A History of Algorithmic Fairness

Mingle Presentation

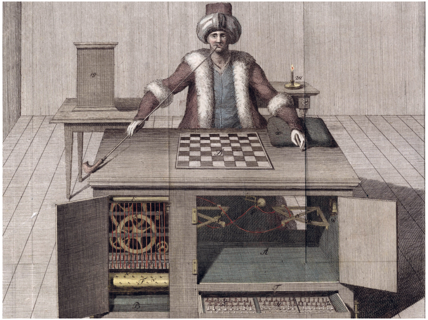
Dom Owens

University of Bristol

September 2020

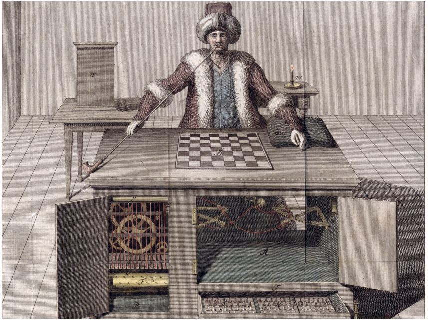
Contents

Introduction: Algorithmic Fairness



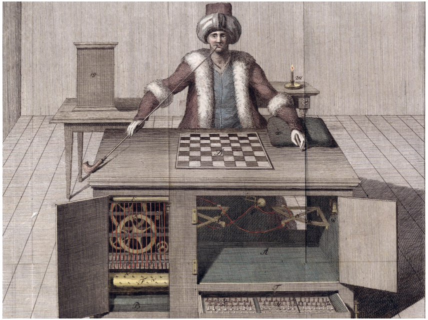
- We now make many decisions aided by **statistical** (**AI**, **ML**, ...) **algorithms**

Introduction: Algorithmic Fairness



- We now make many decisions aided by **statistical** (**AI**, **ML**, ...) **algorithms**
- Some of these affect humans: Parole decisions, hiring, credit scores...
- Algorithms *should* make decisions better than humans: More data, much faster, **less biased?**

Introduction: Algorithmic Fairness



- We now make many decisions aided by **statistical** (**AI**, **ML**, ...) **algorithms**
- Some of these affect humans: Parole decisions, hiring, credit scores...
- Algorithms *should* make decisions better than humans: More data, much faster, **less biased?**
- Often, algorithms show same biases as humans - why?

Definitions and Measures: What are we dealing with?

Disparate Treatment: **intentionally** treating an individual differently based on his/her membership in a protected class

Definitions and Measures: What are we dealing with?

Disparate Treatment: **intentionally** treating an individual differently based on his/her membership in a protected class

Disparate Impact: negatively affecting members of a protected class more than others even if by a **seemingly neutral** policy

Definitions and Measures: What are we dealing with?

Disparate Treatment: **intentionally** treating an individual differently based on his/her membership in a protected class

Disparate Impact: negatively affecting members of a protected class more than others even if by a **seemingly neutral** policy

For some yes/no decision ($Y \in \{0, 1\}$), with class membership $S \in \{0, 1\}$, consider ratio

$$R = \frac{P(\hat{Y} = 1|S = 0)}{P(\hat{Y} = 1|S = 1)}$$

Definitions and Measures: What are we dealing with?

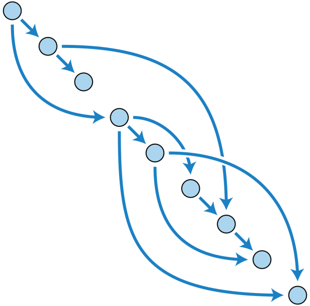
Disparate Treatment: **intentionally** treating an individual differently based on his/her membership in a protected class
Disparate Impact: negatively affecting members of a protected class more than others even if by a **seemingly neutral** policy

For some yes/no decision ($Y \in \{0, 1\}$), with class membership $S \in \{0, 1\}$, consider ratio

$$R = \frac{P(\hat{Y} = 1 | S = 0)}{P(\hat{Y} = 1 | S = 1)}$$

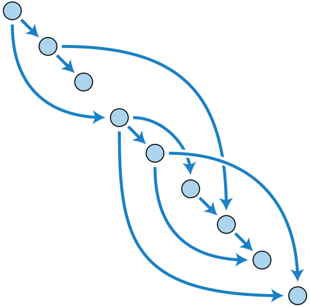
If $R \leq 1 - \epsilon$, we have evidence of discrimination ($\epsilon = 0.2$ for "80% rule")

Causes: Where does this come from?



