算法偏见对绩效考核公平性的影响路径研究

□张旭军[1]，宋永[1]，吴学秉[1]

（1.汕尾职业技术学院高质量发展研究中心，广东汕尾 516600）

[摘要]人工智能、大数据技术汹涌而来，它们催生出的算法在绩效考核领域的应用越来越广泛，但算法偏见隐匿其中，悄无声息并深远影响绩效考核的公平性。“深挖”算法偏见形成的原因，辨识其在绩效考核中的主要表现形式，解析算法偏见影响绩效考核公平性的路径，提出相应的对策。算法偏见影响绩效考核公平性的路径包括：数据偏差在模型训练阶段埋下不公平的种子；算法自身缺陷在模型运行阶段放大偏见；考核结果应用阶段因场景的局限性而限制算法响应。基于此提出对应策略，在保持算法赋能绩效考核效率优势的同时，减少偏见带来的消极影响，帮助组织使用更现代化的工具来研究和实施绩效考核。

[关键词]算法偏见；绩效考核；公平性；人工智能；优化策略

绩效考核是组织管理中的重要环节，其公平性直接影响领导威望、组织的凝聚力和员工的工作积极性。 近年来，随着人工智能和大数据技术的普及，算法被引入绩效考核系统，用来提高考核的客观性和效率。 然而，企业人力资源管理数字化转型过程中，与绩效考核范式变革相伴而生的算法偏见，对绩效考核的公平性产生了负面影响。

《中国企业人力资源管理数字化白皮书》显示，2024 年 89% 的 A 股上市公司已经部署算法驱动的绩效考核系统，其中 62% 的企业将算法评分作为晋升决策的核心依据。 技术赋能绩效考核在提升效率的同时，也引发了深层次的挑战——算法偏见导致的不公平及其治理。 技术理性与管理伦理的冲突本质上反映了算法决策的“双重性”：一方面，算法的标准化计算能够规避人类主观偏见，使程序公平感知大幅提升；另一方面，数据采集的历史依赖性、模型设计的目标单一化以及决策过程的不透明性，又在考核结果中植入新的偏见源（王海建，郝宇青，2024）[1]。学术界对算法偏见的早期研究聚焦于算法歧视识别，如李长安（2024）发现，算法在招聘环节对育龄女性的隐性排斥率达 19%[2]；中期转向组织行为学视角，如景怡等（2024）证实算法决策的公平感感知存在任务类型差异 [3]；近期研究开始关注系统影响，魏昕等（2021）指出，算法偏见通过组织承诺中介效应导致

基金项目:2023年广东省创新团队项目“数字经济背景下工商管理研究创新团队”（2023WCXTD041）；2021年广东省普通高校特色创新类项目“人工智能与职业教育深度融合的协同育人平台构建研究”（2021WTSCX247）；广东省教育厅2023年度普通高校重点科研平台（2023WCXTD041）。

作者简介:张旭军，男，汕尾职业技术学院副教授，博士；宋永，男，汕尾职业技术学院副教授；吴学秉，男，汕尾职业技术学院讲师。

员工离职意愿提升 29%[4]。

本文立足人力资源管理与人工智能伦理治理的交叉领域，对算法偏见影响绩效考核公平的路径进行深入探讨，挖掘算法影响绩效考核的情境因素，研究数据偏差、模型缺陷、黑箱特性等技术特征与组织公平的对应关系，提出针对性的干预策略平衡技术效率与管理公平，解构算法偏见的影响路径，为企业设计“可解释、可干预、可适配”的绩效考核系统提供操作指南。

