arXiv:2304.02351v1 [cs.MA] 5 Apr 2023

偏见的构建与解构：基于智能体的模拟中对特权与指导关系的建模

AndriaL.Smith[*]（**asmith@is.mpg.de**）马克斯·普朗克智能系统研究所，Heisenbergstr. 3，斯图加特，巴登-符腾堡州¨70569，德国

SimonHeuschkel[*]（**simon.heuschkel@student.uni-tuebingen.de**）图宾根大学，Maria-von-Linden-Str. 6，图宾根，巴登-符腾堡州¨72076，德国

Ksenia Keplinger（**kkeplinger@is.mpg.de**）马克斯·普朗克智能系统研究所，Heisenbergstr. 3，斯图加特，巴登-符腾堡州¨70569，德国

Charley M. Wu（**charley.wu@uni-tuebingen.de**）图宾根大学，Maria-von-Linden-Str. 6，图宾根，巴登-符腾堡州¨72076，德国

[*] 这些作者贡献相同

摘要

在挑选领导者、判断谁具有影响力以及选择与谁互动时，偏见不仅存在于社会中，也存在于组织环境中。借鉴领导者涌现理论和社会影响理论，我们探讨了能够支持多元化领导者的潜在干预措施。我们使用基于智能体的模拟，在适应度景观上对集体搜索过程进行建模。智能体结合个体学习与社会学习，并以特征向量表示，其中融合了相关特征（如个体学习特征）与无关特征（如种族或性别）。智能体依据表现预测，运用理性学习原则估算各项特征的权重，并据此动态确定其网络中的社会影响力。我们展示了偏见如何基于历史特权产生，以及如何通过干预措施（如导师指导）大幅减少偏见。本研究揭示了偏见形成与消解背后的认知机制，并为未来实证研究中有待检验的现实干预措施提供了重要见解。

关键词： 社会影响；偏见；特权；社会网络；干预

引言

偏见存在于我们如何挑选领导者、受谁影响以及与谁互动之中。例如，在标准普尔1500指数成分公司中，名叫约翰或戴维的首席执行官人数多于女性首席执行官（Johnson, Hekman, & Chan, 2016）。尽管人们日益关注如何营造更加多元、包容的组织环境，但偏见等诸多障碍依然阻碍着进步（Keplinger & Smith, 2022）。

尽管关于偏见来源及消除偏见的潜在干预措施的实证研究历史悠久（Axelrod, 1997; Freeman, Penner, Saperstein, Scheutz, & Ambady, 2011; Serban et al., 2015; Schelling, 1971），但特权与偏见在何时、为何以及如何产生，仍不明确（Colella, Hebl, & King, 2017）。具体而言，我们关注特权如何阻碍边缘化领导者的涌现（Badura, Galvin, & Lee, 2022）。此处，特权被定义为特定群体无需付出相应努力即可获得奖励和资源（Case, Iuzzini, & Hopkins, 2012; Crevani, 2019）。为整合有关领导者涌现与社会影响的理论，我们采用基于智能体的模拟；这是一种计算方法，在领导力研究领域仍鲜有应用（但参见 Cao et al., 2020）。这些模拟使先前的文字理论更加精确（Samuelson et al., in press; Vancouver, Wang, & Li, 2020），并揭示了对任意智能体特征（如种族、性别、年龄等）形成偏见背后的认知机制。

本研究旨在：1）展示当某些智能体被置于环境中的特权位置时，偏见如何通过多智能体学习的理性原则再度形成；2）探究一种干预措施的影响：我们将高绩效与低绩效智能体之间的临时社会网络连接建模为外部智能体（例如导师或榜样），以消除偏见。我们的模拟表明，我们能够系统性地减少所有智能体的偏见，从而增加新兴领导者的多元化代表性。

方法

我们采用基于智能体的模拟，由一支包含 7 个智能体的团队共同优化具有多个局部最优解的二维适应度景观。我们使用 Ackley 和 Drop Wave 环境作为优化算法的常见测试函数（Surjanovic & Bingham, 2013），此外还使用 Mason 和 Watts（2012）的环境；以往研究将这些环境描述为具有相似的平均收益和

