AI ETHICS

主讲人：路建飞、胡旭伦
张逸然、李为民

AI和大数据分析的广泛应用

- 谷歌翻译
- 聊天机器人
- 自动化股票交易
- 医疗和手术机器人

道德和隐私等问题

- 隐私和数据保护
- 算法的不公平问题
- 算法的不透明问题

一些解决方法

- 对个人数据的保护，例如欧洲通用数据保护条例GDPR提出的一系列措施
- 标准化，例如ACM等机构制定了一些道德准则
- 负责任的研究和创新（RRI），在研究和创新的早期阶段让利益相关者加入进来，确保研究过程更加开放和透明，并要求研究人员处理相关道德和隐私问题

GDPR—Right to explanation

- 商业和政府部门越来越多的应用机器学习算法做出会影响我们日常生活的决策，而机器学习本身对于普通人来说就相当于一个可以输入输出的黑箱。随着对黑箱机器学习系统的歧视的担忧上升，因此在GDPR中制定了一种用户“解释权”的相关规定。

- 解释权：要求每个制定决策的算法必须能够证明决策的正确性。

GDPR—Right to explanation

- 用户角度：
- 1.GDPR中第15、22条仅针对个人的数据，那么除此之外的数据，如交通数据，在决策中作为重要要素，如何审查；法国数据隐私保护引入了权重概念，公开算法系统中的各因素权重避免了这一问题；
- 2.如何评估一个算法所造成的影响，收集对特定的用户的影响困难，对影响本身评估困难，这种困难直接转嫁到用户身上；
- 开发角度：
- 1.引入DPIA数据影响评估概念，在开发过程中涉及“对个人的评价或评分”、“具有法律或类似意义的自动化决策”等10种算法功能开发的情形需要进行DPIA的评估，DPIA与GDPR的结合会帮助开发者合规化。
- 2.自愿认证：算法系统的处理模块和控制模块可以自愿认证其符合规定，签发证书和信托印章。问题是签发机构以认证为生存的手段，以此来收取会费，因此当会员违反规定时，如果处罚有会员离开风险，如果不处罚又会降低认证可信度。
- 第三方机构：
- 1.目的，减轻用户所背负的审查自身数据是否被利用的负担；
- 2.GDPR81： GDPR的第81条规定，当事人可以委托第三方机关提出申诉，刑事司法权力，并代表其接受赔偿；GDPR10:允许第三方机构在不受数据主体的授权下，对数据控制器进行控诉；
- 3.民间机构可以成为监督者，但法院法律系统担心敏感数据泄露不愿公开算法系统

“解释权”or正确？便利？

ML与GDPR合规化

- 1.在DevOps管道（开发、运维两个环节）中利用机器学习发现个人身份信息：
- 2.BigID利用机器学习持续跟踪数据中心或云中跨生产环境和开发环境的个人身份信息变化情况。
- 3.BigID的BigOps使用机器学习跨所有数据存储库发现、背景化和编目个人身份信息。它接入开源DevOps环境（如Jenkins）跨整个开发生命周期自动监控个人身份信息的变更。它使用机器学习将其数据与可疑数据库进行对比，以快速确定哪里存在需要及时通知的违规情况。

在研究中实施RRI

- 欧盟的人脑工程（HBP）有一个子项目专门用于实施RRI计划以应对道德和社会问题，它要求
- HBP要对其工作产生的后果负责
- 开放公众参与，共同讨论HBP和公众相关的问题
- 对道德问题进行处理。
- 除此之外，还必须遵守欧洲关于研究对象知情同意的规定。

RRI的优点

- 建立一种责任文化，让研究人员愿意为其研究和创新工作的过程和结果承担责任。
- 通过将RRI整合到研究过程中，研究人员开始主动反思道德和社会问题。
- RRI提供了一种让我们主动地去塑造AI方法，而不是等待它来塑造我们。

