# 人工智能的偏见与公平性

姓名：张鹏杰 学号：3220231268

**摘要：**随着人工智能（AI）系统在日常生活中的广泛使用，对算法公平性的考量成为不同研究领域关注的焦点问题。由于目前的人工智能系统大多是数据驱动，数据中的偏见对模型公平性有很大影响，此外模型本身的设计也可能导致偏见的产生，由此，本文梳理了人工智能中偏见的类型，并总结了相关的解决方法。

**关键字：**人工智能，数据偏见，公平性

# 引言

人工智能算法已经渗透到我们日常生活的[方方](http://renwu.hexun.com/figure_1792.shtml" \t "https://tech.hexun.com/2021-05-18/_blank)面面。算法作出电影推荐，提供购买产品的建议，并越来越多地用于贷款申请[1]、雇用[2]等高风险决策中。算法决策有明显的好处，与人不同，机器不会变得疲劳或无聊，并且可以比人纳入更多数量级的考虑因素，但是，就像人一样，算法容易受到偏见的影响，这些偏见会使他们的决策“不公平” 。

在决策过程中，公平是指根据个人或群体的固有或后天特性，对个人或群体不存在任何偏见或偏爱，因此，一种不公平的算法是其决策偏向特定人群的算法[3]。一个典型的例子来自美国法院用来做出假释判决的工具，替代性惩戒者犯罪管理剖析软件（COMPAS）可以衡量一个人再次犯下另一种罪行的风险，法官使用COMPAS来决定是释放罪犯还是将他或她关在监狱里。对该软件的一项调查发现，这个软件存在针对非裔美国人的偏见：COMPAS给非裔美国人罪犯分配的风险得分比给具有相同特征的白种人更高的风险得分。在其他领域也存在相似现象，例如用于评选选美冠军的AI系统存在针对肤色较黑的参赛者的偏见，或[数码](http://tech.hexun.com/" \t "https://tech.hexun.com/2021-05-18/_blank)相机中的面部识别软件过度预报亚洲人在眨眼。这些偏见的预测源于数据或算法中隐藏或忽略的偏见。

在本文中，我们确定了机器学习结果不公平的两个潜在因素，即数据偏见和算法偏见[3]。我们回顾了有关数据偏见如何歪曲机器学习算法所学知识的研究，以及算法本身的工作方式上的细微差别（即使数据是无偏见的），这些细微差别使它们无法做出公平的决策。

# 人工智能算法的偏见

人工智能算法的偏见可以归结为两个潜在原因，即数据偏见和算法偏见。

大多数人工智能算法都是数据驱动的，因此数据与算法的功能是紧密耦合的，在训练数据存在偏差的情况下，训练它们的算法将学习到这些偏差，并将偏差反映到算法的结果中。因此，数据中固有的偏差可能会影响使用数据的算法，产生有偏见的结果，算法甚至可以放大数据中已有的偏见。

此外，算法本身的运行机理存在差别，这些差别可能导致偏见（即使数据是无偏见的），由此使算法无法做出公平的决策。

因此，根据偏见的来源，可以将人工智能算法的偏见分为三类：数据到算法、算法到用户、用户到数据。

2.1 数据到算法

本章介绍数据对算法的直接影响，即收集到的数据中存在固有的问题，与真实预期分布存在偏差，这种偏差导致了算法的偏见。

**2.1.1 测量偏差**

在数据收集过程中，测量的指标存在问题[4]。例如在再犯罪奉献预测工具（COMPAS）中，研究人员将过去受逮捕以及周围人受逮捕的次数作为变量衡量一个人的“危险”程度，但是真实情况是少数族裔社区受到更多的监管，由此有更高的逮捕率，根据该指标将所有少数族裔划分为“危险”的群体是一种错误的结论。

**2.1.2 省略变量偏差**

当一个或多个重要变量被排除在模型之外时，就会存在省略变量偏差[5]。例如设计一个模型用于预测客户停止订阅服务的概率，其变量包括其服务的各项指标，倘若此时出现一个新的、强大的竞争对手，导致大量用户停止订阅，那么原有的基于服务指标的预测模型将无法产生正确的预测。

**2.1.3 表示偏差**

表示偏差产生于数据收集过程中，在从整体抽样时缺乏了部分关键数据，使得收集数据集无法反映整体分布[4]。如数据集缺乏地域多样性，在东西方文化方面存在偏见。

**2.1.4 聚合偏差**

从整个群体中得出的一般性结论不适用于所有个体时，就会有聚合偏差。这种偏差在医学诊断中较为常见，如被广泛用于诊断和检测糖尿病的糖化血红蛋白水平在不同性别和种族间存在明显差异，如果只基于训练数据得出一般性的预测结果，模型在具体应用时必然存在偏差。

