

# 人工智能公平性

- 姓名：JY
- 学号：

## 谁来为AI的公平性负责？如何让AI变得更加公平，真正的服务于人？

对于AI的公平性，**算法工程师起主要责任**，其余起次要责任。确保AI的公平性是一个多方面需要共同努力的问题，包括政府组织、科研机构、AI企业，以及所有的AI使用者。

### AI不公平的主要原因

AI不公平的主要原因通常可以归结为以下三个方面：数据偏差、算法偏差和应用场景中的偏见。

- 数据偏差**：数据偏差是AI公平性问题中的一个重要环节，如果训练数据集的收集、处理或选择存在偏差，那么这些偏差将可能会被AI系统所学习并复制，在应用时产生不公的结论或决策。例如，美国刑事司法系统使用的一种风险评估AI工具，被发现在评估犯罪嫌疑人的再犯风险上存在种族歧视。具体来说，该工具对黑人的再犯风险评估高于实际，而对白人的再犯风险评估低于实际。深入研究发现，这是因为该工具的训练数据来自于历史的犯罪记录，而这些记录本身就存在种族歧视的问题。在另一示例中，一款面部识别AI系统在识别非洲和亚洲脸部特征时，发现的错误率远高于识别欧洲面部特征时。原因是这款AI系统对亚洲和非洲面部的训练数据量远小于欧洲面部，导致其在识别时缺乏足够的样本参考从而产生误判。这些例子说明：AI系统的学习方式取决于输入数据的特征。换句话说，如果数据存在不公平性，那么AI系统就会学习并复制这些不公平性。因此，我们需要密切关注数据的收集和处理过程，确保数据集在性别、种族、年龄等各个重要维度上都足够多样化，以此来避免AI系统在实际应用中出现不公平的行为。
- 算法偏差**：是指AI系统在处理信息和学习规律过程中由于算法设计的问题，导致的偏离真实或公正的结果。有时这种偏差可能是因为算法工程师在设计算法时的主观偏见，有时也可能是算法在处理复杂数据和学习模型时的固有局限性。无论何种原因，算法偏差都可能导致AI系统的输出结果存在不公平、歧视等问题。例如，一个AI系统被训练来对取得成功的CEO的特征进行建模，数据集中大多数CEO都是男性，那么该模型可能会倾向于将“男性”视为成为成功CEO的一个重要特征。这是一个典型的算法偏差，因为性别并不应该是决定一个人能否成为成功CEO的关键因素。然而，由于训练数据的不均衡，算法可能学习到的是：“男性”的CEO更常见，这就引入了性别偏见。同样，如果在设计推荐算法时，过分强调用户过去的行为，那么可能会导致算法陷入过度个性化的“过滤泡泡”，用户接收到的信息过于单一，不能全面了解世界，这也是算法偏差的一种。为了避免算法偏差，我们需要在设计算法时考虑公平性，例如，对于重要的决策（如招聘、贷款批准等），我们应该明确规定不能基于种族、性别等敏感属性进行决策。此外，也需要从源头上控制数据质量和多样性，避免训练数据的偏差导致算法学习到错误的模式。同时，我们也需要定期审核和验证AI系统的决策，确保其在实际操作中的表现是公平的。
- 应用场景中的偏见**：是指在实际应用AI系统时，由于使用环境、条件、或处理方式等因素引起的偏见。即使AI系统在设计阶段并未存在明显的算法偏差，但在特定的使用场景和条件下，仍有可能导致不公平或歧视性的结果。例如，在招聘过程中使用AI系统，假设这个AI系统训练得很公平，没有性别、种族或其他偏见。然而，如果它被用于筛选的职位长期以来都是由特定群体（比如白人、男性）主导，那么由于历史倾向和职位要求，AI系统可能更倾向于选择这个群体的候选人，即使他们并非最合适的候选人。这其实反映了应用环境和社会结构带来的偏见。又比如，AI系统在提供贷款批准建议时，由于贫困地区的居民可能无法提供较高的信用评分和正式的金融记录，但这并不意味着他们就不应获得贷款。如果AI系统严格依赖这些信用评分和金融记录，而忽略了社会和经济条件，结果可能会对这些社区的居民产生不公平的影响。因此，在实际应用AI系统时，不仅要关注数据和算法本身，还需要考虑到应用环境和条件可能带来的偏见。我们需要确保AI

