聚知刊出版社
JZK publishing

算法透明度对感知公平的影响

陈一帆

苏州大学，江苏省苏州市，215006；

摘要：尽管人工智能在人力资源实践中的应用越来越多，但求职者往往不明白算法是如何做出决策的。本文研究算法透明度是否以及如何影响感知公平。通过实验，我们发现透明度能提高求职者的感知公平，结果有利性起了调节作用。

关键词：透明度；感知公平；结果

DOI：10.69979/3029-2700.24.10.020

引言

从贷款申请到精准投放广告，人工智能在现代社会的应用越来越广泛。越来越多的商业组织正在整合人工智能，以改善人力资源决策和自动化人力资源管理活动。人工智能之所以有吸引力，是因为它能够在技术、组织和制度框架内自主行动。简而言之，人工智能可以帮助人力资源专业人员更有效地执行人力资源相关的任务。

算法通常是黑匣子，其内部工作原理无法观察或理解。算法在各种环境中的使用可能会给作为最终用户和数据主体的个人和社会带来认知上的挑战。如果不了解算法是如何工作的，消费者无法就是否依赖算法判断做出明智的决定。这种不透明性对公平感构成了严重挑战，这导致人们广泛呼吁提高算法决策系统的透明度。透明度应该被理解为一种认知关注，与认知不透明的概念相对。

尽管人们对人工智能的兴趣日益浓厚，但很少有研究调查如何通过解决算法驱动的人力资源管理的黑箱问题来改善求职者对人工智能的反应。本文以人力资源管理为研究背景，证明了透明度能提高求职者的感知公平，但这主要取决于算法的结果。积极结果下，透明度对感知公平有积极影响，消极结果下，透明度对感知公平有消极影响。

1 理论与假设

1.1 算法透明度和感知公平

组织有责任为所有求职者提供公正的过程，在人力资源管理的情境下，公平可以定义为基于人们的表现或需求平等或公平地对待每个人。组织公平文献讨论了不同类型的公平，包括程序公平和分配公平。程序公正是指工作场所中与资源分配过程相关的公平性，分配公正是指工作场所中与资源实际分配结果相关的公平性。

现存文献中提到的一个伦理机会是通过在整个过程中及时向申请人提供更新和反馈来提高透明度，这可以被认为是公平待遇的一个要素。通常情况下，公司无法及时提供相关信息，或者除了确认已收到候选人的申请外，不提供任何信息，算法的预测和决策过程是不透明的。GDPR 还保证了“解释权”，人们可以通过该权利要求解释算法做出的决定。

可解释性是指向用户解释算法系统所做决策的能力。可解释性的主要目标之一是实现透明度，从而建立系统的可信度。可解释性通常会提高对系统的依赖和信任，公平感知，人与人工智能协作，任务效率，以及用户对系统故障的理解。关于可解释性对用户行为影响的研究主要集中在当代XAI方法如何影响用户对人工智能系统的感知。基于敏感性和基于案例的解释被视为系统的局部解释（结果解释），试图证明为什么特定案例会收到特定结果，而基于输入影响和基于人口统计的解释被视为全局解释（模型解释），描述系统的模型及其工作原理。全局解释增强了用户的公平感，而局部解释更有效地揭示了不同案例之间的公平感差异，并且用户对此类系统的先前信任程度会影响他们对解释的反应。

先前的文献表明，拥有获取信息的能力可以培养授权的心理机制。解释算法如何工作可以产生赋权的心理体验，因为它让人们能够控制获取知识的过程，从而更好地理解算法的决策。透明度和心理授权都可以改善用户对系统的感知，并可能影响他们使用系统的意愿。因此，我们提出如下假设：

H1：算法透明度对感知公平有积极影响。

1.2 结果有利性

对公平的感知也可能取决于算法的结果。根据社会监控系统(SMS)模型，人们有一种归属于他人的驱动力，这种排他性促使他们更多地关注与拒绝相关的社会信

息。因此，在收到积极的结果后，求职者可能不会注意到招聘过程是如何进行的，因为这种积极的结果足以满足人们归属感的需要。但是，在负面结果的情况下，SMS 模型预测人们会更加关注与拒绝相关的社会信息，之前的研究也证明了消极结果下人们更有可能反对算法的意见。此外，人们趋向于认为积极的结果是公平的，消极的结果是不公平的。

