人工智能公平性：

1. 问题与回答

问题：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？

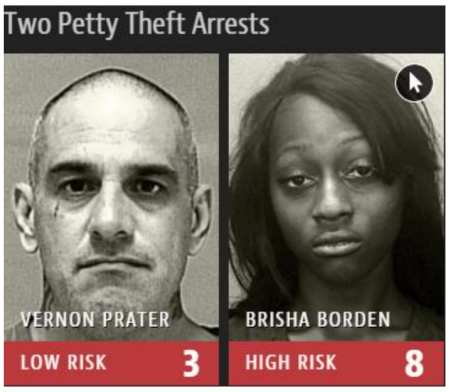
我认为这个答案主要是企业里的算法工程师起主要责任，其余起次要责任

2. 算法工程师

2.1 导致不公平的原因

AI的公平性是指与人类的道德价值和伦理准则相一致，避免歧视和不公平现象。在企业和算法工程师方面，他们是AI模型的**设计开发者**，无论是数据集还是算法设计上，都需要为此承担责任。

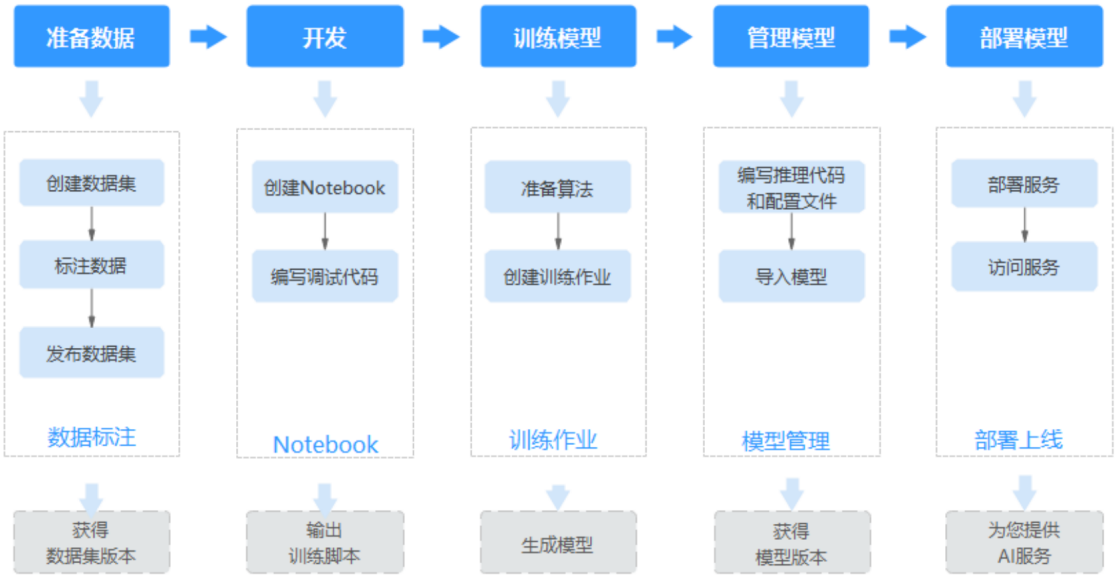
首先，**数据集**上，企业和算法工程师需要对AI的数据进行质量检查，以确保数据集的多样性和平衡。例如，如果数据集只有少数族裔或性别的数据，则AI会重复和放大这种先入为主的偏见，从而导致不公平的偏差。例如，在美国**犯罪情况**的判别中，对黑人和拉丁裔的标注很高，这会导致不公平的偏见；例如**机器翻译**中，又经常把“女性”与“护士“联系起来，造成NLP领域的训练数据固化问题。



此外，在**算法设计**上，企业和算法工程师需要注意算法本身是否存在某种意识形态上的偏见以确保AI决策的公平性。**算法开发人员**本身，无意之中便可能对于某一项数据赋予了**更高的权重**，由于开发者对参数的调整，这使得模型容易输出他所期望的成果，哪怕这个成果往往和事实所相违背。可以说，哪怕数据集标注没有问题，无形之中，这个训练者也在 **强迫**模型往他所期望的方向发展。

最后，在**模型评估**上，由于**深度学习**的黑盒性，谁也无法保证训练出的模型是否放大了某种联系，又或者是在原先的数据集中让他产生了什么误解。此时，需要算法工程师去对模型进行 **应用统计**，需要由他们去甄别训练出的模型是否合理正确。

2.2 促进公平的方法



从本图出发，叙述促进公平的角度

数据集方面，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，常常也有灯下黑的情况。**算法设计师应该对训练数据集进行处理，以消除训练集中的偏见。其方法就是课堂上所提到过的：

- **平衡数据集**：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
- **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

训练方面，可以通过如下的手段，从技术上来引导AI模型向更公平的方向发展：

- **增加公平性限制**：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。
- **对抗训练**：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，可以利用对抗样本来检测和纠正模型的隐含偏见。
- **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。
- **多源数据训练**：为了避免数据来源不均衡带来的偏见，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。
- **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。