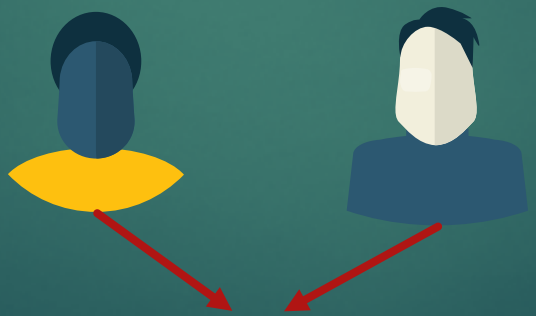
决策算法的公平性

决策算法会对人类造成伤害

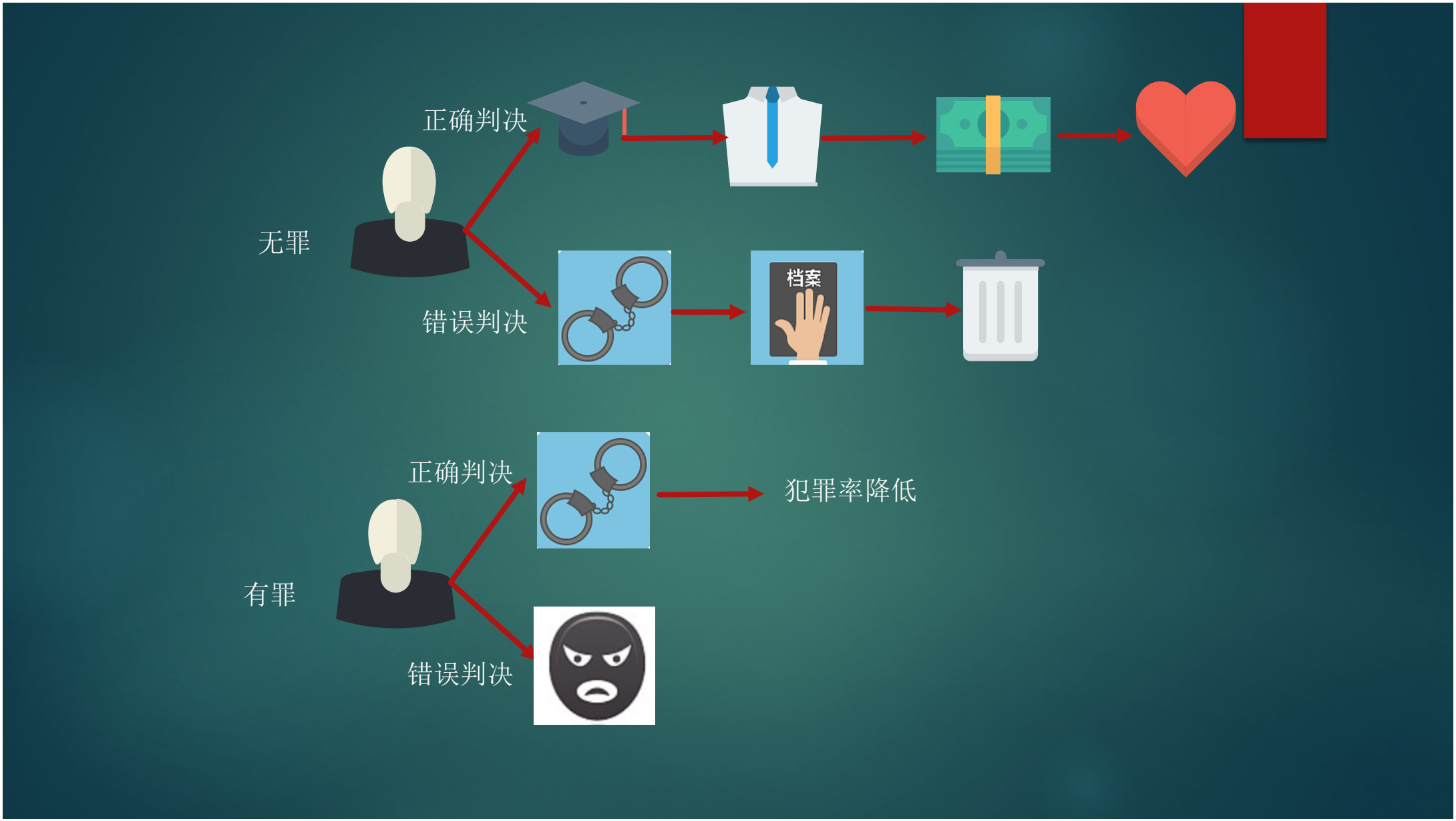
例子：

Northpointe Correctional Offender Management Profiling for Alternative Sanctions（COMPAS）