消费者经常对事件、行动和行为的原因做出推论，并将行为或结果归因于内部或外部原因。收到有利决策的消费者会有进行内部归因的动机，决策者的类型影响消费者对不同的决策结果进行归因推断，这会对消费者的态度产生影响。在自动化的背景下，消费环境的内部或者外部归因解释了他们的产品偏好，在某种程度上人们更容易将有利的消费结果归因于内部，形成对产品更积极的态度。归因理论提出，人们有动机以一种自我服务的方式对自我相关的结果进行归因:为了保持或增强自我价值，人们有动机将有利的结果归因于自己(即“内部归因”)，并将不利结果归因于外部因素(即“外部归因”)。

用户对系统的公平性感知主要受系统输出的影响，系统结果超过解释效果。当系统产生“负面”输出时，无论提供何种解释，它都可能被视为不公平。当系统产生“正面”输出时，它可被视为公平，并且不同解释程度都有效。在候选人认为自己可能适合该职位并期望被推荐的情况下，系统的输出（“积极”或“消极”）可能超过解释的效果，而当候选人没有被推荐时，他们认为输入（候选人描述）和系统的输出之间没有关联。因此，我们提出如下假设：

H2：结果有利性调节了算法透明度对感知公平的影响。

2 实验设计与结果

2.1 实验一

1）实验目的

实验 1 的目的是检验算法透明度（高 vs. 低）对感知公平的影响。

2）被试

本研究从见数平台招募了 240 名被试，最终样本包括 205 名完成调查的参与者。

3）实验流程

被试首先阅读对应的算法筛选简历情境材料。假设参与者发现了冰泉公司的招聘广告，并且他们有兴趣申请这份工作。

低透明度组不提供任何算法信息，高透明组呈现的内容如下：智能招聘算法“IRA-N”基于反向传播神经网络进行电子简历的智能筛选，该算法将应聘人员的综合素质分成不同模块，计算机程序会自动选择最合适的申请人进行面试。

以上材料改编自 Newman 等人(2020)的研究，被试在阅读完情境材料后被要求回答注意力检查题目。随后，参与者报告了相关题项。

4）变量测量

感知透明度的量表改编自 Mourali 等人(2024) 的研究，题项为“我觉得我很了解招聘算法是如何做出决定的；招聘算法如何做出决定是非常清楚的；招聘算法的决策过程是透明的。”

感知公平量表改编自 Conlon 等人(2004)的研究，共包含 3 个题项，如“我觉得招聘过程是公平的；我觉得自己可以有效地使用招聘算法；我觉得自己可以理解招聘算法的决定。”

5）结果与讨论

操纵性检验。为检验透明度是否操纵成功，以透明度为自变量，感知透明度为因变量，进行单因素方差分析。结果表明，不同透明度之间具有显著差异 ($F(1,204)=10.59$, $p=.004$)，低透明度组 ($M_{低透明度}=4.03$, $SD_{低透明度}=1.25$) 的感知透明度显著低于高透明度组 ($M_{高透明度}=5.48$, $SD_{高透明度}=1.05$)。

假设检验。以透明度为自变量，感知公平为因变量，进行单变量分析。结果表明，透明度高低对感知公平具有显著差异 ($F(1,204)=2.97$, $p=.039$)，高透明度组被试的感知公平评分($M=5.56$, $SD=1.36$)显著高于低透明度组($M=4.95$, $SD=2.18$)。假设 H1 得证。

2.2 实验二

1）实验目的

实验 2 在实验 1 的基础上进一步探讨结果有利性在其中发挥的调节作用。

2）实验设计

我们通过 Credamo 平台招募被试，实时剔除没有通过注意检查的被试，最终得到 180 份有效数据。

3）实验流程

在实验一的基础上，向被试揭示算法决策的结果。

积极结果：恭喜您，通过简历筛选！

消极结果：很遗憾，您的资质不符合岗位需求。

4）结果与讨论

操纵性检验。为检验透明度是否操纵成功，以透明度为自变量，感知透明度为因变量，进行单因素方差分析。结果表明，不同透明度之间具有显著差异 ($F(1,179)=6.79$, $p=.005$)，高透明度组 ($M_{高透明度}=5.83$, $SD_{高透明度}=2.25$) 的感知透明度显著高于高

