《人工智能安全》课程作业

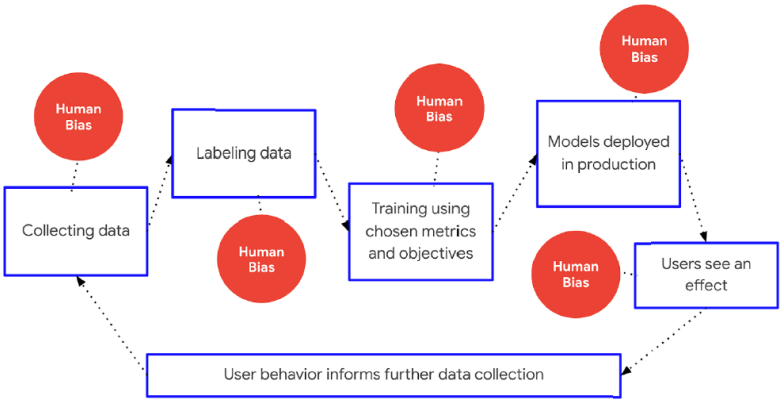
HW 4： 人工智能系统公平性

姓名： 周炜

学号： 3210103790

专业： 计算机科学与技术

邮箱： 3210103790@zju.edu.cn

> **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？**
> - 企业、算法工程师
> - 监管部门/法律
> - 用户
>
> 
>

谁为AI的公平性负责?

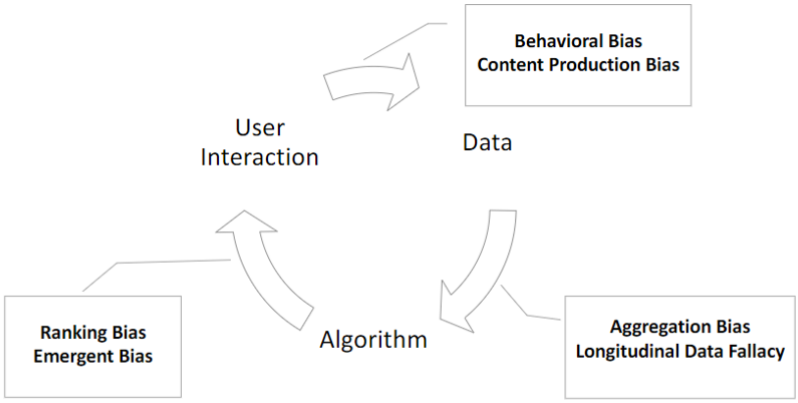
> 参考论文：A Survey on Bias and Fairness in Machine Learning

AI从模型到落地应用的**全过程中的每一个部门**都应该为自己所参与部分的AI的公平负责。AI的公平性责任需要由企业、社会、监管部门和用户等多方共同承担。只有通过各方的合作和努力，才能确保AI系统的设计和应用不会对人类社会产生不公平或歧视性的影响。 AI是一个复杂系统，其部件和服务可以是由多个市场参与者共同来提供的，因此，整个AI的治理是一个需要全产业共建共担、各司其职的多层治理模式

企业、算法工程师

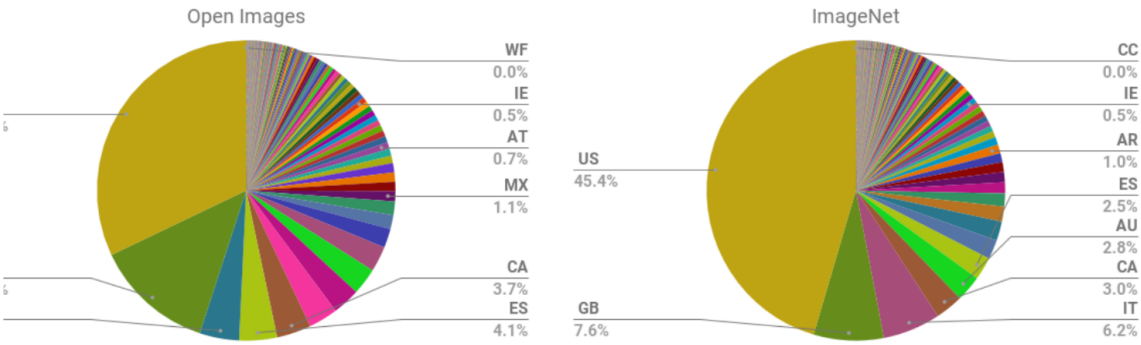
在AI的开发、部署和使用过程中，企业是主要的责任方之一。企业应该承担起确保AI系统公平性的责任。这包括在数据收集和处理过程中避免偏见和歧视，确保模型的**训练数据**集具有多样性和代表性，以及监控和纠正模型在实际应用中可能产生的公平性问题