**2.1.5 抽样偏差**

抽样偏差来源于对子组的非随机抽样，即对不同子组的抽样策略不均匀。例如，对其中一个子组抽样更多的数据，而对其他子组抽样更少的数据，抽样策略的偏向性使不同组的数据量不同，从而使数据更多的组占有更高的权重。

2.2 算法到用户

算法可以影响用户的行为，算法中的任何偏差都可能在用户行为中引入偏差。在本节中，我们将讨论由于算法结果影响用户行为的偏见。

**2.2.1 算法偏差**

算法偏差是指在输入数据中不存在偏差，而纯粹由算法导致的偏差[6]。算法设计的选择，如使用某些优化函数，正则化等，都可能导致有偏见的算法决策。

**2.2.2 用户交互偏见**

用户交互偏见产生于系统的交互界面，用户在与系统交互的过程中，受到窗口界面的引导，在交互过程中会产生偏见的选择[6]。如在Web搜索引擎汇总，排名最高的结果是最相关、最重要的，由此会使用户更多地点击。

**2.2.3 热门偏见**

受欢迎的事物会显露更多的信息，由此使人们更多地了解，由此可能产生偏见[7]。例如线上购物系统中，被更多购买的物品会更受人追捧，而这种偏见是可以通过算法控制产生的，如刷单。

**2.2.4 自发偏差**

在系统投入使用一段时间后，受到用户的能力、性格和习惯的影响，在后续系统更新时，系统可能更偏向于使自身更适应一类群体，从而使系统具有偏见[8]。

**2.2.5 评价偏差**

模型评价时使用了不适当的基准，例如面部识别中的Adience和IJB-A基准[4]。

2.3 用户到数据

许多用于训练模型的数据都是用户生成的，用户的任何固有偏见都可能反映在他们生成的数据中。此外，当用户行为受到算法影响时，该算法中存在的任何偏差都可能在数据生成时引入偏差。

**2.3.1 历史偏差**

历史偏见是世界上已经存在的偏见和社会技术问题，即使有完美的采样和特征选择，历史偏见也会在数据生成过程中渗入[4]。在2018年的图片搜索中，搜索女性CEO时会发现女性CEO的图片较少，原因是《财富》 500强企业中只有5％的CEO是女性，这将导致搜索结果偏向男性CEO。这些搜索结果当然反映了现实，但是搜索算法是否应该反映这一现实是一个值得考虑的问题。

**2.3.2 群体偏见**

当统计数据时，如果目标群体与待统计群体不符时，会出现群体偏见，这种偏见产生于用户本身的差异分布[9]。例如，根据不同社交平台上用户人口统计数据，女性可能更偏向于使用Pinterest，Facebook，Instagram，而男性更偏向于Reddit或Twitter。

**2.3.3 自我选择偏见**

自我选择偏见是指不同用户在产生数据时存在的偏差，常见于统计调查。如在民意调查中，对政治更热情的民众更愿意完成投票。

**2.3.4 社会偏见**

其他人的行为或内容影响我们的判断时，就会产生社会偏见[6]。例如在对一个较差的商品进行评分时，受到其他高评分的影响，我们可能改变我们的原有想法，也对其评为高分。

**2.3.5 行为偏见**

行为偏差来自于不同平台、上下文或不同数据集的不同用户行为[9]。例如在不同平台上展示相同的符号表情可能会获得不同的理解。

# 人工智能偏见的解决方案

从技术角度出发，为了实现算法公平，研究者做出了许多解决人工智能中偏见的努力，一般来说，针对算法中偏差的公平性解决方案可以分为三类：

预处理。预处理技术尝试转换数据，使潜在的偏差被删除，如果允许算法修改训练数据，则可以使用进行预处理技术。

运行时处理。在模型训练或运行时，改进学习算法，引入新的技术手段以对抗偏见，如果允许修改模型或者训练过程，可以修改模型的运行机理，如修改损失函数。

后处理。后处理将模型视为黑盒，即只能进行输入并获取输出结果，并由此仅对输出结果进行处理，使其偏见不再被展示，如设计一个偏见检测模块，将所有带有偏见的结果拦截。

3.1 无偏的数据

为了减轻数据偏见的影响，相关研究者已提出了一些通用的方法，这些方法主张在使用数据时要有良好的做法，例如，使用数据表作为支持文档，阐述数据集创建方法、特征、动机以及偏见[10]。