在模型完成训练后，依旧有一些方法来促进其得到更好的公平性：

- 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，提升个体和组公平性

- 1. 利用验证集$\{x_j|z_j=0\}$，计算个体偏差$\{b_j|z_j=0\}$，若$b_j>\tau$，记录偏差$\beta_j=1$，否则$\beta_j=0$

$$b_j=|f(x_j,z=1)-f(x_j,z=0)|$$

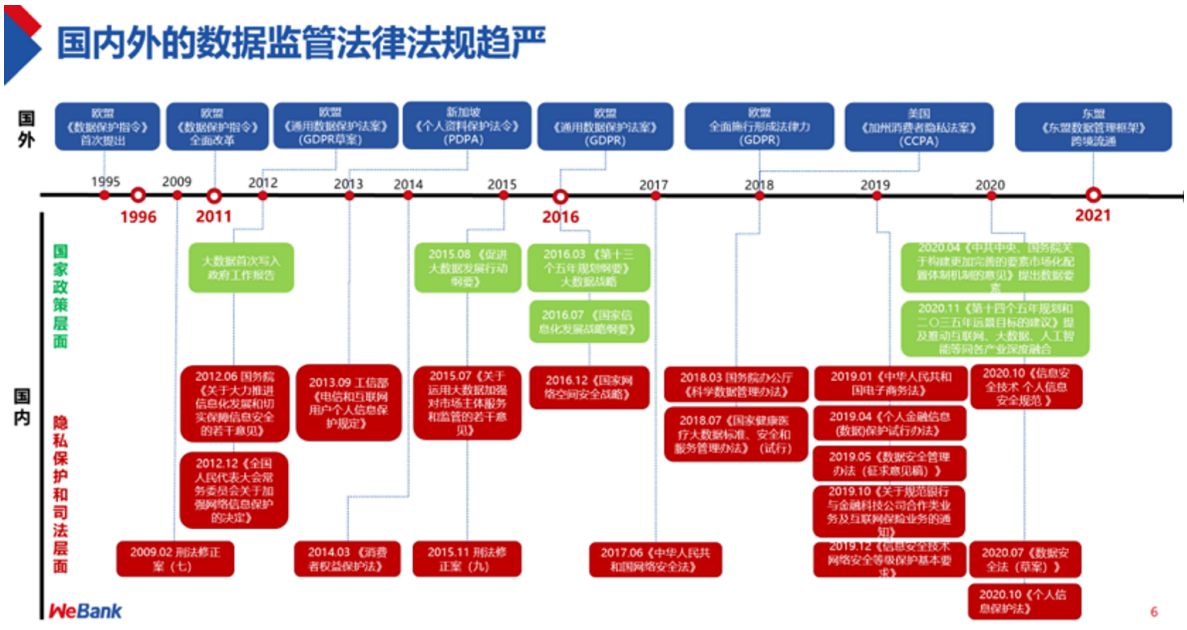
- 2. 创建辅助数据集$\{(x_j,\beta_j)|z_j=0\}$，训练偏差检测器g，输入为x_j，标签为β_j
- 3. 在f运行过程中，对于样本(x_k,z_k)，$\hat{y}_k=f(x_k,z_k)$
- 4. 若$g(x_k)==1$，检测出该样本会存在偏差，则令$\hat{y}'_k=f(x_k,1)$，否则$\hat{y}'_k$=$\hat{y}_k$
- 5. 最终返回的模型f预测结果为$\hat{y}'_k$

- 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。

3. 其它方面的对策

当然，以上我所说的所有都是从技术上分析的。因为再多的法律法规都拦不住真正邪恶的开发者，特别是在当今深度学习难以解释，门槛较高，行政人员难以把控的情况下的。但是，**政府部分**也依旧需要出台一定的政策去促使我上面说的这一切。

比如法律，法律可以要求企业和开发者遵守相关规范，例如美国人工智能技术发展法案（AI in Government Act）就明确要求联邦政府需要建立一个专门机构来监管并推进人工智能的发展。此外，还可以通过制定更多更严格的法规标准、增加行政处罚等手段来建立良好的法制环境，从而确保AI体系的公平性。



比如对违规行为的惩罚，对于那些存在公平性问题的AI应用，监管部门和法律也应该采取相应的措施进行惩罚，以保护消费者的权益和促进AI技术的公平性。例如，可以对那些涉嫌歧视性行为的AI公司进行罚款或停业整顿等处罚。

比如提高宣传，AI公平性是一个基于价值观和伦理的问题，需要引起更广泛的社会关注和认识。监管部门通过媒体宣传、公共教育、学术研究等途径，提高公众对AI公平性的认识，促进社会对AI技术的发展和应用的理解和支持。

而至于个人，我认为其作用相对于较小，但大体上，用户要重视自己的权力，抵制人工智能技术中的任何不公正行为。例如对于某些消费者应该享有的权益或服务有意忽视或歧视某个群体时，用户可以拒绝或寻求相应的法律救济。此外，用于也要有ai能力与风险并存的认知。在越来越智能化的今天，我们依旧不能对ai过度依赖，要拥有自我的是非判别能力，能够从用户端分辨出ai的内在偏见并勇于提出异议。

4. 总结心得

AI安全公平性是AI领域中非常重要的问题，它关乎到个体隐私、社会公平和道德伦理等多方面的问题。作为未来AI领域的从业者，我们应该培养以下三个方面的责任：

1. 技术责任：继续深入了解AI技术的安全与公平性问题，并在算法设计、开发和测试的每个阶段，采取相应的安全和公平性策略，以确保AI系统的正确性、可靠性和公正性。
2. 伦理责任：随着AI技术的快速发展，涉及到伦理和道德问题的情况越来越多。我们要遵守相关的职业道德规范，保证所开发的AI系统不会造成不良后果。
3. 技术透明：应该坚持透明度原则，记录所有的算法训练数据和决策过程，以保证责任的可追溯性。

我希望在未来，我们培养出一份对社会、对用户、对道德伦理的敬畏之心，在尽可能征得用户同意的前提下，确保AI技术的安全公平性，同时为AI技术在人类社会的应用贡献自己的力量。