理挑战等领域,集中体现了技术探索与伦理管控并行的学科特征。此外经过整理统计,算法公平领域的核心作者共计发表了53篇文献,占算法公平研究领域文献总量的12.1%。普赖斯定律指出,当核心作者的发文量占到总数量一半时,才会形成核心作者群[10]。然而当前核心作者发文量占比显著低于该阈值,表明算法公平研究领域尚未形成具有规模效应和显著主导作用的核心作者群。

**表1　核心作者测算**

| 序号 | 作者 | 发文量/篇 | 总被引次数 | $h$ 指数 | $h_m$指数 |
|---|---|---|---|---|---|
| 1 | 李乐民 | 9 | 2 708 | 23 | 23.008 |
| 2 | 王建新 | 7 | 3 787 | 28 | 28.007 |
| 3 | 洪佩琳 | 5 | 1 624 | 21 | 21.012 |
| 4 | 李津生 | 4 | 1 140 | 16 | 16.014 |
| 5 | 丁炜 | 3 | 974 | 14 | 14.014 |
| 6 | 赵春晖 | 3 | 5 311 | 34 | 34.006 |
| 7 | 窦文华 | 3 | 2 044 | 22 | 22.010 |
| 8 | 张德运 | 3 | 2 322 | 24 | 23.010 |
| 9 | 潘雪增 | 3 | 1 667 | 18 | 18.010 |
| 10 | 张治中 | 3 | 1 578 | 15 | 15.009 |
| 11 | 敖发良 | 4 | 588 | 13 | 13.022 |
| 12 | 程雪军 | 3 | 1 991 | 26 | 26.013 |
| 13 | 许可 | 3 | 5 203 | 34 | 34.006 |

## 2.2　算法公平研究的机构共现

一般而言,发文量较高且在相关学术界享有较高认可度的研究机构,其研究成果通常被认为更具权威性和专业性。因此,分析机构的发文情况既有助于厘清算法公平研究领域的主要参与力量机器合作网络,也能在一定程度上反映该领域成果的整体权威性与专业水平。将节点设置为“机构”、时间切片跨度设置为1年、比例参数 $k=25$,中国算法公平研究领域的主要机构分布如图3所示,各机构发文量见表2。

综合图3和表2分析可知,算法公平领域的高产机构以高校为主体,其中北京邮电大学、浙江大学、中国科学技术大学等机构尤为突出,极具代表性。具体而言,信息科技与电子通信领域的高等院校在算法公平的研究中充当了中坚力量,依托其雄厚的科研资源与平台优势,产出成果数量显著,为研究做出了重要贡献。相比之下,以国家数字程控交换工程研究中心为代表的科研院所虽在高产机构之列,但是数量较少,其发文量与影响力均显著低于高校机构。

**表2　算法公平研究领域的高产机构**

| 序号 | 机构 | 发文量/篇 | 序号 | 机构 | 发文量/篇 |
|---|---|---|---|---|---|
| 1 | 北京邮电大学 | 21 | 11 | 东南大学 | 9 |
| 2 | 浙江大学 | 20 | 12 | 武汉大学 | 9 |
| 3 | 中国科学技术大学 | 19 | 13 | 解放军理工大学 | 9 |
| 4 | 清华大学 | 19 | 14 | 国家数字程控交换工程研究中心 | 8 |
| 5 | 电子科技大学 | 17 | 15 | 西安电子科技大学 | 8 |
| 6 | 中南大学 | 14 | 16 | 外贸经济贸易大学 | 8 |
| 7 | 西南交通大学 | 10 | 17 | 重庆邮电大学 | 8 |
| 8 | 华中科技大学 | 10 | 18 | 云南大学 | 8 |
| 9 | 上海交通大学 | 9 | 19 | 解放军信息工程大学 | 7 |
| 10 | 南京邮电大学 | 9 | 20 | 南京航空航天大学 | 7 |

①普莱斯测定方法是指,若同一主题中半数论文为高生产力作者所撰,而这一作者的数量约等于作者总数的平方根,可以用公式表现为 $N\approx0.749\sqrt{N_{max}}$($N_{max}$为最高产作者的论文发表数量)。

②文献计量学中通常采用 $h$ 指数、$h_m$指数等方式对核心作者进行更准确的测定。根据赫希定义,一位学者的 $h$ 指数等于其发表了 $h$ 篇至少被引用 $h$ 次的论文数量,即最多有 $h$ 篇论文被引用了至少 $h$ 次。为了得到更准确的结果,引入了 $h_m$指数。具体测定公式为 $h_m=h+h/N_{c,tot}$,其中 $h_m$为对 $h$ 指数的准确修正,$N_{c,tot}$为学者总被引次数。由 $h_m$和 $h$ 的差值即可得知该学者在其研究领域中的影响程度,差值越小影响越大,反之差值越大影响越小。