系统的使用过程是透明的，有持续的审查和公众参与，公认的公平原则和伦理标准应该被纳入到整个使用流程中，以避免或减少不必要的偏见与歧视。

为了解决AI系统不公平的问题，我们需要更加公平、多样化的数据来源，更加全面、透明的算法，以及更公正的使用环境。更为重要的是，我们需要认识到AI系统并非绝对公正的决策者，而是反映人类决策的工具，我们需要持续地检查和纠正其可能出现的不公问题。

## 促进AI更加公平的方法

让AI变得更加公平，主要有以下几个步骤：

1. **数据公平性**：AI的工作依赖于数据。我们要确保AI的训练数据是多元化的，反映不同群体的经验和观点。例如，如果一个面部识别算法只在单一的种族或性别上进行训练，那么它可能无法正确识别其他未充分代表的群体。因此，确保多元化，无偏见的数据源是非常重要的。
2. **算法的透明度**：公开算法工作原理及其决策过程，使用者能够理解AI的输出结果是怎么产生的。这成就了算法的可解释性和可信赖性。
3. **可解释性和课堂化**：AI的结论应该是可以理解的，用户有权了解AI推荐或决策的原因。比如，一款贷款授信AI，如果拒绝了用户的申请，那么用户就有权了解被拒绝的原因。
4. **合规性和伦理**：遵循相关法律法规，尊重用户隐私权，公平对待用户。比如，AI在做出决策时，不能因用户的性别、种族、年龄等特征而有所偏见。

例如，Google在2018年发布了人工智能的公平性原则，这个原则要求所有的AI系统必须公平对待所有用户，不能因为用户的种族、性别等因素做出不公平的行为。另外一个例子是IBM，他们发布了一款公开的，用于检测AI偏见的工具包AI Fairness 360，帮助开发者发现并纠正各类AI算法偏见。

直面AI不公平性的问题，需要多方面的努力。在数据、算法、应用等全链条中加强公平性考虑、引入公平性设计和评估机制，是解决这一问题的关键。以下将从数据集、训练、评估等环节，详述一些具体的方法和例子。

### 1. 数据集公平性

数据是AI的基础，一个公平的数据集应保证多样性和代表性。对数据集的处理基于课堂上提到的一个基本共识：数据集大不代表数据合理公平。

- **平衡数据集样本多样性**：包含不同背景、种族、性别、年龄等多元化的样本。例如，面部识别技术的训练数据集应包含不同种族、年龄、性别、表情等多样性的人脸图片，确保系统能够公平、准确地识别不同群体的面部特征。标注图片中的敏感标签，在调取训练数据时，挑选出平衡的数据，保证多样性，比如对于男女，肤色、年龄这些敏感的数据做处理。
- **数据无偏见**：避免采集和使用包含社会偏见的数据。例如，在进行员工性能预测的AI系统中，应避免使用包含性别、种族等偏见信息的数据。
- **审查调整数据集**：使用例如REVISE这种半自动化的数据集公平性审查工具，来进行自动审查合成公平的替代数据：利用原始数据集和GAN这类生成工具，由算法工程师自己重新对数据进行一遍合理的清洗与生成，以此来解决原先数据集中的问题。
- **合成成对数据进行数据增强**：由于原先数据集中，一些敏感的数据采样过少，比如犯罪分析中少有白人的，那么可以利用过采样的方法，将数据复制多份，来促使整个数据集的公平合理

### 2. 算法设计公平性

算法是数据驱动的，我们需要确保在算法设计中排除偏见。

- **增强算法透明度**：公开算法的工作原理及其决策过程，可增强算法的可解释性和可信赖性。Google的Explainable AI就是一款可以提供特定模型预测洞察和全局特征重要性分析的工具，提升算法的透明度。
- **公平性优化算法**：引入公平性约束，降低对某些群体的不公平影响。比如，Google的TCAV(TensorFlow Concept Activation Vectors)工具可以帮助算法工程师了解机器学习模型如何理解某个概念，通过修改这些概念，减少算法对特定群体的不公平倾向。