针对替代性制裁的诺森特惩教罪犯管理概况



种族歧视？


正确判决
无罪
错误判决
档案
正确判决
犯罪率降低
有罪
错误判决

决策算法为什么会产生不公平性、不公平体现在哪里

是因为算法有种族歧视？

是因为算法没有保护数据提供者隐私？

是因为算法不如人类判决？

是因为算法不够好？

生命历程分析

潜在因果框架

因果推理方法

反事实模型

评估分析算法

反事实场景1：“如果决策不是基于种族”

反事实场景与真实场景之间差异极小

预测结果不准确

反事实场景2：“如果参与者使用差别隐私保护”

没有提供期望的保护

其他因素造成的伤害要比隐私相关的伤害更为严重

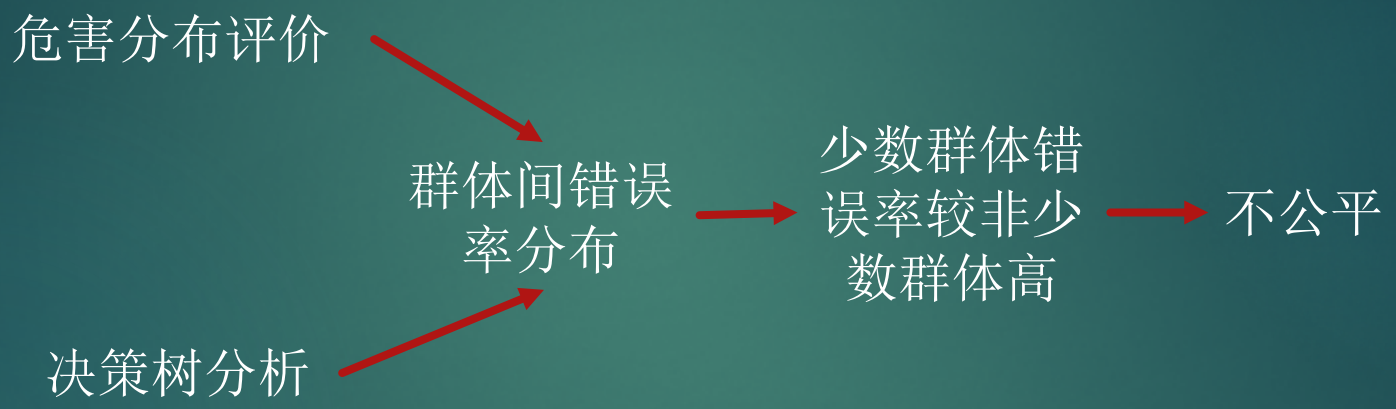
反事实场景3：“如果算法被人类判决取代”

人类决策远差于算法决策

反事实场景4：“如果使用更好的算法”