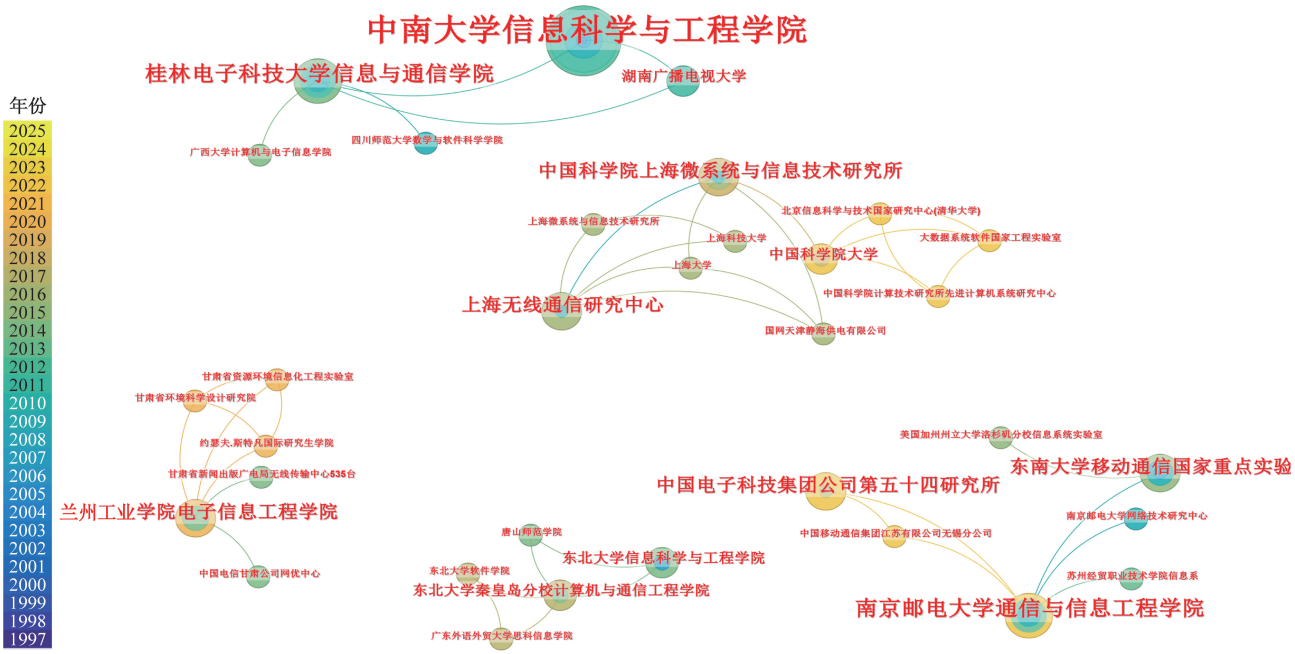



**图 3　算法公平研究领域主要机构共现**

此外,通过进一步观察机构合作图谱还可知,尽管研究机构已经形成了一定的社群聚类特征,但因整体规模有限,致使其影响范围相对受限,可能对算法公平研究的创新性发展形成一定制约。同时,各研究机构之间的连线稀疏;且多为简单的双边关系,表明各研究机构之间的合作网络密度较低,协同关系相对松散,尚未形成紧密的多元合作关系,一定程度上限制了协同效应的发挥。

## 3　算法公平研究的热点挖掘

前沿热点是指在特定时间内研究学者集中关注并密集研究资源的主题,集中体现了当时的研究兴趣点[11]。因此,挖掘前沿热点有助于系统把握研究领域现状,并对于预测其后续发展趋势至关重要。

### 3.1　研究热点:关键词共现分析与词频分析

关键词是对研究领域中主题、核心内容、研究前沿以及关注热点的高度凝练。分析其出现频率有助于有效识别目前该领域的研究兴趣点,进而为把握研究动态、推测未来趋势提供重要依据。将节点设置为"关键词"、时间切片跨度设置为 1 年,算法公平研究领域的关键词共现情况如图 4 所示。

由图 4 可知,算法公平研究领域的关键词共现网络具有较高可信度,其模块值 $Q$(0.766 7)和轮廓值 $S$(0.95)均远超可信阈值($Q>0.3$,图谱结构显著;$S>0.5$,聚类结果合理[12])。聚焦关键词的出现频次,"公平性"以 70 的频次位居首位,凸显了其核心关注度,而如"比例公平""服务质量""算法公平""资源分配""公平算法""调度算法""算法歧视""公平调度""公平"等其他高频词也集中反映了当前研究的焦点。同时,根据中介中心度识别网络中的关键节点可知(中介中心度>0.1,则表明该关键词处于重要位置、扮演重要角色[13]),"公平性""算法""公平调度""人工智能""算法公平""公平""比例公平""服务质量""调度算法"等关键词中心度较高,表明它们在贡献网络中起到连接不同研究主题的作用,是重要的核心节点;此外值得注意的是,大多高频关键词也具有较高中心度,说明算法公平领域的研究热点同时也构成了知识网络的核心结构。综合关于高频词和高中心度关键词的分析,中国算法公平研究的热点主要集中在算法公平性理论、公平算法设计与优化、服务质量保障及实际应用中的算法公平性问题等方面。

除此之外,关键词词频分析还强调节点集合大小与网络密度。根据图 4,中国算法公平研究呈现出以下结构特征:网络规模较大(节点数为 451),反映出研究主题的广泛性与多样性;网络整体连接相对稀疏(连线数为 737,网络密度为 0.007 3),表明不同主题间的关联尚不紧密。然而,特定主题内的关键词联系相对紧密,表明网络中依然存在一定的密集性,对于理解算法公平的热点主题具有关键意义。

### 3.2　热点主题:关键词聚类分析

关键词聚类旨在将内容相似且关联紧密的关

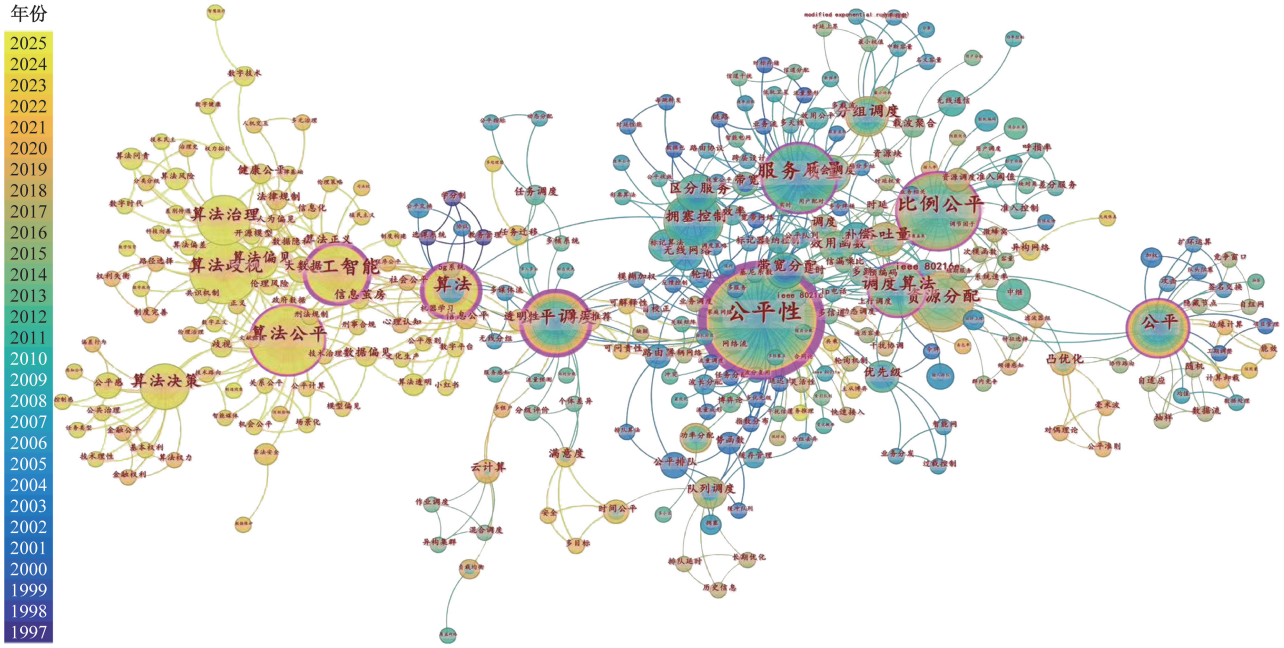



**图 4　关键词共现情况**

键词进行聚集归类，从而识别出不同主题的类别。通过聚类结果，不仅可以观察某一节点在样本文献中出现的频次及其连接关系[13]，更重要的是能够有效识别出算法公平领域的核心研究主题[14]。具体如图 5 所示。