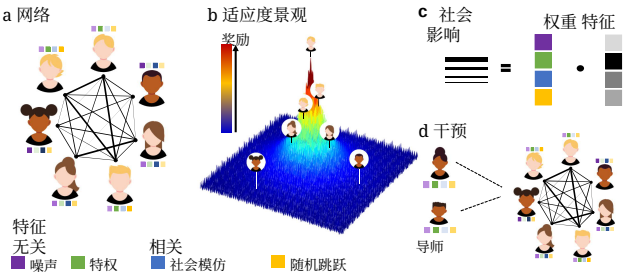

图 1： 基于智能体的模拟。**a)** 网络结构和智能体特征。**b)** 适应度景观。**c)** 社会影响力学习与所学特征权重的关系。**d)** 通过导师指导进行干预。

局部最优解数量（Barkoczi, Analytis, & Wu, 2016）。我们将每个景观定义为包含 $1000 \cdot 1000$ 个离散位置。

智能体由加权的社会影响力矩阵A连接，并由一组特征定义（图 1a）。在每次迭代中，智能体使用结合个体学习与社会学习的搜索策略更新其在适应度景观$\mathbf{x}_i$上的位置以及相应的适应度值y_i（图 1b），然后根据学到的特征权重W更新社会影响力（图 1c）。

社会影响力。每个智能体i由一组特征$\mathbf{f}_i = [\gamma, \eta, \rho, \nu]$定义，这些特征代表个人属性，可能与策略相关（即反映学习策略的特征：γ、η），表明特权（即初始条件 ρ），或完全无关（即噪声 ν）。

在每次迭代中，每个智能体都会尝试通过对其他智能体j的特征$\hat{r}_{i,j} = \mathbf{w}_i^\top \mathbf{f}_j + \varepsilon$进行线性加权求和，来估计其表现，其中 $\varepsilon = \min_j r_j$是群体中所有智能体当前最低奖励的偏移量。其他智能体的表现通过相对于当前时间点T进行时间折扣后的历史奖励来衡量：$r_{i,T} = (\sum_{t=0}^{T} y_{i,t} \cdot \lambda^{T-t}) / \sum_{t=0}^{T} \lambda^t$，其中时间折扣设为 $\lambda = 0.9$。通过梯度下降，最小化实际奖励与预测奖励之间的均方预测误差，以更新权重：

$$\mathbf{w}_i \leftarrow \mathbf{w}_i - \alpha \cdot \frac{\partial}{\partial \mathbf{w}_i} \mathcal{L}_{MSE}(\mathbf{r}, \hat{\mathbf{r}}_i) \tag{1}$$

学习率为 $\alpha = 0.1$。

因此，权重的学习反映了社会影响中针对特征的偏差：权重较高的特征会使智能体产生更大的社会影响。我们在每次迭代中根据预测表现更新社会影响矩阵：$A \leftarrow A + \beta \cdot W \cdot F^\top$，其中 $\beta = 0.5$用于控制更新率。

个体与社会学习。智能体结合社会学习和个体学习策略，以优化其在适应度景观中的位置。每个智能体i首先以$P(\gamma_i)$的概率进行社会模仿，并根据社会影响权重采用 softmax 模仿策略：

$$\pi_{\text{imitation}}(\mathbf{x}_j) \propto exp(\frac{a_{i,j}}{\sum_k a_{i,k}} / \tau) \tag{2}$$

直观而言，智能体更有可能模仿那些被认为具有较高影响力的其他个体$a_{i,j}$。

如果未实施社会政策，智能体会以概率 $1 - P(\gamma_i)$采用个体学习。首先，智能体会以$P(\eta_i)$的概率尝试随机跳跃：评估搜索空间中的一个随机位置，并在该位置能提高其适应度值时跳转过去。如果智能体不进行随机跳跃，则会使用随机爬山法（SHC）对所有相邻位置进行局部优化：

$$\pi_{\text{SHC}}(\mathbf{x}') \propto \exp(y_i' / \tau), \tag{3}$$

其中，每个y_i都是相邻解$\mathbf{x}'$的适应度值，且适应度值较高的解更有可能被选中。所有情况下，我们都设定 $\tau = 0.01$。

政策相关特征 γ 和 η 从$\mathcal{U}(0, 0.1)$中均匀采样，而特权 ρ 和噪声 ν 则从具有两个不同均值$\in (0.03, 0.07)$的多峰高斯分布中采样，每个均值对应的方差为 0.01，并在 $[0, 0.1]$之间截断。这两项政策相关特征都通过帮助逃离局部最优来影响表现，但 γ 取决于社会信息的质量。