对于被修正个体来说，能减少危害
伤害的跨群体分布会不公平

不公平性究竟体现何处、原因是什么



原因

- 精度在群体间不一定平衡
- 错误率在群体间不一定平均分布
- 即使错误率能平衡，也不能保证公平结果

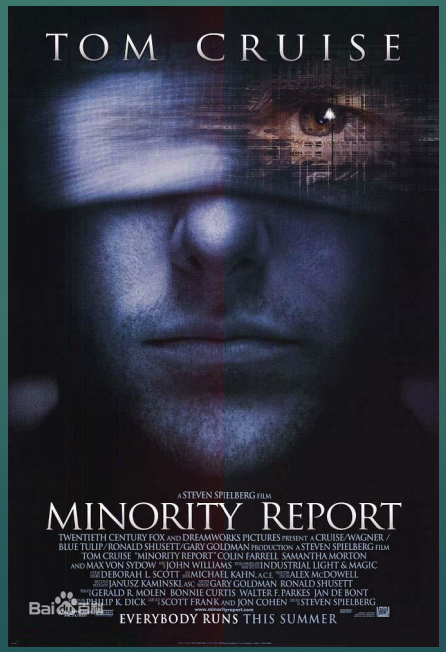
解决算法不公平——算法管理评估层面

- 所设计的算法应便于评估跨群体不公平性
- 反事实分析法预测算法决策产生的结果
- 反事实分析法解释算法的决策

算法什么时候是透明的？

少数派报告-史蒂文·斯皮尔伯格
电影讲述了2054年的华盛顿特区，谋杀已经消失了。未来是可以预知的，而罪犯在实施犯罪前就已受到了惩罚。司法部内的专职精英们----预防犯罪小组负责破译所有犯罪的证据----从间接的意象到时间、地点和其它的细节，这些证据都由“预测人”负责解析。他们是三个超自然的人，在预测谋杀想象方面还从未失过手。

假设现在他们预测你将要实施犯罪，你会怎么想？怎么做？



普遍认为的透明度的方法

20年来学者们认为，自动化处理需要更高的透明度，但是这种透明度应采取何种形式，却还远不清楚。数据挖掘和预测分析的兴起使得透明度问题变得更加迫切。因为决策往往脱离人类的直接控制。在这种情况下，人工控制仅包含在预测分析算法中内置的设计决策中，以及可能存在的任何决策后评审程序中。

1.公开源代码（将所有的代码都开源，让每个人都可以看到。）
2.分析但是不公开源代码（系统确定的输入与输出，属性，算法结构，可以完全解释其功能）
3.信息规范（规范，协调规范，信息协调规范）

预测分析的两个方面

数据去文本化

一个例子，假设Sally违约了50,000美元的信用卡债务。她承担了为她八岁的女儿支付救生治疗费用的债务，虽然她一直在付钱，但她负担不起最低金额。当信用卡公司开始收款时，她宣布破产。罗杰也违约了5万美元的信用卡债务，这是他赌博引起的，罗杰也宣布破产。 他们各自的信用报告记录了破产事件，但没有说明导致他们破产的不同背景。信用报告使信息脱离背景。脱离语境的破产报告意味着他们都难以获得信贷，尽管莎莉实际上是一个良好的信用者，而罗杰不是。

预测分析使数据脱离背景

abortionfacts.com（堕胎网站）

例如，搜索对象可能是一名寻求堕胎的孕妇，一名支持堕胎的激进分子，一名反堕胎的激进分子，或者一名学术研究人员。

您可以通过添加更多的数据(例如，搜索者是男性)来消除其中的一些可能性，但无论多少数据都无法提供充分的解释。

你理解并解释为什么人们通过构建将他们的价值观、目的和意图以及他们发生在其中的背景整合成一个有意义的模式的叙事来思考和行动。

增强预测能力

因为尽管错误率很高，但预测模型通常优于人类在没有他们帮助的情况下做出的预测。预测系统往往有着更高的准确率，速度和不会受外在因素影响（歧视，心情等）

虽然歧视数据令人讨厌，但是它仍对预测有着很大的帮助。顺便说一下，并非所有令人反感的歧视都是非法的。例如，许多人发现至少有一些完全合法的价格歧视案例（对不同的人收取同一产品或服务的不同价格）。

普遍认为的透明度的方法

1.公开源代码（将所有的代码都开源，让每个人都可以看到。用户往往不看也不懂，专业人士有时也可能不懂，并且算法可能脱离经济和文化背景）

2.分析但是不公开源代码（系统确定的输入与输出，属性，算法结构，可以完全解释其功能，但是保证其公平性是一个问题，歧视数据往往能提升算法的准确度，没有法律要求算法必须满足反歧视法，也是在确定该系统是否能在相关成本和效益之间产生可接受的权衡方面效果有限，仍然需要算法运行时的经济和文化背景。）

