机器学习与用户行为中的偏差问题：知偏识正的洞察

研究成果：机器学习与用户行为中的偏差问题: 知偏识正的洞察
作者：郭迅华，吴鼎，**卫强，陈国青**
发表期刊：*《管理世界》*，2023 年第 5 期，145-159+199

1. 研究背景和意义

面向大数据和人工智能的科学研究和实践探索在全世界范围内广泛展开，推动了数智赋能的价值创造和模式创新，已成为经济增长和国家发展的新引擎。对此，我国在战略层面上予以了高度重视和前瞻部署。国家“十四五”规划和 2035 年远景目标纲要将大数据列入重点产业，强调推动大数据技术创新，培育大数据全生命周期产业体系。国务院《新一代人工智能发展规划》提出，构建开放协同的人工智能科技创新体系，全面支撑科技、经济、社会发展和国家安全。

大数据环境下基于数智技术的管理决策广泛使用机器学习方法，利用用户行为数据以训练模型。然而，机器学习方法在数据使用和模型训练中不可避免地存在着模型偏差（包括数据观测和提取中的采样偏差以及模型构建中的拟合偏差）；同时，在情境应用中，模型所做出的预测和推荐会对不同用户的感知、态度、行为选择和观点表达产生不同的影响，形成行为偏差（包括用户活动的选择偏差和行为表现的表达偏差）。一方面，行为偏差被带入到用户行为数据中，通过算法训练的过程被机器学习模型所习得，产生新的模型偏差；另一方面，模型偏差在系统应用中被输出，在情境交互中干扰用户的自选择及行动，引发新的行为偏差。这样的循环使得模型偏差和行为偏差往复重塑，持续放大和扩散，不仅会导致模型算法的效力降低，也会干扰用户认知与决策。

在数字经济环境下，对于这些偏差问题的研究是各界关注的前沿课题。偏差问题内涵丰富、成因繁杂、机理复杂，存在着重大的研究挑战。现有研究的视角多样，但方向零散，不利于形成对于偏差问题的系统性认知与体系性解决方案。本文对机器学习和用户行为中的偏差问题进行了系统性的讨论，界定分析了其科学内涵、结构机理、性质特征、实践影响和管理意义，并对若干前沿研究和重要进展进行了梳理凝练，对相关偏差辨知和纠正策略进行了解析阐释，对未来研究方向和探索路径进行了前瞻展望，为学界研究和业界应用提供“知偏识正”的洞察。本文工作为信息系统乃至整个管理学领域的方法创新和赋能创新提供了一个互动式新视角，也为数智赋能管理实践提供了理论与应用的启示。

2. 偏差问题的内涵、结构与挑战

在数智化管理与决策系统中，用户行为数据作为系统输入影响机器学习模型的训练和预测，而机器学习结果的呈现又会影响用户行为与决策，两者相互影响，形成一个循环往复的反馈回路，共同构成数智化管理与决策的基本框架。然而，机器学习与用户行为中的偏差问题也通过这个循环体系不断累积放大，最终可能导致机器学习的失效和行为与决策的困境。

“偏差”（bias）一词内涵丰富。不同学科领域对于“偏差”的侧重和定义有所不同，包括认知偏差、统计偏差、利益冲突、歧视偏见等多种理解。本文所关注的偏差，主要指机器学习与用户行为中的系统性偏误（systematic error），即系统性偏离真实事态的结果或发现，或者引发系统性偏离的某种过程。这主要包括两类偏差问题：行为偏差与模型偏差，二者通过偏差循环机制相互影响。

行为偏差问题聚焦于用户行为产生及数据化过程中的准确性、全面性和一致性，即用户行为数据在多大程度上真实、完整、无干扰地反映了现实世界中的情境要素以及用户在这些要素条件下的属性、认知和行动意图。模型偏差问题则聚焦于模型的正确性、代表性和可靠性，即机器学习模型在多大程度上准确、充分、稳定地拟合了基本事实和逻辑并在此基础上做出预测。偏差循环是指机器学习模型与用户行为数据之间的相互影响机制，主要关注机器学习模型对于用户行为偏差的习得以及模型偏差对于用户认知行为的干扰。

本文对行为偏差和模型偏差的结构以及影响机制进行了分析阐述，提炼出四类基本偏差，即选择偏差、表达偏差、采样偏差与拟合偏差，以及两类偏差循环机制，即偏差习得机制与偏差交互机制，并通过形式化表示分析了偏差问题的一般性特征，结合相应实例展示了偏差问题的情境化表现。