图 5 呈现了算法公平领域的关键词聚类效果。该网络显示出较广的主题覆盖度和较强的主题间关联性。根据分析，共出现了 12 个主要聚类：#0 公平性、#1 算法歧视、#2 服务质量、#3 比例公平、#4 公平调度、#5 公平、#6 算法、#7 公平算法、#8 算法决策、#9 轮询、#10 流量监测、#11 公平排队。各聚类轮廓值均大于 0.7，显示结果可靠。因此基于对上述聚类所代表的内容进行深入解读和综合归纳，可将中国算法公平的热点研究主题划分为“算法公平的内涵与成因”“算法公平的构成主体及风险防范”和“算法公平的运用”3 个方面。

（1）算法公平的内涵和成因。作为国内学者进行算法公平研究的基础，厘清其内涵与成因是构成理论探索的首要环节。该主题主要涉及聚类#0 公平性、#6 算法、#5 公平、#1 算法歧视和#8 算法决策，是相关理论的重要组成。党的十八届三中全会强调“紧紧围绕更好保障和改善民生、促进社会公平正义深化社会体制改革[15]。”这一指示在数字时代延伸为对算法公平的保障要求，而明确其内涵则是前提。其中，“算法歧视”作为数字时代公平缺失的典型现象，在研究中占比显著并可作为关键切入，其本质是传统歧视随技术演化的新形态[16]，由不合理数据分析引发对特定群体的不公正待遇[17]。据此，算法公平可定义为“不受算法演算或数据偏见影响，对个体/群体给予同等对待”，即“相似者得相似待遇”[18]。然而在现实中，算法公平这一目标却面临着根本矛盾：算法模型训练之初因依赖带有分类决策与群体标签等固有缺陷的数据集，使得技术设计难以规避历史偏见的再生产，同时也形成了“理论理想-实践受限”的认知框架。因此，算法公平的内涵需同时涵盖政策要求、多维公平界定、技术载体约束以及现实困境。

而要突破这一框架，则需要厘清算法公平的成因。算法公平的构建需回应当前中国特色社会主义初级阶段“以起点公平、机会公平为主，兼顾过程公平和结果公平，逐步向结果公平迈进[19]”的要求，但是在技术层面面临严峻挑战：算法模型在训练之初其数据集合就因分类决策、群体标签等不可避免地存在歧视性因素[20]。更关键的是，算法固有的技术局限（如算法黑箱、迭代偏差等）会进一步放大训练数据中的隐性历史偏见。这种数据初始污染与技术放大的双重困境，从根本上制约了算法公平的技术实现路径。

（2）算法公平的主体构成及风险防范。主体构成和风险防范是算法公平相关研究的重点，涉及#3 比例公平、#7 公平算法、#2 服务质量和#10 流量监测等聚类，共同构成了“核心要素支撑构成＋

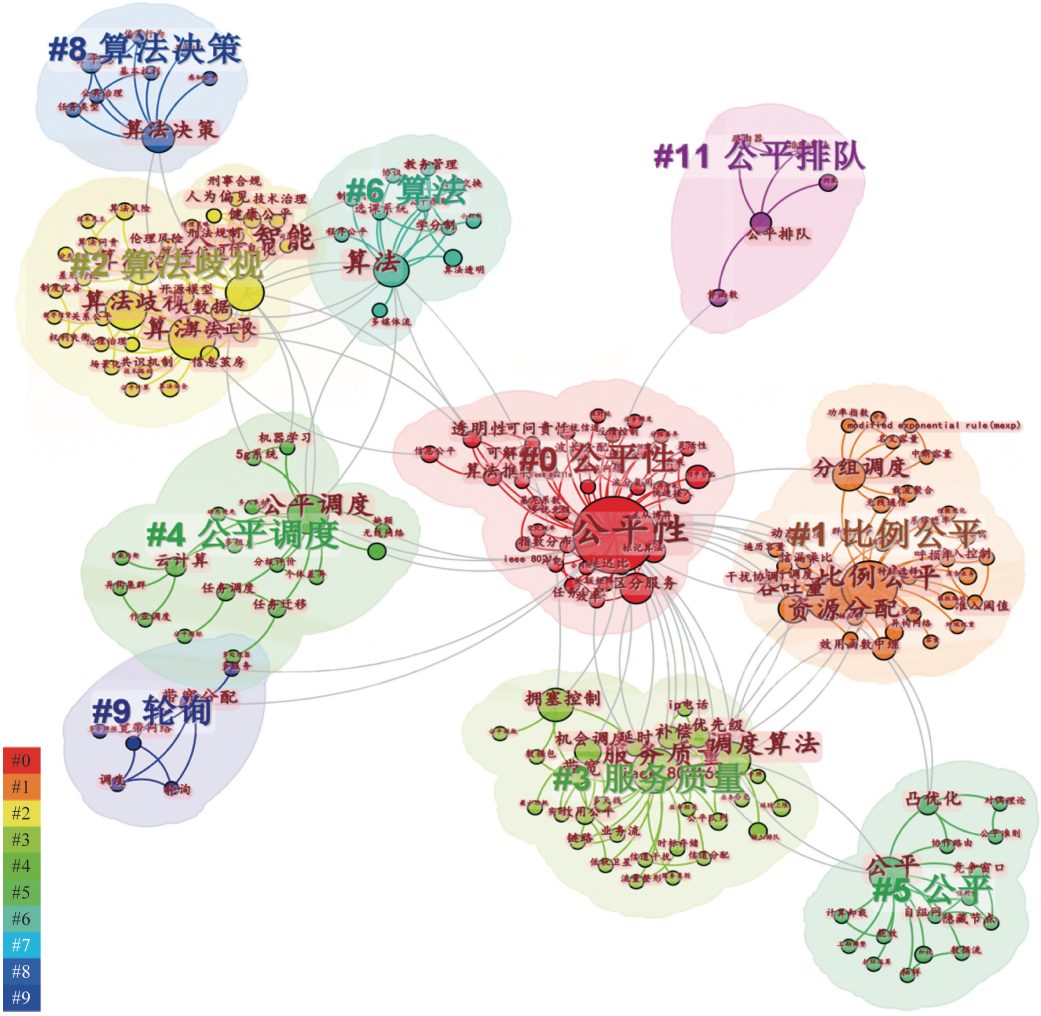



**图 5　关键词聚类分析**

技术手段保障防范”的协同关系。算法公平问题的主体包括数据提供者、服务提供者、监管者及用户[21],其交互流程共同构成了算法公平的社会基础:数据提供者收集、提供数据;服务提供者以此训练算法并提供服务;但由于模型偏差、主观偏见嵌入等因素[22],需要政府、法院等充当监管者,确定公平标准并实施约束。以上交互流程揭示了算法不公的深层根源,主要包含偏见根源及算法根源[19]。为防范此类问题,需构建从原则定义到落地执行的技术框架。首先,作为可量化的公平准则,“比例公平”是算法公平构成的核心维度,为公平算法设计提供标尺;而“公平算法”作为执行载体,通过编码比例公平等准则将抽象目标转化为可运行逻辑,二者共同形成算法公平的“核心执行体系”。其次需要强调的是,算法公平的实现需要兼顾效率与约束。因此“服务质量”作为算法性能边界,需要公平目标协同优化,形成“效率-公平平衡机制”。最后,通过运用风险防范手段“流量监测”以实时追踪资源分配与任务调度数据、识别因偏见根源所导致的数据偏差或因算法根源所导致的逻辑漏洞引发的执行不公平,为动态调整提供依据。在此基础上,针对识别出的偏见根源与算法根源,前者需要采用侵入性监管措施获取所需信息并追究主体的法律责任[23],而后者则需要运用算法设计优化,最终形成“识别-监视-修正”的闭环。