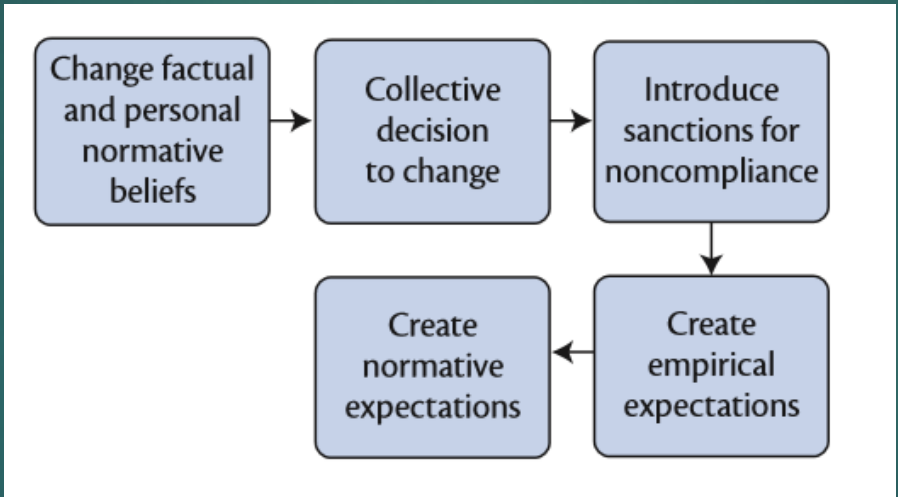
3.信息规范（规范，协调规范，信息协调规范，管理决策过程的问责机制和法律标准没有跟上技术的步伐，没有足够规范的信息规范）

信息规范，为了确保消费者能够轻松确定与信息隐私相关的预测系统的风险和收益。信息规范无处不在，它们限定了在给定的上下文中，允许、期望、甚至要求被揭示的关于不同个体的信息的类型或性质。信息规范限制了信息的收集、使用和分配，其方式取决于规范各方互动的社会角色。规范在信息处理的成本和收益之间取得平衡，允许一些但不是所有的信息处理。适当的信息规范可以使消费者轻松确定与其所处的预测系统相关的风险和收益，因此是消费者透明度的重要来源。

信息规范是消费者透明度的源泉主要讲了，规范是群体中的一种行为规则，只有集体整合才能确保和谐。目前缺乏规范，预测分析的许多(如果不是大多数的话)用途不受相关规范的制约。

规范
协调规范
信息协调规范
适当的信息规范可以使消费者轻松确定与其所处的预测系统相关的风险和收益，因此是消费者透明度的重要来源
实现这一目标需要有适当选择的信息流。通过遵守以下信息协调规范，各成员共同确保所需的选择性信息流：**只披露适当的信息**信息规范限制了信息的收集、使用和分配，其方式取决于规范各方互动的社会角色。只有集体整合才能确保和谐，所以除非有足够的成员顺从，否则任何一个成员的顺从都没有什么意义。

一个例子
汤姆一家要办一个家庭聚会，每个人都知道不应该说，不开心的话（规范）
每个人都应该不说不开心的话，且认为其他人也不会说（协调规范）
现在汤姆在吃饭前明确说了，大家不要说不开心的话，而且要求大家要做到（信息协调规范）



最后作者建议可以使用法律来促进信息规范

总结

从AI自身来看，它只是一种工具和技术。在社会的方方面面，AI都起到了推动行业发展和便利生活的作用，但是同时AI的广泛使用也会产生一系列前所未有的安全问题，危害用户与社会安全，就像是一把双刃剑。

这不仅需要我们从技术层面考虑安全问题，安全以人为中心，更要从法律、规约、道德等方面，更全面的考虑AI带来的安全问题。

- 谢谢观看