特权与干预。为模拟特权，我们使用 ρ 来定义智能体的起始位置，将其置于奖励排名前$(\rho \times 10)$分位数 ± 0.005对应的位置。这种初始化使特权与模拟开始时的表现呈正相关（例如，$\rho = 0.5$ 将在奖励景观中接近中位数的位置开始），但不会影响学习能力。

我们假设，智能体会形成一种偏见，倾向于为特权学习较大的特征权重。因此，我们引入了一种导师指导干预。导师被建模为特权较少的智能体 $\rho \sim \mathcal{N}(0.03, 0.01)$，但具有较高的政策相关特征 $\eta, \gamma \sim \mathcal{N}(0.08, .005) \in [0, 0.1]$ 和较高的表现 r_{mentor} >90% of all fitness values。在每次迭代中，特权较少的智能体（$\rho < .05$）有 0.2 的概率被分配一位导师。除群体成员外，还会利用导师的特征和r_{mentor}来优化受指导者的社会权重$\mathbf{w}_i$。因此，导师不参与集体优化，也不会成为社会模仿的对象或社会影响网络的一部分，而仅用于支持社会特征学习（式 1）。不过，导师会向群体表明，他们意识到噪声特征的存在，也意识到某些偏见可以被打破（Freeman Jr & Kochan, 2019；Ivey & Dupr´e, 2022；Raza & Onyesoh, 2020；Williams et al., 2020）。

结果

我们在每种环境中进行了 1000 次模拟，每次模拟包含 150 次迭代。图2显示，尽管智能体起初没有偏见，但到了第 25 次迭代，它们便学会了为特权特征赋予较高的权重 ρ。随着模拟继续进行，随机跳跃 η 被证明是有用的，智能体也学会了为这一特征赋予越来越高的权重。以 γ 形式出现的社会模仿起初似乎不受重视，其权重持续下降，直到干预之后才有所变化——

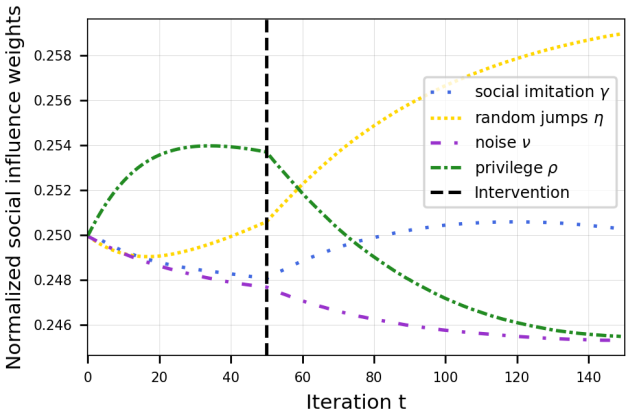

图2： 结果。干预对不同适应度景观中特征平均权重的影响。归一化权重表示各特征对绩效的贡献程度。

这可能是由于较高的特权权重产生了掩蔽效应。

干预后（垂直虚线处），特权权重显著下降，而社会模仿权重开始上升。这表明，导师不仅能降低策略无关特征（即特权）的影响，还能帮助智能体学会更多依赖策略相关特征γ和η。尽管干预后特权和噪声权重下降得更快，智能体对特权的依赖仍高于对噪声的依赖ρ ν，这说明要彻底摆脱习得的偏见有多么困难。没有干预的模拟中，特权权重也会下降，但幅度小得多，而且无法促使社会模仿权重上升。

结论

我们提出的计算方法为理解学习的理性原则如何影响偏见的形成提供了一种工具，也揭示了这些原则如何决定绩效应归功于哪些特征。我们展示了偏见如何基于历史形成的特权自然产生，以及如何通过为表现较好的边缘化智能体与表现较差的边缘化智能体建立指导关系来加以缓解。本研究深入揭示了偏见形成与消除背后的认知机制，并指出了可在未来实证研究中检验的现实干预措施。