(3)算法公平的运用。该主题主要包括聚类#4 公平调度、#11 公平排队和#9 轮询,是“算法公平”研究在具体实践场景中的 3 种核心路径。具体而言,“公平调度”是算法公平在任务或资源调度场景中的核心应用形态,其本质是通过设计调度算法,确保不同主体随着时间的推移都能得到相等的资源[24],避免因算法偏向导致部分主体被系统性忽视,直接体现了算法公平中“无偏倚分配”的核心诉求。“公平排队”则聚焦于队列管理场景,是算法公

平在请求排队环节的细化运用。本质上是将算法公平的原则嵌入排队逻辑，保障队列中各主体的“机会公平”。“轮询”作为一种基础且直观的调度策略，通过轮流为每个主体分配资源或处理时间，避免单个主体独占资源，是算法公平中“平等对待”原则的极简实现。虽简单但能有效防止极端不公平，属于算法公平运用的基础手段。总的来说，以上 3 种路径构成了算法公平落地的技术谱系。“轮询”奠定了平等性基础，是算法公平在调度中最基础的运用形式；“公平排队”实现了场景化机会公平，是算法公平在队列场景的专项运用；而“公平调度”则是算法公平在更广泛调度领域的综合运用，实现了系统分配正义。这三者均通过具体算法将“公平”从抽象原则转化为可操作的规则，使算法公平从理论诉求转化为技术实体。

## 4　中国算法公平研究的演进趋势

### 4.1　研究脉络：Timeline 分析

CiteSpace 中的 Timeline 分析是一种聚焦时间维度的分析方式，其核心作用在于将静态的文献数据转化为动态的知识演进图谱，帮助研究者系统性地把握研究领域的历史脉络、当前热点与未来趋势。算法公平研究领域的 Timeline 分析结果如图 6 所示。

根据图 6 能够直观了解每年的研究主题，根据主题与时间可以归纳为以下 3 个显著阶段。

第 1 阶段（1997—2009 年）为算法公平研究的初步探索阶段。这一阶段恰逢中国互联网建设的初始时期，研究大多围绕网络与系统中的公平性问题展开。其核心议题主要聚焦于资源受限场景下公平性与效率、服务质量的平衡，并由此衍生出多样化的实践算法，如涂晓东和李乐民[25]关于公平排队算法的研究、刘刚等[26]关于公平分配算法的研究、江勇和吴建平[27]关于公平调度算法的研究等。这些早期研究成果共同奠定了算法公平研究的理论框架，不仅为具体算法设计提供了逻辑依据，也为不同场景下公平性问题的解决提供了初步的分析视角。总的来看，这一阶段主要聚焦算法层面的技术实现，在具体研究领域上存在一定局限。虽然研究对“算法公平”理论进行了初步探索并提出了诸多实践算法，但尚未与社会学、管理学等理论相结合，同时也未与“算法歧视”“算法偏见”等相近概念进行明确辨析。