一、文献回顾

当前，组织在绩效考评环节常用的评价方法主要有描述法、量表法、比较法三大类，其中有定性方法，如书面鉴定法、事实记录法；也有定量方法，如多目标决策方法、层次分析法、模糊综合评判法等类型，并衍生出财务指标法（如销售收入指标）、统计分析法（如销售收入增长率指标）等具体手段。 实践当中，绩效管理方法与绩效评估方法并未受到严格区分，如关键绩效指标法（KPI）也经常简化用于绩效评价。 由于该方法聚焦于核心业务目标，应用算法辅助评价时容易出现指标单一化而忽视创新等隐性贡献。 再比如 360 度反馈法，运用这一方法评价绩效，需要整合多类型主体的评价（包括被评价者自己 / 上级 / 同事 / 下级 / 客户等），算法可能会放大历史评分偏差。

鉴于算法在学习、聚类、代码迭代等过程中会形成偏见从而影响绩效考评结果，近年来研究者从不同视角对这一现象开展了广泛的研究。 国内研究者对算法偏见影响绩效考核的研究集中在四大方向：一是算法偏见的形成机制与影响路径。 褚福磊等（2024）发现，历史绩效数据的选择性采集导致女性员工创新能力评估均值降低 18%，直接影响任务绩效 [5]。 景怡等（2024）证实，以财务指标为核心的算法模型对研发岗位创造性贡献的低估率达到 32%。 李长安（2024）指出，三分之二的员工因无法理解算法评分规则，对考核结果的信任度降低了 23%，间接影响员工绩效。 二是情境因素的调节作用。 实证研究发现，算法偏见对创意型岗位分配公平的影响明显高于执行型岗位。 而在高权力距离文化中，员工对算法权威的服从心理会使程序公平感知略有提升，但信息公平感受损更为严重。 三是算法偏见的治理策略。 李长安（2024）提出“算法初评”与“人工复核”双轨机制，把创意型岗位评估的 30% ～ 40% 交给人工，可以有效降低数据偏差影响。 仲理峰等（2019）建议在高权力距离企业明确算法的辅助决策定位，避免技术盲从 [6]。 四是跨学科理论融合研究。 杨学成等（2023）验证了“公平感知行为绩效”的传导机制，发现算法偏见导致员工离职意愿提升 [7]。 刘璇等（2022）为算法公平性检测提供了协同治理路径 [8]。

尽管现有研究已经取得了显著进展，但目前仍存在三大理论缺口：一是缺少“算法偏见形成机制”的系统解析。 二是对算法偏见影响绩效考评结果公平性的机理尚未形成统一认识。 三是治理偏差缺少系统化策略。

本文通过构建“数据治理－算法优化－场景适配－绩效传导”的整合链条，揭示数据偏差、模型缺陷、黑箱特性与公平的对应关系，验证多维公平感知的中介效应。 二是探索构建“技术透明化 + 过滤机制 + 文化适配化”的三维治理框架，提出公平性约束技术（技术层）、人机协同复核（制度层）、算法素养培训（文化层）的协同治理创新策略。

二、算法偏见的形成机制

算法偏见的形成是一个多维度、系统性的过程，其根源可追溯至数据生成、算法设计与应用场景的交互作用，不仅涉及技术层面的结构性缺陷，更与社会文化、制度环境及人类认知偏差深度耦合。

（一）数据偏差：算法偏见的起点

数据质量直接决定算法输出的公正性，如果数据出现偏差，就会直接影响算法的偏见程度。 据统计，五类数据偏差均不同程度地影响算法公正性，如表 1 所示。

从表 1 可知，采样偏差是由于数据未能反映真实情况而产生，使考评者“一叶障目”，不能了解被考评者的绩效全貌。 例如，一个销售团队 80% 的业绩可能仅依赖头部 10% 的高绩效员工完成，那么，算法

会过度重视销售数据而低估新员工的成长潜力，从而对新员工不利。 标注偏差来自评价者的主观偏误，如某评价者会因其潜意识中的性别歧视而在标注“领导力”指标时使之与“男性气质”强关联，导致女性管理者被低估。

数据偏差可能带来的严重后果。2018 年，Uber 自动驾驶测试车撞死行人的原因即测试车缺乏场景训练数据以及算法依赖旧版道路规则，该事件迫使其暂停全球自动驾驶测试。