致谢

本研究获得德国联邦教育与研究部（BMBF）资助：T¨ übingen AI Center，资助编号：01IS18039A；并由德国研究联合会（DFG）在德国卓越战略项目 EXC2064/1–390727645 项下资助。

参考文献

Axelrod，R.（1997）。文化的传播：一种具有局部趋同和全球极化特征的模型。冲突解决期刊，*41*（2），203–226。

Badura, K. L., Galvin, B. M., & Lee, M. Y. (2022). 领导力的涌现：一项综合性综述。应用心理学杂志, *107*(11), 2069. Barkoczi, D., Analytis, P. P., & Wu, C. (2016). 起伏地形中的集体搜索：跨环境分析。载于第*38*届认知科学学会年会论文集 (pp. 918–923). Cao, S., MacLaren, N. G., Cao, Y., Dong, Y., Sayama, H., Yammarino, F. J., . . . 等人 (2020). 集体中的领导者涌现与领导力感知的基于智能体模型。复杂性,*2020*. Case, K. A., Iuzzini, J., & Hopkins, M. (2012). 特权体系：交叉性、意识与应用。社会问题杂志,*68*(1), 1–10. Colella, A., Hebl, M., & King, E. (2017). 《应用心理学杂志》百年歧视研究：一份发人深省的综述。应用心理学杂志,*102*(3), 500. Crevani, L. (2019). 场域中的特权：组织实践如何促成特权在场域中的交织。斯堪的纳维亚管理学杂志,*35*(2), 101035. Freeman, J. B., Penner, A. M., Saperstein, A., Scheutz, M., & Ambady, N. (2011). 看起来像那么回事：社会地位线索塑造种族认知。*PloS one*,*6*(9), e25107. Freeman Jr, S., & Kochan, F. (2019). 探讨高等教育中跨性别、种族与世代的导师指导：一项民族志研究。国际导师指导与教练教育杂志. Ivey, G. W., & Dupr´e, K. E. (2022). 职场导师指导：批判性综述。职业发展杂志, *49*(3), 714– 729. Johnson, S. K., Hekman, D. R., & Chan, E. T. (2016). 如果候选人中只有一位女性，那么她在统计意义上几乎不可能被录用。《哈佛商业评论》,*26*(04), 1–7. Keplinger, K., & Smith, A. (2022). 职场女性的污名化：污名的来源及其在个人、组织和社会层面的后果。载于多元化实践 (pp. 23–38). Emerald Publishing Li mited. Mason, W., & Watts, D. J. (2012). 网络中的协作学习。美国国家科学院院刊, *109*(3), 764–769. doi: 10.1073/pnas.1110069108 Raza, A., & Onyesoh, K. (2020). NHS高级领导者的反向导师指导：一种新型关系。未来医疗保健杂志, *7*(1), 94. Samuelson, H. L., Lee, J., Wessel, J. L., Grand, J. A., Samuel- son, H., Lee, J., . . . Grand, J. (in press). 组织多元化与包容性的计算建模。Schelling, T. C. (1971). 隔离的动态模型。数学社会学杂志,*1*(2), 143–186. Serban, A., Yammarino, F. J., Dionne, S. D., Kahai, S. S., Hao, C., McHugh, K. A., . . . Peterson, D. R. (2015). 面对面与虚拟团队中的领导力涌现：一个包含基于智能体模拟、准

实验和实验测试。领导力季刊,*26*(3), 402–418. Surjanovic, S., & Bingham, D. (2013). 模拟实验虚拟库：测试函数与数据集。 检索于2023年3月31日，网址：http://www.sfu.ca/ ssurjano. Vancouver, J. B., Wang, M., & Li, X. (2020). 将非正式理论转化为形式理论：工作动机综合模型动态计算模型的案例。组织研究方法, *23*(2), 238–274. Williams, N., Ravenell, J., Duncan, A. F., Butler, M., Jean- Louis, G., & Kalet, A. (2020). 同伴导师培养项目：种族/族裔少数群体教师导师指导的经验教训。族裔与疾病, *30*(2), 321.