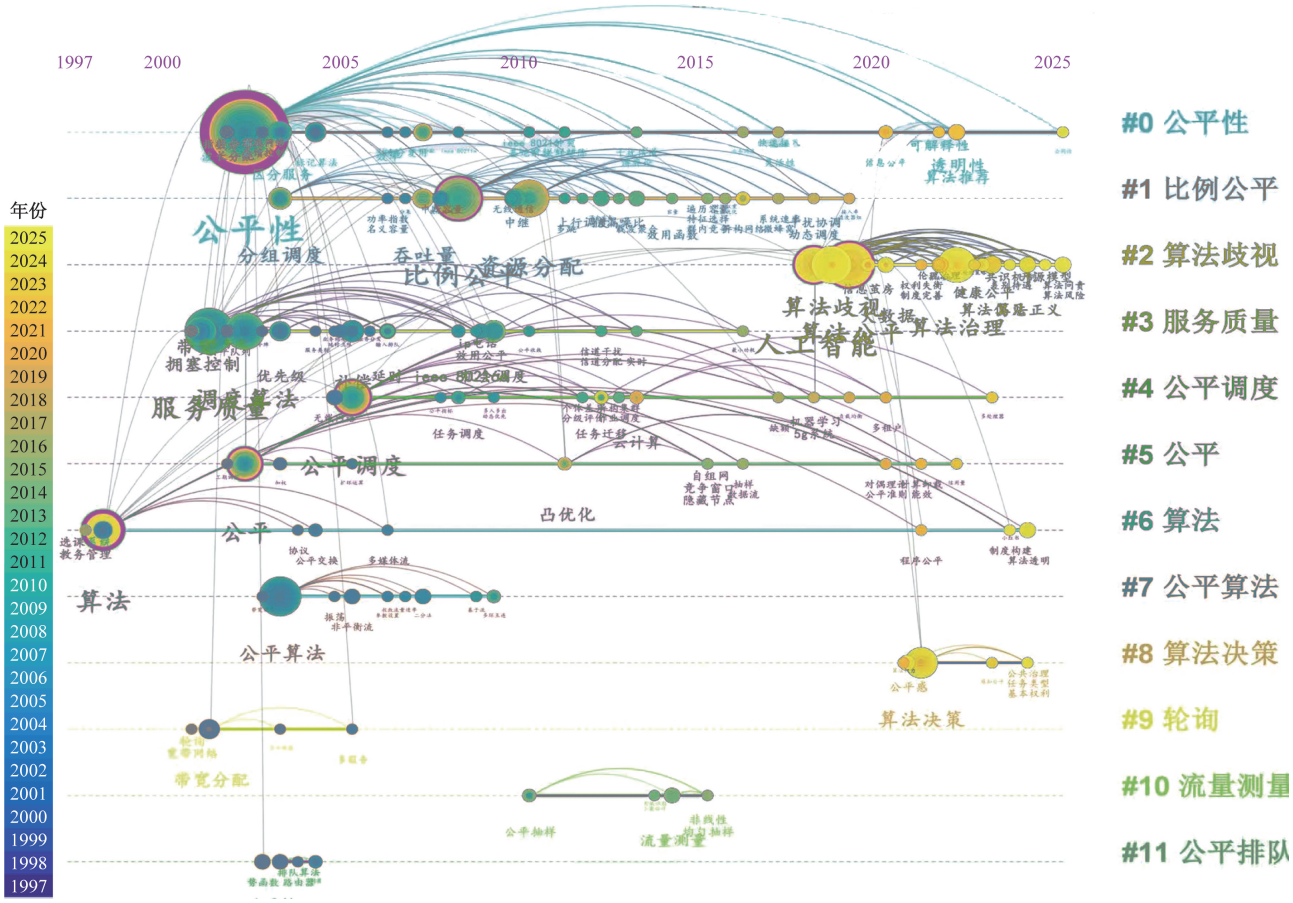



图 6　算法公平领域的 Timeline 分析结果

第 2 阶段(2010—2021 年)为算法公平研究的发展调整阶段。尽管这一阶段中关于实践算法的研究依然占据多数,但随着移动互联网的兴起和发展,相关研究的侧重逐渐倒向“互联网平台及其催生的新型社会生态”[28],并开始与“算法偏见”“算法歧视”“法律规制”“伦理反思”等主题融合,跨领域理论体系初见雏形。代表性研究包括张莉莉和朱子升[29]关于运用法律规制算法歧视时路径与制度的完善研究、汪建[30]对于人工智能算法歧视问题的评述研究、刘培和池忠军[31]对于算法歧视的伦理反思研究等。这些跨学科研究推动算法公平从技术问题上升为社会治理议题,拓展了其在法律、伦理、金融、社会管理等多元场景中的应用适配,共同为构建从技术算法到社会公平的全链条保障体系提供了支撑。

第 3 阶段(2022 年至今)是算法公平研究的转型发展阶段。随着经济社会数字化的转型加速,互联网发展进入了智能物联时代,数字算法成为社会关键基础设施[28],在网约车、外卖平台、网络购物等多个民生领域发挥着核心作用。但是算法的广泛应用也加剧了不同个体间分享信息的差异化程度,扩大了公众的认知鸿沟,并可能消解社会共识的信息基础。这一过程最终导致公众信息公平权被侵蚀,信息资源为互联网寡头所垄断,加剧了严重的算法不公问题[32]。在此背景下,如行政学、伦理学、法学等在内的社会科学领域开始关注算法公平的相关研究,推动该领域从技术层面向社会治理层面的转型。在这一阶段中,相关成果数量逐年攀升,出现了如魏远山[33]关于算法歧视的政府治理研究、陈雄燊[34]关于算法伦理维度的研究、任颖[35]关于算法规制的立法研究等成果。此类研究凸显了跨学科合作的重要性,并使其逐渐成为该领域的主导研究新范式。

### 4.2 研究前沿:关键词突现分析

突现分析能有效检测特定时间内的关键词突变,有助于发现新型研究主题和提高学者应对科技发展的动态能力[36]。运用关键词突现分析,可以揭示算法公平领域的研究热点演变态势,为准确研判其未来发展趋势提供重要依据。分析结果如图 7 所示。

图 7 显示,自 1997 年起的高突现关键词主要包括“算法公平”“比例公平”“算法歧视”“资源分配”“人工智能”“算法决策”“算法治理”等,反映出相关研究的前沿议题。从突现强度与研究时间来看,

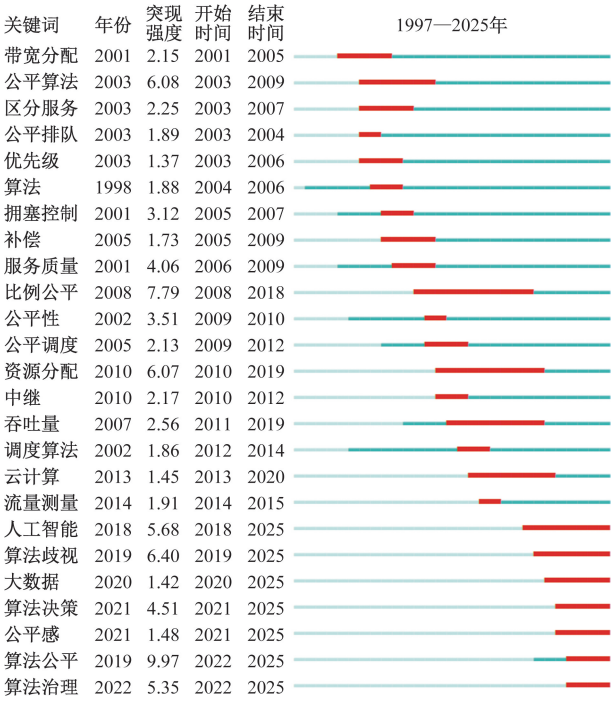



**图 7 关键词突现图谱**

“比例公平”拥有较高突现率(7.79)与最长的研究时间(2008—2018 年),是“算法公平”演进过程中的重点关注对象与核心焦点,在早期发展中发挥了重要的支撑与推动作用。

而近 5 年算法公平领域的主要突现词则转向“大数据”“算法决策”“公平感”“算法公平”和“算法治理”,其中突现率最高的是“算法公平”,达到了 9.97。作为一个综合性概念,“算法公平”与研究前期出现的“比例公平”“资源分配”“算法歧视”等关键词联系密切,能够有效整合分散的研究点,使得学者能够从更宏观、系统的视角探索算法公平问题,推动了该领域的整合与深化。“算法公平”的高突现强度不仅体现了其作为核心议题的重要性,也凸显了其在理论研究和实践应用中的双重影响力。在理论上,它促使学者们持续完善算法公平的基础理论体系,包括定义界定、标准度量、实现机制等;在实践中,它显著提升了社会学、伦理学、法学等多学科对算法公平的关注,促使社会治理、伦理反思、法律规制等应用场景更加注重算法的公平性设计以避免算法歧视等不公平现象。

## 5 结论与展望

### 5.1 研究结论

基于 CiteSpace 分析,系统梳理了中国算法公平领域的研究演进,深入挖掘了当前热点主题和研究前沿并揭示了其中的互动关系、影响及价值,得

出以下结论:算法公平领域的研究产出在经历了初期的快速增长、中期的显著波动后,正处于复苏阶段,近5年的发文数量正呈现出先快速增长后逐步回落的变化趋势。同时,相关研究也经历了从初步探寻到深层挖掘、从单一视角到跨领域融合的演进过程。尽管算法公平研究领域目前已经取得了丰硕的成果,但依旧存在一些薄弱环节、面临着诸多挑战。