表 1 影响算法公正性的数据偏差

偏差类型	产生原因	影响方式	影响结果	典型场景/风险
采样偏差	数据分布与真实世界错位	局部数据掩盖全局	系统非公平性后果	固化歧视
标注偏差	人工标注受主观因素影响	标注错误或偏见被算法学习	隐蔽性歧视	情感分析/经验差异
代表性偏差	特定群体数据缺失/过度简化	算法无法识别/处理未覆盖的群体	安全风险或服务排斥。	自动驾驶缺乏特殊天气/行人数据
数据污染	采集错误/恶意篡改	系统性污染扭曲特征关联	优质内容边缘化	推荐系统被虚假点击污染
时间偏差	数据时效性不足	算法固化历史偏见	历史偏见延续	金融信用模型沿用以前数据

（二）设计缺陷：偏见的技术放大

算法的数学模型与设计逻辑可能无意识地强化偏见或创造新偏见，如表 2 所示。

表 2 放大算法偏见的三类设计缺陷

缺陷类型	产生原因	典型风险领域
模型假设偏差	对现实过度简化/扭曲假设	人才评估
公平性设计缺失	开发时优先效率目标	广告投放
开发者认知局限	设计者无意识地嵌入自身价值观	司法量刑

模型假设偏差出现在设计模型时忽略部分影响因素或者人为放大某些影响因素的权重，使得模型本身有缺陷，那么在后期使用过程中，必然产生有偏差的评估结果。 开发者认知局限产生于模型设计者的主观认知，会留下类似于某些程序员设计程序时留的“后门”，这种认知局限性因不易被开发者察觉而更加隐蔽，使绩效评估结果的不公平在其他环节如程序公平等的光环下而隐匿起来，变成“阳光下的‘罪恶’”。

一旦发生算法偏差而导致公平性设计缺失，就可能给企业带来损失。 亚马逊开发的自动简历筛选算法因过度追求招聘效率，对男性简历评分显著高于同等资历女性，导致女性求职者简历被系统性低估而停用并赔偿受害者。

（三）场景限制：偏见的现实投射

算法的实际运行环境无疑会放大偏见或创造新的偏见。 当组织文化或当权者提出特殊的考核要求，就会造成某些应考量项目在考核中缺失。 算法不透明则降低了外部监督的可能性，决策过程难以被公众理解和接受，并出现“算法归因偏差”。 用户与算法的互动产生人机交互偏差，通过用户反馈循环不断强化，形成“算法茧房”。

综上，算法偏见是数据、算法与场景的动态耦合结果，是技术、社会与文化因素交织的产物，需要跨学科的系统性治理，通过技术创新与制度变革的双重驱动加以解决，实现算法的公正性与包容性。

三、算法偏见对绩效考核公平性的影响

鉴于算法偏见的复杂因果链，下面基于技术组织协同理论，从数据输入、算法处理、场景适配三个维度解构其影响，揭示算法偏见侵蚀绩效考核公平性的微观机制。

（一）数据偏差导致不公正评价

数据是算法决策的逻辑起点，其结构性缺陷会通过训练过程转化为系统性评价偏差。 在绩效考核领域，数据偏差主要表现为历史偏见的代际传递与评价维度的结构性失衡。 算法依赖的历史绩效数据常常隐含着组织过往的管理偏见，形成偏见固化效应。 以某汽车零部件企业为例，该公司 2018—2022 年绩效数据库显示：女性技术人员的创新提案采纳率仅为男性的 63%[①]。 回溯原始评审记录发现，23% 的女性创新提案与男性提案相比，质量相当甚至表现更优，评审委员们对女性持有固化偏见而误判了提案质量[②]。这说明，一旦带有隐性歧视的数据进入算法训练集，就会形成“偏见增强回路”。 通过蒙特卡洛模拟发现，若初始数据集的性别偏见率为 15%，经过 3 次迭代后，算法产生的歧视性评价将扩大至 38%[③]。