研发者本身主观上并无意造成偏见，但往往有一些偏见是无意识产生的，从而导致将偏见引入AI系统的开发和设计中



在数据、算法和用户互动反馈回路中放置偏见定义的例子

比如，在使用`ImageNet`和`Open Images`这两个机器学习中广泛使用的数据集的时候就需要额外注意，因为其数据来源主要来自美国和英国



社会

社会应该对AI系统的设计和应用提供监督，确保它们不会对不同群体产生不公平或歧视性影响

政府和监管机构

政府机构和监管机构应制定和实施相关政策和法规，**设置红线**，并且**明确问责机制**，以确保AI系统的设计和使用符合公平原则。加强对相关企业的监督，要求透明度和问责制，对可能存在的歧视性和不公平行为进行调查和制裁

用户

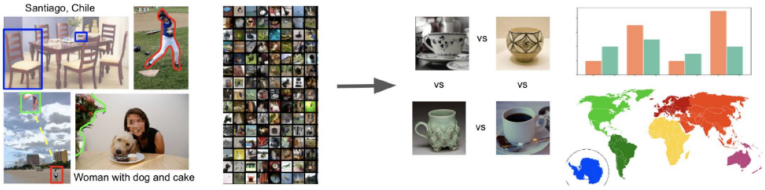
用户作为AI系统的使用者也应该对公平性负有责任。用户应该了解AI系统的工作原理和潜在偏见，并使用这些系统时保持警惕。用户应该提供反馈和投诉机制，以帮助揭示和解决AI系统中的公平性问题

如何让AI更公平?

数据方面

1. 使用更具有多样性和代表性的数据集：确保AI系统的培训数据集具有多样性和代表性，包括不同种族、性别、年龄、文化背景等的数据。避免数据集中的偏见和歧视，防止其在AI系统中被放大或复制
2. 预处理，使用预处理算法自校验训练数据，检测偏见和歧视，对数据使用模糊打标等方式进行公平化处理，比如使用**上课提到的REVISE半自动化的数据集公平性审查工具**

   - REVISE：半自动化的数据集公平性审查工具

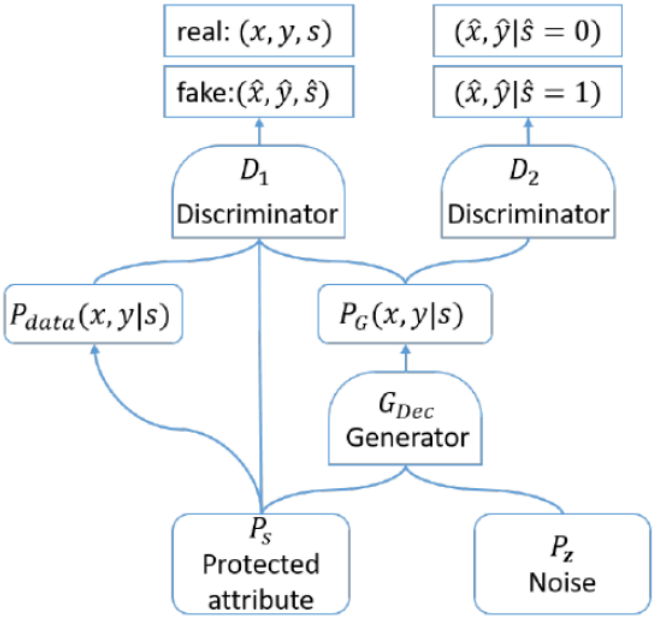


REVISE输入为含有标注的图片　　计算物体、性别、地域这三大类的公平性指标值

REVISE工具功能

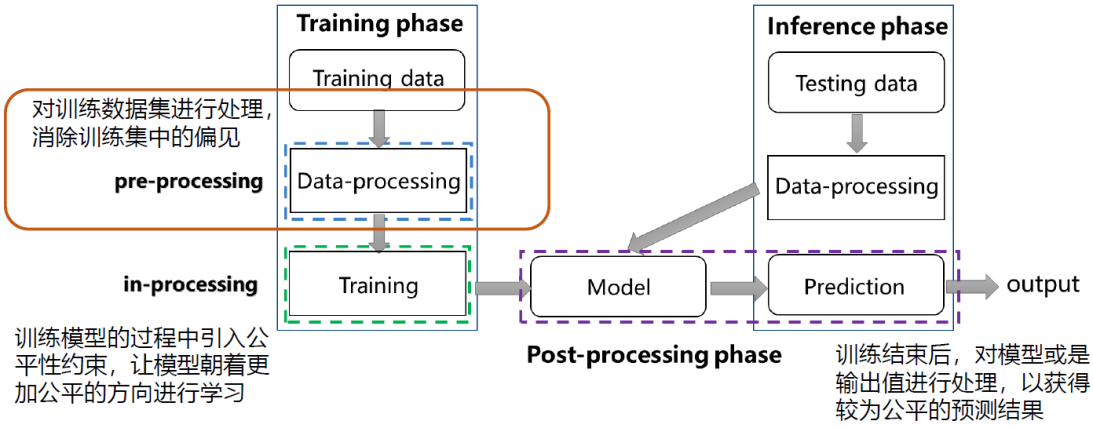
- ✓ 根据数据集的语义分割和标注、标签信息，使用内置的公平性指标进行计算，分析数据集中潜在偏见
- ✓ 提供可能的解决方案，辅助数据收集者进一步收集数据以消除偏见（公平最终需要依靠人的干预）
- ✓ 将偏见缓解融入到数据集创建的全流程中