(1)从主要作者与研究机构角度来看,近30年来,经过学者对算法公平的深入研究与剖析,相关领域积累了一批高质量学术成果,同时跨领域探索也初见端倪。作为中国算法公平研究领域的核心作者,李乐民、王建新、洪佩琳、李津生、丁炜等学者的成果在演进过程中发挥了关键作用。其中,许可的《算法规制体系的中国构建与理论反思》为完善算法相关法律、调和算法向善做出了重要贡献,是算法公平本土化维护方面的里程碑。此外,新一代学者的不断涌现也为相关研究注入活力。例如,近5年的研究成果中,任颖的《算法规制的立法论研究》、梅傲的《积极伦理观下算法歧视治理模式的革新》、陈昌凤的《算法伦理研究:视角、框架和原则》等不仅拥有较高被引量,也有效促进了跨领域研究发展。然而,现有研究成果表明,中国的算法公平研究尚未形成稳定的核心作者群或紧密的科学共同体,研究机构间也缺乏多元合作,这对协同研究产生了一定制约,跨领域知识融合与交流合作等仍有待强化。

(2)从研究主题层面来看,相关研究主要聚焦公平调度、比例公平、公平算法、分配算法、算法研究、算法歧视等,构成了“算法公平”领域的主要研究主题。通过多角度、多层次的探究,相关研究逐渐揭示了中国算法公平领域的运行机制与实践范式。同时,研究主题还呈现出与政策等场景扩展密切相关的特征,其中“公平调度”“公平算法”“算法歧视”等尤为突出。这些研究主题与目前数字中国政策方针高度契合,不仅反映出相关研究与政策设计的高度一致,也彰显了其时代价值和实践意义。

(3)从研究脉络方面看,算法公平研究的演进脉络呈现出显著的波动特征,历经了由浅至深、由单一学科到跨领域的过程。中国算法公平研究的演进可以划分为3个阶段:算法公平研究的初步探索阶段(1997—2009年)、算法公平研究的发展调整阶段(2010—2021年)和算法公平研究的转型发展阶段(2022年至今)。在各阶段中,政策方针、学术研究与实践探索之间的联系愈发紧密,初步形成了“政策引领研究方向-学术研究提供理论与技术支撑-实践探索检验并反馈”的闭环,为推动算法公平落地提供了技术方案和制度规制,反映了相关研究与时代同步的特性。但需要指出的是,当前研究主要聚焦在社会科学、经济学、信息技术科学等方面,在跨学科广度上仍存在拓展空间。因此,未来研究应进一步深化跨领域协作、拓宽研究视野以推动相关研究向更加全面、均衡的方向发展。

### 5.2 前沿热点预测

目前,尽管“调度算法”“公平调度”等传统主题仍保持一定研究热度,但随着“算法歧视”“算法偏见”等新兴议题的涌现,研究重心已显著转向后者。这一转变直接反映了算法公平领域的研究趋势已从理论框架构建转向实践方案开发。由此可预测的是,未来算法公平将聚焦“算法歧视”问题,以“去歧视化”为前沿热点,致力于通过内部算法优化及外部机制适配构建可信的算法生态。

(1)从内部去歧视化。从算法内部实现去歧视化是涉及技术实现和伦理哲学等多方面的复杂过程,涵盖数据处理、算法设计、伦理界定诸多环节。从技术方面去歧视化需要对算法进行再设计,是一种直接优化路径[34]:这就要求结合算法决策具体目的,并依据治理主体的类型及其作用,建立包括分类标准、性能标准和设计标准的技术框架;而在伦理哲学方面,则需要重构价值基础:关键在于将透明原则和程序公平融入算法设计与应用过程[37]。一方面,需公开算法设计意图和运行机制以接受监督,践行透明原则;另一方面,需构建涵盖先验公平、准确性公平与差异公平等多维度的公平性标准体系[31],以避免采用单一的、绝对的评判准则。

(2)从外部去歧视化。数字时代下,算法深度融入社会治理的全流程,依托大数据等手段深刻改变了传统行政的运行规律和基本范式。尽管带来了效率提升等诸多便利,但算法决策对模型的严重依赖意味着其固有缺陷将持续反映在最终结果中。因此,需要引入法律等外部规制手段,与内部技术优化形成互补。作为强制规范手段,法律主要作用于算法的应用过程,不仅为算法运行提供了合法性基础并赋予其强制力,更关键的是能够提高设定明确规范来引导、约束算法的道德价值[38]。当前,围绕“算法公平”的法律研究正日益与社会学、管理学等领域交叉融合,在跨学科探索聚焦于界定算法的使用底线、规制算法设计者的责任,旨在减少歧视

性后果并力求实现个体公平和群组公平、过程公平和结果公平的有机统一[39]。

尽管上述内部技术优化与外部法律规制共同构成了“去歧视化”的双重路径,一方面,内部的技术优化为外部的社会治理提供了可操作工具、量化依据与落地载体,另一方面,外部的社会治理为内部的技术优化设定边界、明确方向、提供纠偏机制,但实现算法公平的深层关键在于这两者能否形成有效协同与良性互动。研究指出,传统算法设计主要关注诸如模型速度、鲁棒性等技术目标,缺乏对社会目标的系统性响应[40]。这种脱节使得技术优化难以匹配治理需求;同时,当前算法审计主要依赖静态文档审查,监管机构缺乏验证因果推断模型的技术能力[41]。由此可见,目前“去歧视化”的关键挑战在于如何有效衔接内部技术实现与外部社会治理机制,唯有通过技术与治理的深度耦合,才能逐步消解公众信任赤字,推动算法公平从理论走向实践。以招聘场景为例,在技术优化环节中,可以运用对抗训练技术生成“女性”“大龄求职者”等样本以削弱算法模型对性别、年龄等敏感数据的歧视。其实践效果则可以通过对样本覆盖率、特征权重调整等因素的算法审查进行判定。若依旧存在歧视类问题,可以依据“算法责任追溯机制”,要求开发团队重新训练模型并提交整改报告,对算法进行进一步调整迭代,从而在具体场景中形成“技术优化-治理验证-反馈迭代”的双向协同体系。

### 5.3 拓展空间展望

早在2018年,皮尤研究中心调查已揭示公众对算法公平的深度忧虑。58%美国受访者认为计算机程序始终会反映人为偏见[42],这种信任缺失严重阻碍了先进算法的应用推广。在此背景下,算法公平领域的研究主要聚焦于“去歧视化”及其多维实践。然而,去歧视化依旧面临着数据偏见固化、算法修正与功能效率冲突等挑战,亟须学者突破单一研究视角,整合跨学科知识开辟新路径。由此,“学科融合”就成为了破解公平瓶颈的关键路径。通过整合计算机、社会学等学科,既为技术优化提供创新方法论,也通过多视角保障全链条公平。基于此,相关研究领域需要在推动跨学科融合、构建科学共同体、更新治理体系等维度上进行深层研究。