进而，本文总结了偏差问题应对过程中的若干重要挑战，并结合作者团队近年来的若干研究工作，阐释和讨论了具体场景下的应对思路、问题建模与求解策略。所阐述的四个研究主题包括用户行为偏差的干预治理、考虑行为偏差的机器学习建模、考虑模型偏差的智能预测与推荐、人机交互过程中的偏差现象与机理，以期对未来研究的开展提供借鉴意义。

3. 未来研究方向与关键议题

在系统性分析的基础上，本文对偏差问题研究的未来方向和关键议题进行了展望，以期提供前瞻性的理论洞察。总体而言，机器学习与用户行为中的偏差问题结构复杂、影响广泛、重要性凸显。未来研究可以沿着下面三个方向深化和拓展。

第一个方向是用户行为偏差的形成机理与应对策略。该方向上的探索具有三个特点。首先是多情境：在不同商务情境下，用户行为偏差的具体表现和成因有所不同，因此未来研究可以探索揭示不同商务情境下用户行为偏差机理，并建立考虑用户行为偏差的经济计量模型。其次是多角度：用户行为偏差的成因丰富，既包括内在行为动因，也可能包括外部推力影响，因此未来研究可以从不同角度探讨典型行为偏差的成因与机理。第三是多路径：用户行为偏差的治理应当结合事前治理与事后治理，未来研究可以通过人机交互设计等手段探讨事前治理机制，也应探索开发行为记录数据中的偏差检测和校正方法。

第二个方向是模型偏差的防止技术与模型应用策略。该方向的探索可以从两个方面展开。一方面是审

视机器学习模型的算法逻辑与训练样本，对样本数据偏差进行校正或控制，对算法逻辑予以修正和改良。另一方面是在机器学习过程中，对用户行为偏差的形成过程予以建模，在控制偏差的情况下进行更可靠的模型训练。

第三个方向是发展整体性理论，聚焦偏差循环，建立机器学习模型与用户行为偏差的治理体系。整体性治理理论的基本目标是在情境动态演变的过程中，将智能模型对基本事实的拟合偏离程度控制在可接受的范围内，并予以逐步优化。其基本手段是以算法为核心，结合行为科学理论，逐步实现各系统要素间的有机协同。未来研究在智能模型的评价过程中，将更加注重真实的用户评价和长期的模型表现。

4. 研究贡献

本文的贡献可以概括为三个方面：第一，对机器学习与用户行为中的偏差结构进行刻画，提炼出四类基本偏差及其循环机制，并通过形式化表示呈现偏差问题的一般性特征；第二，结合作者团队的相关工作，对偏差问题的若干挑战和研究探索进行阐释，通过具象化描述呈现问题建模和求解路径的学理思路；第三，对偏差问题研究的未来方向和可能议题进行展望，通过前瞻性视野呈现值得学界和业界进一步关注的学术空间和实践创新。

本文中的讨论有助于学界和业界更深刻地解析机器学习/人工智能等现代科技在赋能经济社会活动中的作用，形成对于偏差问题的系统性认知（即“知偏”）和体系性解决方案（即“识正”），促进理论和实践层面的探索创新，为我国以及全球数字经济的健康发展贡献力量。

5. 思考与启示

在大数据与人工智能的时代，算法技术与人类行为相互依存、相互塑造。“数智赋能”的管理学研究，离不开技术视角与行为视角的紧密融合。本文所探讨的偏差问题，凸显了新时代对管理学研究领域范式所带来的新机遇和新挑战。特别是生成式人工智能大模型技术的快速发展，其在智能“涌现”的同时也伴生着“AI 幻觉”，进一步凸显偏差问题的重要性。

针对这一现象，本文认为，一方面，算法技术的设计与开发，需要以行为规律作为基础依托和根本驱动力；另一方面，对行为规律和机理的探索和揭示，离不开对算法技术形态属性的审视和理解。唯其如此，我们才能洞察行为机理的演化，把握算法技术的潜力与风险。技术视角与行为视角的协同交互，不仅是应对本文所探讨的偏差问题的关键，更是管理学研究与实践在“数智赋能”时代中突破瓶颈、跃迁发展的方向所在。

供稿：科研事务办公室　编辑：高晨卉　责编：吴淑媛 赵霞

School of Economics and Management
Tsinghua University
Beijing, China 100084
Tel 8610-6278 5515
Fax 8610-6278 5876
清华大学经济管理学院
中国·北京 100084
电话 8610-6278 5515
传真 8610-6278 5876