透明度组（M 低透明度=5.03，SD 低透明度=2.05）。

假设检验。以透明度为自变量，感知公平为因变量，进行单变量分析。结果表明，透明度对感知公平具有显著差异（F(1,179)=3.08，p=.021）。

以透明度为自变量，感知公平为因变量，使用 Bootstrap 程序检验授权的中介效应，样本量为 10000，模型为 1。检验结果显示，在 95% 的置信区间下，置信区间结果不包含 0（LLCI=-.9611，ULCI=-.2168），调节效应显著。假设 H2 得证。

3 结论、启示与展望

3.1 讨论

我们的研究为人力资源管理中算法透明度提供了理论和实践意义，并为政策实施提供了关于如何披露招聘算法信息的见解，尤其是涉及到消极结果。目前的研究表明，算法透明度对感知公平有积极的影响，算法透明度的有效性取决于算法结果。消费者对算法驱动下的人力资源管理的反应取决于决策结果的有利性，这种影响是由不同的归因驱动的：消费者发现相对更难内化算法做出的有利决定，而容易将不利因素外部化。

3.2 理论贡献

我们的研究扩展了算法解释对感知的影响。先前的研究主要基于能力感知的途径，关注各种情境下的消费者反应。相比之下，可解释性感知同样值得关注。为了填补这一空白，本研究重点关注可解释性对用户感知的影响，并指出人工智能系统的可解释性也会推动求职者的反应。

其次，我们的研究为市场营销和消费者行为领域关于消费者对算法系统的接受度的文献做出了贡献。算法的黑盒性质可能会使人们不愿意依赖算法的决定，先前的文献研究了打开黑盒的方法，并提供帮助消费者理解算法如何工作的知识，从而增加他们对算法的接受度。我们扩展了这方面的文献，表明通过对决策结果提供充分的解释也可以减轻对算法系统的厌恶。

3.3 实践贡献

本文对人力资源部门具有重要意义。公司应该发展可解释的人工智能应用程序，模型的透明度是提高感知公平的关键。此外，可以帮助理解基于算法的人力资源决策如何既高效，同时又合乎道德。

为了保护专有信息，组织可能希望避免解释算法决策。然而，我们的研究结果表明，企业不必担心潜在的披露要求。相反，解释有可能增强消费者对组织的信任。这些研究结果还为寻求对算法决策提供有效解释的组织提出了明确的建议。具体而言，公司应该选择包含更具可操作性的变量，通过输入因素-输出结果之间的因果关系，改善消费者的反应。最后，我们的结果表明，对于产生负面结果的决策，应该优先努力做出解释，减少对算法的厌恶。

3.4 未来研究

未来研究可以集中在两个方向：(a)探索解释对信息披露的影响；(b)探索不同信息解释类型的影响。如何更好地解释人工智能模型的结果是一项具有挑战性的任务，这一领域值得进一步研究。

参考文献

[1]包艳，马伟博，廖建桥，赵海涛．算法权力在管理领域的研究回顾、探索与展望[j/ol]．外国经济与管理。

[2]刘潇肖,鲁辰扬,薛贺.求职者对AI面试的接受度及其影响因素:基于公平视角的质性研究[J].中国人力资源开发,2023,40(03):117-130.DOI:10.16471/j.cnki.11-2822/c.2023.3.008.

[3]裴嘉良,刘善仕,钟楚燕,等.AI算法决策能提高员工的程序公平感知吗？[J].外国经济与管理,2021,43(11):41-55.DOI:10.16538/j.cnki.fem.20210818.103.

[4]Ostinelli M, Bonezzi A, Lisjak M. Unintended effects of algorithmic transparency: The mere prospect of an explanation can foster the illusion of understanding how an algorithm works [J]. Journal of Consumer Psychology, 2024.

[5]Alufaisan Y, Marusich L R, Bakdash J Z, et al. Does explainable artificial intelligence improve human decision-making?[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(8): 6618-6626.

作者简介：陈一帆（1998.10-），女，汉，江苏常州，学生，硕士，苏州大学，研究方向消费者行为。