除了通用技术，一些工作还针对具体类型的偏见提出解决方案。例如，将因果模型和因果图用于模型设计中，以检测数据中的歧视[11]。

3.2 公平的学习算法

为了解决算法本身的公平性问题，人们根据不同的模型提出了各种方法来满足一些已有的公平性定义，下面举例介绍：

公平的分类。由于分类是机器学习中的一项典型任务，并且广泛应用于可以与人类直接接触的不同领域，因此这类方法必须产生公平的结果，消除可能伤害某些人群的偏见，对此研究者提出了某些满足公平性定义的方法，例如分类公平性、机会均等、概率相等[12]。

公平的回归。对于回归问题，有作者提出POF准则帮助实现准确性和公平性间的合理权衡[13]。

# 总结

本文主要讨论了人工智能（AI）系统中的偏见问题，特别是这些偏见如何影响算法的公平性。文章指出了数据偏见和算法偏见是导致AI系统不公平的两个主要原因，并提出了相应的解决方法。

对于数据偏见，本文总结了测量偏差、省略变量偏差、表示偏差、聚合偏差和抽样偏差等问题，这些问题都是因为数据本身存在问题或者数据收集方式导致的。为了解决这些问题，需要更加科学和公正地收集数据，并尽可能地考虑所有相关变量。

对于算法偏见，本文总结了算法设计本身可能导致偏见，例如算法的选择、优化函数的选择、正则化的方法等都可能影响算法的公平性。为了解决这些问题，需要更加审慎地设计算法，并考虑其可能对用户行为产生的影响。

此外，本文还总结了用户交互偏见的问题，这是由于用户在与系统交互的过程中受到界面引导而产生的偏见。为了解决这个问题，需要优化交互界面，尽可能减少对用户选择的引导，让用户自主做出决策。

总的来说，要解决AI系统中的偏见问题，需要从数据收集、算法设计、用户交互等多个方面入手，全面提升AI系统的公平性和公正性。

# 参考文献

[1] Amitabha Mukerjee, Rita Biswas, Kalyanmoy Deb, and Amrit P Mathur. 2002. Multi–objective evolutionary algorithms for the risk–return trade–off in bank loan management. International Transactions in operational research 9, 5 (2002), 583–597.

[2] Miranda Bogen and Aaron Rieke. 2018. Help wanted: an examination of hiring algorithms, equity. Technical Report. and bias. Technical report, Upturn.

[3] Mehrabi N, Morstatter F, Saxena N, et al. A survey on bias and fairness in machine learning[J]. ACM computing surveys (CSUR), 2021, 54(6): 1-35.

[4] Harini Suresh and John V Guttag. 2019. A Framework for Understanding Unintended Consequences of Machine Learning. arXiv preprint arXiv:1901.10002 (2019).

[5] Kevin A Clarke. 2005. The phantom menace: Omitted variable bias in econometric research. Conflict management and peace science 22, 4 (2005), 341–352.

[6] Ricardo Baeza-Yates. 2018. Bias on the Web. Commun. ACM 61, 6 (May 2018), 54–61. https://doi.org/10.1145/ 3209581

[7] Azadeh Nematzadeh, Giovanni Luca Ciampaglia, Filippo Menczer, and Alessandro Flammini. 2017. How algorithmic popularity bias hinders or promotes quality. arXiv preprint arXiv:1707.00574 (2017).

[8] Batya Friedman and Helen Nissenbaum. 1996. Bias in Computer Systems. ACM Trans. Inf. Syst. 14, 3 (July 1996), 330–347. https://doi.org/10.1145/230538.230561

[9] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2016. Social data: Biases, methodological pitfalls, and ethical boundaries. (2016).

[10] Misha Benjamin, Paul Gagnon, Negar Rostamzadeh, Chris Pal, Yoshua Bengio, and Alex Shee. [n.d.]. TOWARDS STANDARDIZATION OF DATA LICENSES: THE MONTREAL DATA LICENSE. ([n. d.]).

[11] Lu Zhang, Yongkai Wu, and Xintao Wu. 2017. Achieving non-discrimination in data release. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 1335–1344.

[12] Berk Ustun, Yang Liu, and David Parkes. 2019. Fairness without Harm: Decoupled Classifiers with Preference Guarantees. In Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97), Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, Long Beach, California, USA, 6373–6382.http://proceedings.mlr.press/v97/ustun19a.html

[13] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, JamieMorgenstern, Seth Neel, and Aaron Roth. 2017. A Convex Framework for Fair Regression. arXiv:1706.02409[cs.LG]