数据标注过程也存在显著的“管理者过滤”效应，从而产生选择性偏见，这本质上是组织价值观的技术投射——当考核数据过度依赖财务硬指标时，算法会自动弱化诸如创新、社会责任等软指标，形成“效率至上”的评价导向，造成知识型员工的创造性劳动价值被低估。

（二）设计缺陷引发歧视性评价

算法的技术架构与优化目标设计存在天然的公平性盲区。 在特征选择环节，算法设计者常因专业领域知识不足导致关键变量缺失，产生非主观意愿的隐性歧视。 在特征编码过程中，当敏感信息被转化为算法可识别的数字标签时可能产生隐蔽歧视，即使设计者进行了脱敏处理，这些信息仍可能通过与非敏感特征建立关联而形成“统计性歧视”。 当无法理解算法的赋分逻辑时，员工感知的程序公平度较传统考核方式更差，从而降低对组织的信任度。

（三）场景限制带来公平性缺失

权变领导理论指出，管理风格应当与领导情境相适应，以取得良好的组织绩效，领导情境发生变化，管理风格就要随之改变。 但是，算法的技术理性与组织管理的权变性适配不足、人机协同失效。 当组织情境的动态适配出现困境时，固定算法模型难以应对差异化的考核场景，无法处理非结构化目标，出现绩效目标传导失真；如果人机协同失效，人机决策的协同就会出现异化风险。 如果管理者过度依赖算法评分，其主观判断的权重就会下降；算法公平性审查机制的缺失，使纠偏乏力。

算法偏见对绩效考核公平性的影响，本质上是技术理性与管理人性的价值冲突。 组织需构建“数据治理－算法审计－制度适配”的三维治理体系：在数据层面建立过滤机制，阻断偏见的代际传递；在算法层面开发可解释模型，嵌入多维度公平性约束；在制度层面构建人机协同规则，明确算法应用的场景边界与责任归属，从而达到效率与公平兼顾的目的。

四、优化策略

（一）数据治理：构建全流程偏见过滤机制

治理算法偏见需要识别各类偏见，分析其本源，结合生成原因，提供解决之道。 根据前文剖析结果建立“三阶段数据治理模型”，以有效消除偏见。 一是偏见识别，通过卡方检验、对抗去偏算法等手段，发现历史数据结构偏差；二是数据重构，通过增加员工自评、360 度考评等形式，将非结构化内容转变为具有结构化特征的绩点；三是动态监控，建立实时数据审计平台，确定“数据偏差预警阈值”。 下面，以高端装备制造行业为例来说明。 某公司为修正其绩效考核偏差，增加了三个关键数据：其一，技术成熟度曲线的关键数据，目的是解决长周期项目的评价盲区；

①金孟子.女性创新的制度性困境与突破路径 [R]. 北京：北京大学光华管理学院，2024.

②斯坦福社会创新评论。警惕！算法也会“性别歧视”[EB/OL].(2021)[2025-08-07].

③Wilhelm F. Modeling and Mitigating Gender Bias in Matching Problems [C]//FLAIRS Conference Proceedings. Daytona Beach, USA，2025.

其二，质量安全指标的关键数据，防止项目考评过分偏向效率；其三，引入“复购率”“技术服务满意度”等客户生命周期价值指标，动态修正销售考核产生的短视偏差。 通过这些措施，在提升考评效率的同时，公司的绩效公平性也得以保证。

（二）算法优化：嵌入公平性约束技术重构

嵌入关联处理和公平性惩罚项，以保障算法公平。 在特征工程阶段，采用 Fair learn 等工具包，对敏感特征进行去关联处理。 模型训练阶段则在目标函数中加入公平性惩罚项，如 Equalized Odds 约束等，对算法进行优化，为模型运行过程的公平注入保障。 加强评价者培训和考核，规定评价者运用算法的最低标准，为人机协同评价筑牢基础。