### 3. 训练过程公平性

在训练阶段，我们应用公平的算法和无偏见的数据进行模型训练。

- **公平的优化目标增加公平性限制**：考虑在目标函数中引入公平性指标，如平等机会，平等预测等。例如，预测工作人员的绩效对于所有性别应该有相同的精度。在训练模型时，可以引入公平性限制项来确保模型不会因为隐含的偏见而产生不公平的结果。
- **多元化训练数据**：使用更广泛和多样化的训练数据，以代表尽可能多的用户群体和条件。举例来说，自动语音识别系统就应该使用包括所有地域、性别、年龄、方言的音频样本进行训练。通过引入对抗样本来迫使模型更好地学习数据分布和规律。对于公平性问题，可以利用对抗本来检测和纠正模型的隐含偏见。为了避免数据来源不均带来的偏见，可以尝试在多个数据集上进行训练，从而生成更具有代表性的数据，并提高模型的公平性。
- **领域独立训练**：在某些情况下，模型在特定的数据集上表现良好，但是在其他数据集上却表现不佳。这种现象称为领域依赖性，会导致模型的公平性受到威胁，那么便需要尝试利用领域独立的训练数据集，来提高模型对不同领域数据的泛化性能和公平性。
- **监督信息的引入**：在某些情况下，可以通过引入一些有监督的信息来纠正模型的不公平性。例如在图像分类问题中，可以对某些具有代表性的特征进行监督，以避免对某些特定群体的偏见，程序员可以在epoch的迭代中设置监督检查。

### 4. AI评估和监督公平性

在评估阶段，我们应确保AI系统在各种情况下都表现得公平。

- **多角度评估**：从多个角度进行公平性评估，包括预测结果的准确性、公平性、可解释性等，对结果进行多角度的解读。比如，MIT的“what-if”工具可以让研发者进行预测结果的多角度分析，
- **公平性审计**：定期通过公平性审计来评估AI系统的表现。例如，IBM的AI Fairness 360 工具提供了一组用于检测和缓解机器学习模型和数据中偏差的工具及库。
- **修改模型敏感属性**，调整预测结果，这里可以运用课堂上提到过的算法：后处理算法  
IGD（Iterative Gradient Descent）是一种用于优化分类器预测以减少决策偏见的方法。它在已有预测结果的基础上进行迭代和调整，以接近一个预设的公平性目标。通过优化特定的公平性度量准则，如平等机会或演算平均，IGD能够减少不同群体之间的预测结果差异，以消除潜在的偏见进行公平性纠正，进而增加AI模型的预测公平性。
- 在测试中，**调整相关系数**：可以通过调整分类器的阈值来改变模型的决策，从而达到公平性的效果。如果在测试中发现，其针对少数群体有很大的偏见，那么不妨对这个分类器进行调整阈值来改变模型的决策，从而达到公平性的效果。
- 进行**模型可解释性研究**：对模型进行可解释性研究，查看模型的关注点，以此来促使工程师了解模型，在公布模型和进一步修改模型时都更为有所依据。

总的来说，AI公平性的实现需要在多个环节多方面共同努力，包括选择无偏见的数据源，公平的算法设计，公正的使用环境等，并在每个环节都引入评估和反馈机制，通过持续的跟踪和调整，以保证AI系统的公平性。

## 政府和法律对AI公平的保证

政府和法律对AI公平的保证是建立在对AI技术深入了解和科学有效管理的基础上的。此外，他们也能通过立法、规范和政策引导，促进AI公平的发展。

### 1. 法律立法：

在AI公平性的问题上，政府可以通过制定相关的法律法规来对其进行规范和管控，以保证AI的发展不会导致不公平的情况。首先，政府可以制定涵盖各个方面的AI法律法规，包括但不限于个人隐私保护、数据使用及存储、算法透明性等方面。这样不仅可以为AI的公平性提供法律保障，也能对AI的使用和应用设定一个明确的权利和责任边界。以欧洲联盟为例，他们于2021年4月提出了AI法规草案，这是全球首个针对AI的全面性法规框架。其中，规定了AI系统如何分类，对于符合特定规定的AI系统，必须要满足监管的所有要求，其中就包括规定AI系统不能违反公正，必须遵守保护所有人类权利的原则，以

及避免对某些群体的不公正对待。除此之外，AI法规还要求AI系统的提供者需要提供一份详细的AI系统说明，公开AI系统的目的、功能、预期使用、效用等（对于在公共空间被部署用于生物识别的AI系统，还要提供额外的信息）。这样的规定对于提高AI系统的透明度，促进其公平性具有重要的作用。其次，为了保障法规的有效实施，政府可以设立专门的法规监管机构，负责审查AI系统，对于违反法律法规的行为进行监测和惩罚。例如，欧洲AI法规草案中提出，需要成立欧洲人工智能委员会，这个委员会将负责监管整个欧洲范围内的AI系统安全，并可以对违反规则的AI系统进行制裁。

