Machine Learning and Incentives

Daniel L. Chen
Toulouse School of Economics

Henrik Sigstad
University of Oslo

SIOE
Jun 24, 2022

Machine learning improves decisions ...

Hiring

Lending


Credit Score
Overview
Personal Data
My Score
Your Credit Score Is
811
Excellent
330
590
930
Click Here For More Details
Legal Information
Terms Of Use
Copyright

Justice

... but also causes problems

Algorithmic Bias

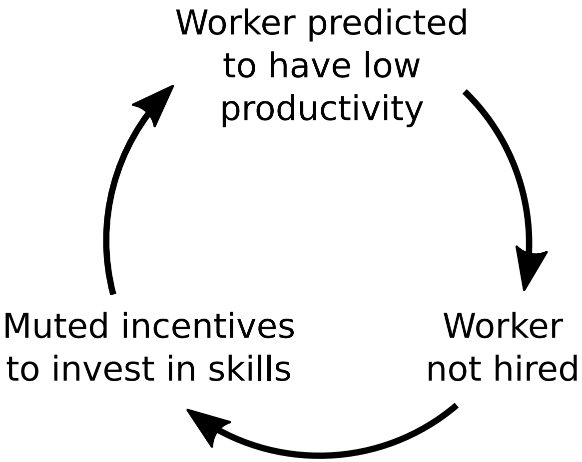
Mehrabi et al (2022), Kleinberg et al (2018), ...

Machine Learning and Incentives

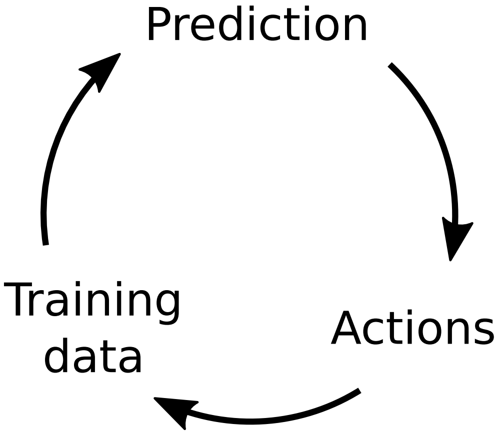
How does ML affect incentives to

- Repay debts?
- Comply with the law?
- Exert effort on the job?
- ...

Statistical Discrimination (Arrow 1973)

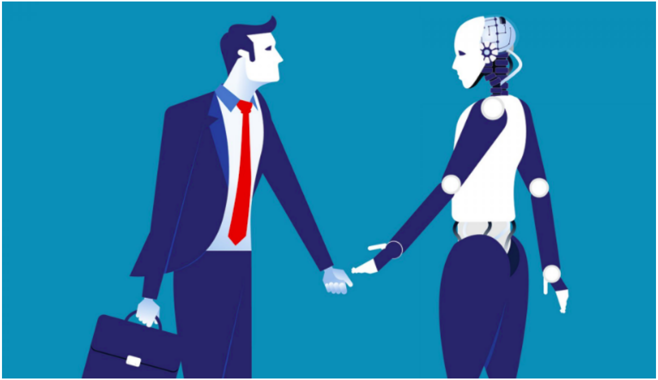


Arrow (1973) Applied to Machine Learning



Today: Justice

The robot lawyers are here - and they're winning

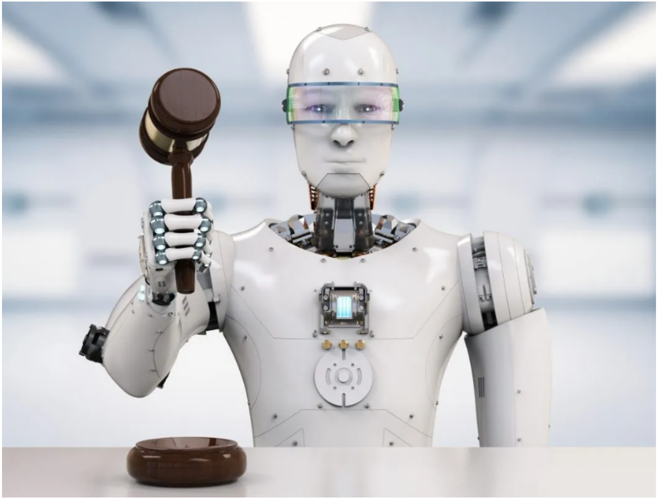


(source: BBC)

Predictions Encouraging Settlements

“Providing the parties with personalized outcome predictions doubles settlement rates and reduces average case duration” (Sadka, Seira, and Woodruff 2018)

What if an Artificial Intelligence *Decides* Court Cases?