(1)深化跨学科合作机制与研究范式。“算法公平”目前已经逐步从信息工程学领域逐渐转向社会科学领域,研究内容上从计算机的客观公平排序算法向管理学领域的算法歧视、算法黑箱等概念拓展。由此可见,关于算法公平的研究已不再是技术领域的单一问题,而是全社会需要共同面临的重要目标。因此,深化跨学科合作机制与研究范式至关重要。其一,通过构建“跨学科数据共享与标准对接”机制,统一相关研究中如案例数据库、公平性评估指标库等基础数据或标准,打通不同学科的数据壁垒,避免各学科因数据口径不一导致的研究断层。例如,将计算机学科的“统计公平性指标”与法学的“差别影响标准”对接以形成可跨学科解读的评估体系。其二,以具体的算法公平问题为核心,融合不同学科的研究方法。例如,将计算机科学的“算法公平性量化模型”与教育学的“教育机会均等理论”相结合,构建适合教育场景的公平性评估框架等。

(2)构建聚焦算法公平问题的科学共同体。“算法公平”研究的跨学科拓展为科学共同体的实现提供了理论依据,同时也为应对数字时代的复杂社会问题提供了组织基础。作为由具有共同科学信仰、范式和目标的科学家群体所构成的组合[43],科学共同体本质上是基于共享学术范式与伦理目标的研究者协作网络。随着学科发展越来越成为一个系统工程,这就需要不同领域的科学家进行合作,从而形成跨学科的科学共同体[44]。科学共同体不仅覆盖技术环节,更贯穿“社会需求分析-技术开发-政策适配”的全链条,有助于解决当前日益复杂的科学及社会问题。作为数字时代亟须突破的现实问题,“算法公平”只有通过多学科的协同努力才能成为现实,这无疑强调了科学共同体在资源整合以及知识共享方面的重要价值。其一,通过建立跨学科数据平台(如脱敏医疗数据与司法记录的联合调用),破解算法训练中的数据孤岛问题;其二,组织跨领域学者明确核心概念、伦理底线,在规范敏感数据的使用边界的同时为分散研究提供统一对话框架;其三,通过推动高校与企业共建联合实验室,由企业提供真实数据与工程支持、高校提供理论与技术方案,形成“问题-成果-实践”闭环。

(3)推动适应技术发展的社会治理体系更新。算法公平领域的跨学科融合推进了社会治理从“工具理性”转向“价值理性”,其核心在于通过跨学科协作与知识整合优化政策设计、执行与评估全流程,确保算法公平嵌入社会治理体系。其一,建立场景化规制与动态复核机制,实现精准治理。针对不同的应用场景,需制定差异化规制标准。例如,金融领域侧重于运用信贷以评估算法的群体公平

性，就业领域则强调运用性别等要素来检验算法是否公平。其二，推动跨部门协作生态，凝聚治理合力。例如，打破网信、司法、行业主管部门、科研机构间的壁垒，建立“算法公平治理联席会议”制度，明确各主体权责。网信部门统筹算法备案与风险预警、司法部门完善算法歧视纠纷的裁判标准、行业主管部门制定领域内技术规范、科研机构根据规范进行算法设计和迭代等。其三，加强复合型人才培养与伦理审查机制，夯实理论基础。例如，在高校开设“算法公平与社会治理”跨学科专业，融合计算机科学、法学、伦理学等知识，通过“学术＋行业”的双指导模式培养既懂技术原理又熟悉法律伦理的人才。同时将伦理审查嵌入算法全生命周期，通过设立伦理审查委员会，对设计目标、训练数据、输出结果等进行全流程评估。

以信用评分中算法系统性低估农村群体的情况为例。首先需深化信息工程学、社会学、管理学等的跨学科合作机制，打破学科壁垒。从信息工程学角度看，信用评分算法往往基于城市群体的金融行为数据，其变量选取与农村实际情况存在偏差，能通过模型无关的局部可解释性（local interpretable model-agnostic explanations，LIME）算法等定位出其中的关键技术点；社会学则能揭示农村群体在金融参与中的结构性障碍，比如农村地区金融服务覆盖率低，很多人缺乏正规信贷记录，大量非正规借贷行为未被纳入采集范围，导致算法无据可依而给出低分；管理学视角则能分析金融机构在算法设计中对风险控制的过度侧重如何忽视了农村群体的特殊性，使得算法在目标设定上存在偏差。跨学科合作能够整合不同层面的分析，避免单一视角局限，从而全面把握问题成因。其次，应构建聚焦算法公平问题的科学共同体。科学共同体需涵盖技术研发者、社会研究者、金融机构代表、政府监管部门、农村群体代表等多方主体。技术研发者负责根据跨学科成果优化算法，如开发能识别农村特色资产等的针对性算法；社会研究者和农村群体代表则持续反馈农村群体的实际需求和算法应用中的问题，为迭代优化提供方向；金融机构作为算法使用者，需转变观念，将算法公平纳入自身发展目标；政府监管部门则起到引导和规范作用，确保各方协同合作。科学共同体通过常设沟通平台与联合研究项目整合分散资源，形成技术与需求的协同合力。最后需推动适应技术发展的社会治理体系更新。当前针对信用评分算法的治理体系还存在滞后性，难以应对算法快速发展带来的新问题，因此立法层面需明确算法在信用评分中不得存在针对农村群体的歧视性设定，同时要求其设计与应用过程中的可解释性和透明度；政策层面应加强农村金融数据的基础设施建设，推动相关数据的规范采集共享，为算法提供全面、准确的数据；监督层面要建立包括政府、行业、社会等多方参与的监督机制，定期评估相关算法并及时发现和纠正其中的公平性问题。

总的来看，以上“深化跨学科合作-构建科学共同体-推动治理更新”的解决方案能为算法公平问题的解决提供良好的内外部环境，确保技术优化和多方协作真正转化为正向社会效应，逐步实现算法公平。

## 参考文献


[1] 杨又，陈小念. 算法资本主义的全球殖民机制及其批判[J]. 深圳社会科学，2024，7(5)：148-160.

[2] 张力，郑丽云. 算法推荐的歧视与偏见[J]. 中国报业，2020(13)：48-49.

[3] 肖田. 人工智能时代算法偏见浅析[J]. 科技传播，2020，12(7)：122-123.

[4] 国家互联网信息办公室，中华人民共和国工业和信息化部，中华人民共和国公安部，国家市场监督管理总局. 互联网信息服务算法推荐管理规定[EB/OL]. (2022-01-04)[2025-05-11]. https://www.cac.gov.cn/2022-01/04/c_1642894606364259.htm.

[5] 侯剑华，胡志刚. CiteSpace软件应用研究的回顾与展望[J]. 现代情报，2013，33(4)：99-103.

[6] 赵建保. CiteSpace可视化流程与分析范式研究[J]. 知识经济，2014(16)：105-107.

[7] 邱均平，杨思洛，刘敏. 改革开放30年来我国情报学研究的回顾与展望(一)：情报学研究论文的年度分布与期刊分布分析[J]. 图书情报研究，2009，2(2)：1-7.