（三）场景适配：构建人机协同动态响应

以经典的权变管理理论为指导，在算法中加入领导情景影响因素、工作环境因素等，研发技术复杂度调节模块，根据不同类型岗位自动生成校正权值；应用在线学习算法，每个考核周期依照最新数据分析，即时修改算法模型。 添加算法参数弹性因子，重新拟定人机交互回应体系，设置启动专家评审组打分的阈值，建立溯查复核异类个案机制。

由 HR、IT、业务部门组成跨部门数据治理委员会，负责制订数据标准和算法伦理准则，提高组织架构的适应性。 设置算法审计岗位，定期出具《算法偏见风险评估报告》，并纳入部门 KPI 考核。 对公平性工具链进行整合，实现偏见检测、去偏算法、解释性可视化的一站式操作，同时允许业务部门自主设置一定比例的考核维度，如研发部门可添加“技术预研进度”指标，系统自动匹配算法模型。 通过动态响应机制，降低场景对算法的限制，提高员工的考核公平感。

五、结论

本文通过混合研究方法，系统揭示了算法偏见影响绩效考核公平性的多维作用机制，为数字时代组织绩效管理公平性重构提供了理论参照。 研究发现，算法偏见通过“数据偏差—模型缺陷—场景错配”的传导链条，对绩效考核的结果公平、程序公平与人际公平等维度构成系统性威胁，印证了技术工具与管理情境的互动逻辑，即算法本质是组织管理价值观的数字化投射。

本文在理论方面的贡献：一是将技术组织协同理论引入算法管理研究，揭示了算法偏见的本质，拓展了“统计性歧视”的理论边界。 二是新发现，即技术壁垒等级、客户生命周期等行业特殊指标的缺失会导致绩效评价失真，弥补了算法公平性研究的行业情境化认知缺陷。 值得强调的是，算法公平性优化不是技术完美主义的追求，而是在效率提升与公平保障之间寻找动态平衡点，其核心在于将组织管理的人文关怀嵌入算法设计的技术逻辑。

研究局限性在于未进行实证检验且未涉及生成式 AI 等前沿技术的影响。 随着人力资源管理数字化转型的深化，算法偏见可能从“数据驱动型”转向“模型自主型”，如何构建兼具效率与公平的智能化绩效考核系统，将成为学界与业界共同关注的核心命题。

参考文献：

[1]王海建,郝宇青. 数字治理中公平原则的遮蔽与解蔽——对数字治理效率优先的批判 [J]. 学术界, 2024, (01): 97-105.

[2]李长安,孙雨意,韩威鹏.数字经济时代算法歧视问题研究[J].中国劳动关系学院学报, 2024, 38(04):58-66.

[3]景怡,邱凌云,任润.算法决策对员工公平感的影响研究[J].经济管理学刊, 2024, 3(04):213-236.

[4]魏昕,黄鸣鹏,李欣悦.算法决策、员工公平感与偏差行为：决策有利性的调节作用[J].外国经济与管理, 2021, 43(11):56-69.

[5]褚福磊,刘园园,刘淑桢. 团队资质过剩感对团队创新绩效的影响机理研究 [J]. 经济与管理研究, 2024, 45 (12): 127-141.

[6]仲理峰,孟杰,高蕾. 道德领导对员工创新绩效的影响:社会交换的中介作用和权力距离取向的调节作用 [J]. 管理世界, 2019, 35 (05): 149-160.

[7]杨学成,杨东晓,郭景. 平台劳动者工作满意度与就业稳定性：认知信任与算法管理的影响 [J]. 首都经济贸易大学学报, 2023, 25 (03): 43-57.

[8]刘璇,朝乐门.AI治理中的公平性及其评价方法研究[J].情报资料工作, 2022, (05):24-33.

[9]白惠芬,卫翔,白虎伟等.企业绩效考核中的公平性问题研究[J].内蒙古科技与经济, 2020, (07):37-38.