3. 利用GAN ， 合成公平的替代训练集，在该训练集上训练模型



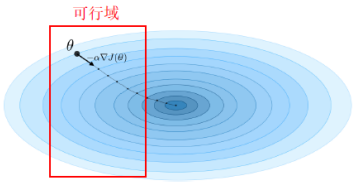
4. 合成数据集,进行成对增强,消除其中的偏见信息

算法方面

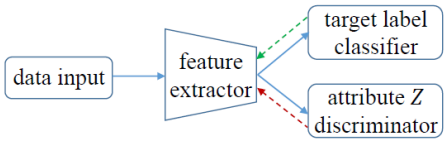


1. **算法改进**，将更改合并到目标函数中或施加约束，以下仅仅举1例(课堂上老师也提到过了**公平分类、对抗训练、独立领域训练、通过修改敏感属性以调整模型预测结果**，因此这里不再赘述)，更多可见于综述性质论文《A Survey on Bias and Fairness in Machine Learning》

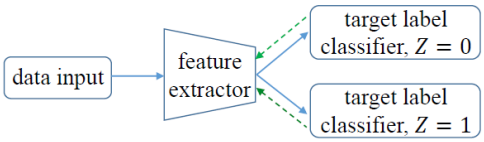
> 公平回归
>
> 参考论文《Towards Standardization of Data Licenses: The Montreal Data License》，将公平性指标(比如人际公平，群体公平，混合公平）引入机器学习的过程中去，**作为惩罚函数**
>
> 有两个群体S_1和S_2，其群体公平的惩罚函数就可以定义为：
>
> $$f(w,S)=\left(\frac{1}{n_1 n_2}\sum_{\substack{(x_i,y_i)\in S_1\\(x_j,y_j)\in S_2}} d(y_i,y_j)(w.x_i-w.x_j)\right)^2$$



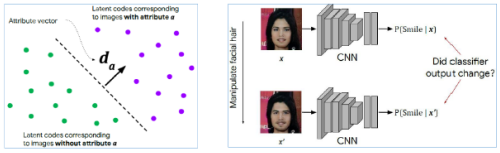
案例五：增加公平性限制项



案例六：对抗训练



案例七：领域独立训练



案例八：修改敏感数据

2. 训练后，通过访问模型训练过程中未涉及的保持集来进行**后期处理。**如果某种算法只能将学习的模型视为黑盒，而没有任何能力修改训练数据或学习算法，则只能使用后期处理，在该后期处理中，黑盒模型最初分配的标签则会根据在该阶段中的功能重新分配
3. 训练过程采取**分布式学习**（DLPs），通过在训练阶段允许受保护的敏感属性，但避免他们用来预测时间来同时满足训练差异和影响差异

其他

1. 增强透明度和解释性：AI系统应该具有透明度和解释性，使用户和相关利益相关者能够理解其工作原理和决策过程。这有助于发现和解决潜在的公平性问题，并建立用户对AI系统的信任
2. 加强审查和评估：对AI系统进行定期审查和评估，以识别和纠正潜在的公平性问题。这包括使用度量指标来评估系统对不同群体的影响，并及时修正和改进系统的设计和算法。这有助于避免单一视角和潜在的偏见，并确保AI系统的公平性和人类价值导向
3. 建立良好的反馈和问责机制：建立反馈和问责机制，让用户和相关利益相关者能够报告AI系统中的公平性问题和不公正行为，并采取适当的措施来纠正和改进AI系统
4. 加强监管和完善法规：政府和监管机构应制定和实施相关的监管和法规，以确保AI系统的设计和应用符合公平原则。这包括制定反歧视法律和隐私保护法规，加强对AI系统的监督和审查
5. 加强教育和帮助大众梳理意识：提高公众对AI公平性的意识和理解，通过教育和宣传活动，推动社会对AI公平性的重视，并促使各方更加关注和采取行动