[8] 钟文娟. 基于普赖斯定律与综合指数法的核心作者测评：以《图书馆建设》为例[J]. 科技管理研究，2012，32(2)：57-60.

[9] 刘永胜，朱亚丽，梁小枫，等. 数字图书馆研究文献的增长态势及其核心著者[J]. 晋图学刊，2010(3)：6-11，39.

[10] 王泓凯，张烈. 国际视阈下的博物馆交互设计研究现状、热点与趋势：基于CiteSpace可视化计量分析[J]. 南京艺术学院学报(美术与设计)，2024(4)：90-96.

[11] 范蕴玉. 国外社区韧性研究的知识图谱与热点前沿：基于Web of Science(2001—2020年)的可视化分析[J]. 温州大学学报(社会科学版)，2022，35(1)：81-95.

[12] 陈悦，陈超美，刘则渊，等. CiteSpace知识图谱的方法论功能[J]. 科学学研究，2015，33(2)：242-253.

[13] 章志鹏，曾志勇，李忠. 知识图谱的应用研究现状与发展趋势：基于CiteSpace的可视化分析[J]. 信息与电脑

(理论版),2020,32(11):118-120.
[14] 李占乐,张燕. 当前国内数字政府研究进展与趋势:基于 CiteSpace 软件的文献计量分析[J]. 长沙大学学报,2021,35(5):65-71.
[15] 中国共产党第十八届中央委员会第三次全体会议公报[EB/OL].(2013-11-12)[2025-05-20]. https://www.xuexi.cn/8678b2383e6993870a760de7eb5c0d12/e43e220633a65f9b6d8b53712cba9caa.html.
[16] 石颖. 算法歧视的发生逻辑与法律规制[J]. 理论探索,2022(3):122-128.
[17] 谢永江,杨永兴. 人工智能时代下的算法歧视及其治理研究[J]. 北京邮电大学学报(社会科学版),2022,24(5):18-25.
[18] 张欣. 算法公平的类型构建与制度实现[J]. 中外法学,2024,36(4):866-883.
[19] 张跃进. 公平:政府行政的目标和手段:关于公平与效率关系的思考[J]. 沈阳师范学院学报(社会科学版),2001(1):11-12.
[20] 赵精武,陆睿. 从公示到透明:算法公平的风险挑战与治理路径[J]. 暨南学报(哲学社会科学版),2023,45(9):69-82.
[21] 范卓娅,孟小峰. 算法公平与公平计算[J]. 计算机研究与发展,2023,60(9):2048-2066.
[22] 刘朝. 算法歧视的表现、成因与治理策略[J]. 人民论坛,2022(2):64-68.
[23] 张凌寒. 平台"穿透式监管"的理据及限度[J]. 社会科学文摘,2022(2):75-77.
[24] 李千目,张晟骁,陆路,等. 一种 Hadoop 平台下的调度算法及混合调度策略[J]. 计算机研究与发展,2013,50(S1):361-368.
[25] 涂晓东,李乐民. 分组公平排队算法的仿真研究[J]. 电子科技大学学报,2000(4):440-444.
[26] 刘刚,张德运,刘晶. 基于用户公平的随机早期检测算法改进[J]. 微电子学与计算机,2004(6):119-122.
[27] 江勇,吴建平. 基于多服务类比例公平的分组调度算法[J]. 中国科学 E 辑:技术科学,2003(10):865-875.
[28] 谭天. 数字文明的社会化指向:中国互联网 30 年回顾与展望[J]. 海河传媒,2024(3):1-6.
[29] 张莉莉,朱子升. 算法歧视的法律规制:动因、路径和制度完善[J]. 科技与法律(中英文),2021(2):15-21.
[30] 汪建. 智能社会公平为先:试评人工智能算法歧视问题[J]. 中国电信业,2020(1):10-11.
[31] 刘培,池忠军. 算法歧视的伦理反思[J]. 自然辩证法通讯,2019,41(10):16-23.
[32] 申楠. 算法时代的信息茧房与信息公平[J]. 西安交通大学学报(社会科学版),2020,40(2):139-144.
[33] 魏远山. 智能时代政府治理算法化的优势、困境与应对策略[J]. 公共管理与政策评论,2024,13(4):24-38.
[34] 陈雄桑. 人工智能伦理风险及其治理:基于算法审计制度的路径[J]. 自然辩证法研究,2023,39(10):138-141.
[35] 任颖. 算法规制的立法论研究[J]. 政治与法律,2022(9):98-111.
[36] 彭国超,孔泳欣,王玉文. 多维指标融合的主题突变检测研究[J]. 情报学报,2022,41(6):584-593.
[37] 韩勇. 算法歧视的成因及治理路径[J]. 青年记者,2021(8):94-95.
[38] 崔靖梓. 算法歧视挑战下平等权保护的危机与应对[J]. 法律科学(西北政法大学学报),2019,37(3):29-42.
[39] 许可. 算法规制体系的中国建构与理论反思[J]. 法律科学(西北政法大学报),2022,40(1):124-132.
[40] 张吉豫. 构建多元共治的算法治理体系[J]. 法律科学,2022(1):115.
[41] 张欣. 算法影响评估制度的构建机理与中国方案[J]. 法商研究,2021(2):102.
[42] 汪怀君,汝绪华. 人工智能算法歧视及其治理[J]. 科学技术哲学研究,2020,37(2):101-106.
[43] 达亚·瑞迪. 科学共同体使命与价值共识[J]. 科技导报,2021,39(2):18-19.
[44] 吕洁,于建星. 科学共同体的形成及机制优化[J]. 科学管理研究,1999(1):42-44.

# Research Hot Spots and Expansion Space of Algorithmic Equity in China: A CiteSpace-based Bibliometric Analysis

ZHOU Chenyang[1], CAO Hanyang[1,2]

(1. School of Marxism, Yangzhou University, Yangzhou 225009, China;
2. Research Center for Government Governance and Public Policy, Yangzhou University, Yangzhou 225127, China)

**Abstract**: Algorithmic fairness plays a crucial role in promoting social development. By taking the literature from 1997 to 2025 in CNKI(China National Knowledge Infrastructure) as the data source and using the visualization tool CiteSpace, the research found that the research on algorithm fairness in China has shown significant fluctuations since 1997, and subsequent research will continue to grow. The research hotspots cover the "de-discrimination" dimensions such as algorithmic bias and algorithmic discrimination. From the perspective of the research context, the related research has gone through the exploration stage, the adjustment stage and the recovery stage. The frontier hotspots and research expansion have distinct problem-oriented characteristics, reflecting the development trend of diversity and focus. Looking ahead, it is necessary to establish an interdisciplinary cooperation mechanism, build a scientific community and update the social governance system.

**Keywords**: algorithmic fairness; knowledge graph; research hotspots; cutting-edge trends