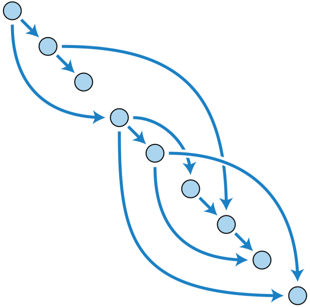
- **"Optimality"** - algorithm aims for accuracy for majority groups

Causes: Where does this come from?



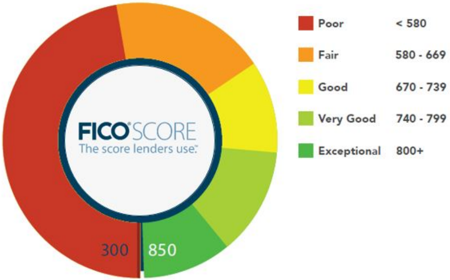
- **"Optimality"** - algorithm aims for accuracy for majority groups
- **Reconstructing** protected characteristics from correlated features

Causes: Where does this come from?



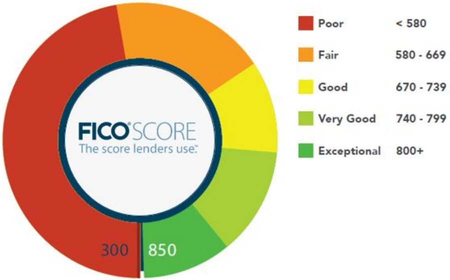
- **"Optimality"** - algorithm aims for accuracy for majority groups
- **Reconstructing** protected characteristics from correlated features
- **Biases in data set** (bad measurements, historical decisions...)

1989: FICO Credit Scoring



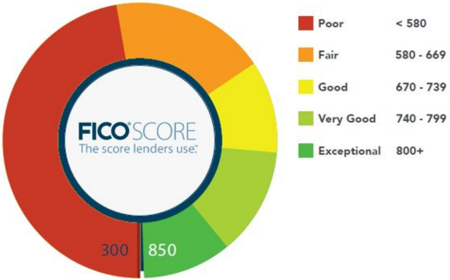
- Credit scoring: should we approve a loan to this person?

1989: FICO Credit Scoring



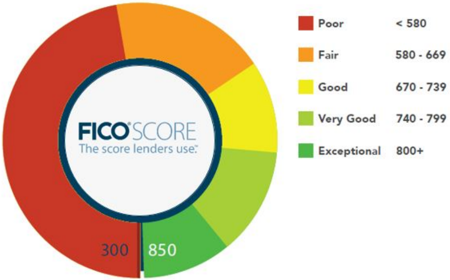
- Credit scoring: should we approve a loan to this person?
- Uses bank data, "ignores" **protected characteristics**

1989: FICO Credit Scoring



- Credit scoring: should we approve a loan to this person?
- Uses bank data, "ignores" **protected characteristics**
- Evidence shows algorithm still discriminates along the same lines

1989: FICO Credit Scoring



- Credit scoring: should we approve a loan to this person?
- Uses bank data, "ignores" **protected characteristics**
- Evidence shows algorithm still discriminates along the same lines
- **Reconstructs** from address, university, web activity...

2014-2018: Technical Hiring at Amazon

- Amazon filtered applications for technical jobs

2014-2018: Technical Hiring at Amazon

- Amazon filtered applications for technical jobs
- Algorithm looked for **similarities to previous hires**

2014-2018: Technical Hiring at Amazon

- Amazon filtered applications for technical jobs
- Algorithm looked for **similarities to previous hires**
- Learned e.g. gender through choice of language

2020: A-level Grades

- A-level exams cancelled

2020: A-level Grades

- A-level exams cancelled
- Ofqual proposed grades from algorithm fed with historical data on **school achievement**

2020: A-level Grades

- A-level exams cancelled
- Ofqual proposed grades from algorithm fed with historical data on **school achievement**
- Were schools on a **fair footing?** Resources, class sizes, student backgrounds...

2020: A-level Grades

- A-level exams cancelled
- Ofqual proposed grades from algorithm fed with historical data on **school achievement**
- Were schools on a **fair footing?** Resources, class sizes, student backgrounds...
- Small classes not re-evaluated by algorithm

Conclusion

- Used unwisely, algorithms can discriminate

Conclusion

- Used unwisely, algorithms can discriminate
- This can have potentially very unfair outcomes

Conclusion

- Used unwisely, algorithms can discriminate
- This can have potentially very unfair outcomes
- **Don't despair!** Ongoing research, increased awareness