



北京大学前沿计算研究中心  
Center on Frontiers of Computing Studies, Peking University

---

# A Snapshot of the Frontiers of Fairness in Machine Learning

---

Author: Chouldechova Alexandra and Aaron Roth

Lecturer: 邢捷 (@jxing0831)

Aug. 24<sup>th</sup>, 2022

# Authors

---

- Alexandra Chouldechova



<https://www.andrew.cmu.edu/~achoulde>

- Aaron Roth



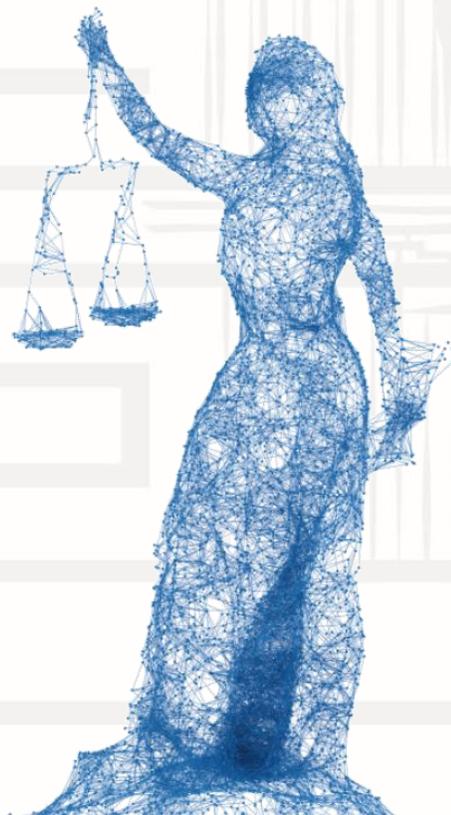
<https://www.cis.upenn.edu/~aaroht/>

# Introduction

---

过去十年机器学习应用的多样性在增加，ML也从早期广告和垃圾的过滤转变为筛选贷款申请人、警力部署、保释和假释通知等一些新的领域。

然而采用数据驱动方式可能会引入“歧视性”概念，此类决策可能在无意中编码了现有ML中的偏见，从而导致引入新的潜在性偏见，因此保证ML决策的公平性显得尤为重要。



# Introduction

---

在过去的几年中，学术界对于ML公平性的讨论逐渐热门起来。尽管如此，我们对机器学习与公平性相关的问题也只是停留在起步阶段。

- What should fairness mean?
- What are the causes that introduce unfairness in machine learning?
- How best should we modify our algorithms to avoid unfairness?
- What are the corresponding tradeoffs with which we must grapple?



*A group of industry, academic, and government experts convene in Philadelphia to explore the roots of algorithmic bias.*

—2018.03



# *What we know*

---

## *Causes of Unfairness*

我们定义的“公平”可能存在偏差，这导致了现有机器学习的“扭曲”行为：

- Bias encoded in data
- Minimizing average error fits majority populations
- The need to explore



# What we know

---

## *Definitions of fairness*

- 机器学习公平性的绝大多数工作都集中在批次分类的任务上
  - 统计定义
    - 大多数、简单的
  - 个人定义 (Dwork\_相似的个体应该受到相似对待)
    - 约束、特定的

以上公平概念的局限性，是否存在“两全其美”的方法？



# Questions at the Research Frontier

---

## *Between Statistical and Individual Fairness*

- 如何做到同样为个人提供保证而无需对知识作出假设即可实现约束？

Kearns和Hèbert-Johnson尝试要求公平的统计定义适用范围扩大到函数定义，该方法显著减轻了统计定义下的一些弱点。

另一项工作削弱了Dwork的假设，算法可以根据未知量返回估计量，并确保公平的统计概念。

以上方法的公平性定义对于某些指标是完美定义的，能否推广到不与任何指标统一的环境？



# Questions at the Research Frontier

---

## *Data Evolution and Dynamics of Fairness*

- 关于算法公平性的研究绝大多数集中在一次性分类任务中，我们需要了解更复杂的系统中公平的动态性。

eg: 广告投标人各自的决定是公平的，那么广告的分配何时公平，何时不公平？

- 研究发现多个公平的组合根本不满足任何公平性约束，同样，一个公平系统的组成部分独立看起来可能也是不公平的。
- 因此找到满意的公平定义和表现良好的框架结构是重要的。



# *Questions at the Research Frontier*

---

## *Data Evolution and Dynamics of Fairness*

- 了解算法如何动态影响环境和人类的动机

eg: 降低大学录取的门槛，随着时间的推移，下一代中会有更大比例孩子生活在父母受过高等教育的家庭。

- 决策通常分布在大量目标不同且难以协调的参与者之中。我们无法直接控制决策过程，而是考虑如何公平地行事。



# *Questions at the Research Frontier*

---

## *Modeling and Correcting Bias in the Data*

- 数据本身是对某些群体不利的社会和历史过程的产物，现有的ML方法可能会强化数据中的偏见。我们的任务则是了解数据中偏见的产生过程，以及如何纠正偏见。
- 使用机器学习训练模型时，模型结果会影响决策过程，进而影响未来训练所得结果。
- 纠正数据偏差通常需要了如何产生偏差，或者判断数据在“无偏差”世界中满足的属性。



# *Questions at the Research Frontier*

---

## *Fair Representations*

- 公平表示学习是一种数据去偏过程，在尽可能多保留与任务相关信息时删除敏感或受保护的信息，引入去偏数据集后，不会产生其他风险。
- 公平表示学习促进公平的好处主要取决于转换后和受保护的特征之间的关联度。



# *Questions at the Research Frontier*

---

## *Beyond Classification*

- 尽管机器学习公平性的工作大都集中在批次分类上，但这只是机器学习使用的一个方面，大部分机器学习（在线学习与强化学习等）都关注动态设置，以此捕获公平性问题。
- 强盗环境中的学习理论需要在探索与利用之间权衡，算法并不总是作出最优决策，有时需要采用次优策略以收集更多数据，这也是不公平的根源。
- 更多关于机器学习公平性的工作——排名、选择、个性化、强盗学习、人机混合预测系统、强化学习等。





---

*Thanks for watching!*

---