## 2. 政策引导：

除立法外，政府也可以通过政策引导，推动AI的公平研发和应用。首先，政府可以通过设置相关的资金支持政策，激励公平的AI研发。这些资金支持可能包括政府直接资助的研发基金、税收优惠政策等。例如，政府可以为经过公平性审计的AI项目提供研发基金，以鼓励企业和研究机构更积极地从事公平性方面的AI开发工作。同时，政府也可以通过提供税收优惠政策，激励企业将资源投入到公平AI的研发中。其次，政府也可以通过采购政策，推动公平AI的应用。例如，政府在采购AI产品或服务时，可以将公平性作为一个重要的考核标准，优先考虑那些经过公平性审计的AI项目，以此来鼓励企业在AI产品或服务的研发和设计过程中，更多地考虑公平性问题。政府还可以通过组织相关的技术挑战赛和奖励机制，激发公众、学术界和工业界对公平AI的关注和研究。例如，政府可以设立专项奖励，如“AI公平性技术创新奖”，对在公平AI方面做出贡献的个人和团队予以表彰和奖励。此外，政府可以通过公共服务领域的开发和应用，提高公众对AI公平性的认识和期待，推动社会公共舆论对公平AI的支持。例如，政府可以利用AI技术改善公共服务，例如在医疗、教育、法律等领域的公平性问题，通过效果显著提高公众对AI公平性的认知。

## 3. 法规执行：

法规的制定仅仅是第一步，有效的执行和监督则是确保AI公平性的关键所在。首先，政府应当设立或指定专门的机构负责对AI公平性进行监督和评估。这个机构可能包括来自法律、科技和伦理等各个领域的专家学者，他们共同的任务是积极监督AI系统是否符合公平性要求，评估AI系统是否存在可能导致不公平的风险。例如，机构可能设立专门的AI公平性审查程序，对于新开发的AI系统，需要在上市前接受公平性审查，仅当通过审查才能进入市场。在审查过程中，机构可能会对AI系统的数据采集、处理、分析等各个环节进行全面的检查，确保其符合公平性原则。这个机构也需要对那些已经在市场中的AI系统进行持续的监督。例如，他们可能会在一段时间后，再次对AI系统进行公平性审查，以便找到和解决可能因为时间推移而产生的新的公平性问题。其次，对于那些违反公平性原则的AI系统，政府需要给予严厉的处罚。处罚可能包括罚款、下架商品、取消许可证等，甚至对于严重违反公平性原则的公司或个人，可以采取刑事处罚。另外，政府还需要配备充足的制度和资源，来确保公众有足够的途径来投诉和申诉可能存在的AI不公平的问题。例如，可以设立专门的AI公平投诉渠道，公众可以通过这个渠道向政府反映他们在使用AI产品或服务中所遇到的可能的不公平性问题。

## 4. 公共参与：

将公众参与到AI公平性问题的讨论中来，能增加其对AI公平性问题的关注，帮助政府理解和聆听不同群体的需求和想法，也能使得决策更为公正和合理。首先，政府可以通过不同的管道，邀请公众参与到AI公平性的讨论中。例如，可以设立线上论坛，开设专题讨论，吸引广大网民参与讨论；也可以设立咨询热线，接受公众对AI公平性问题的反馈；还可以通过社交媒体等方式发布关于AI公平性的信息，提高社会各界的关注度。其次，政府可以举办公开的讨论会或听证会，让公众、专家和企业有机会公开表达自己的观点和建议。这种面对面的交流方式，能让政府更直观地感受到公众的需求，也能让公众看到政府对于这个问题的态度和决心。此外，政府还可以通过社区问卷调查、公众投票等方式，获取公众对于AI公平性问题的意见。这些数据能帮助政府更好地了解社会上对AI公平性的态度，为政策制定和法规执行提供参考。在公众参与的过程中，应注意尊重每一个人的意见，尤其是那些常被忽视的弱势群体，他们的声音同样重要。公众参与的目的是让尽可能多的人有权参与决策，保证公平性问题的处理更为公正和全面。

总结，政府和法律在保证AI公平性方面发挥着重要作用。他们通过制定和执行相关的法规，引导和鼓励AI的公平发展，同时也通过公开透明的方式让公众参与到AI公平性的讨论中来，最终实现一个公平的AI Society。