Model

Consider randomly drawing an *agent* from a population.

Random variables:

- $A \in \{0, 1\}$ agent violates the law
- $F \in \mathcal{F}$ vector of fixed characteristics of agent
- $Z \in \mathcal{Z}$ vector of *evidence*
- Z_1 and Z_0: *potential evidence* if A is set to 1 and 0

$$Z = AZ_1 + (1 - A)\, Z_0$$

- $X = \{F, Z\}$

Punishment rule: $\pi(X) \in \{0, 1\} = \{\text{not punish}, \text{punish}\}$

Machine Learning Punishment

Assumption

We can perfectly estimate $\mathrm{E}[A \mid X]$ *by machine learning.*

Definition

A *machine learning punishment rule* punishes if $\mathrm{E}[A \mid X] > k$ for a constant k.

Machine Learning Optimally Reduces Errors

Proposition

A machine learning punishment rule "optimally reduces errors"

(=no other rule with lower type I and type II error rates)

Incentives

Assumption

Agent engages in crime ($A = 1$) if profit ($\Pi > 0$) is above increase in expected cost of punishment:

$$\Pi \geq \mathrm{E}\left[\pi\left(F, Z_1\right) - \pi\left(F, Z_0\right) \mid F\right]$$

Assumption

Potential evidence don't vary across types: $Z_1, Z_0 \perp F$

Assumption

A share ε *always engages in crime.*

Optimal Punishment

$$s(z) \equiv \frac{\Pr[Z_1 = z]}{\Pr[Z_0 = z]} = \text{strength of evidence } z$$

Proposition

"Optimal" to punish iff strength of evidence $s(z)$ is above threshold.

Optimal=deters all at minimal punishment costs $\mathrm{E}[\pi(X)]$.

Proposition

Threshold might depend on f.

Proposition

Optimal "non-discriminatory" rule has same threshold for all f

Robot Judges—the Short Run Effect

Proposition

An agent with evidence z and fixed characteristic f is punished by a machine learning punishment rule iff

$$s(z) > \frac{1 - \mathrm{E}[A \mid F = f]}{\mathrm{E}[A \mid F = f]} \frac{k}{1-k}$$

Robot Judges—the Short Run Effect

Proposition

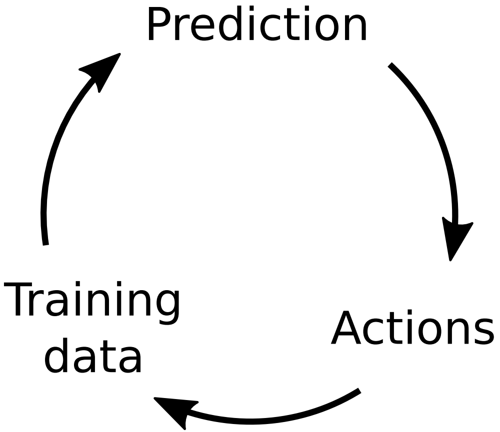
An agent with evidence z and fixed characteristic f is punished by a machine learning punishment rule iff

$$s(z) > \frac{1 - \mathrm{E}[A \mid F = f]}{\mathrm{E}[A \mid F = f]} \frac{k}{1-k}$$

Statistical discrimination and sub-optimal deterrence:

- "Innocent types" ($\mathrm{E}[A \mid F = f] = 0$) never punished
- "Guilty types" ($\mathrm{E}[A \mid F = f] = 1$) always punished
- ...

Arrow (1973) Applied to Machine Learning



Robot Judges—the Long Run Effect

Proposition

Assume profit from crime is observable ($\Pi = h(F)$) and machine learning punishment. Then all agents engage in crime in equilibrium.

Robot Judges—the Long Run Effect

Proposition

Assume profit from crime is observable ($\Pi = h(F)$) and machine learning punishment. Then all agents engage in crime in equilibrium.

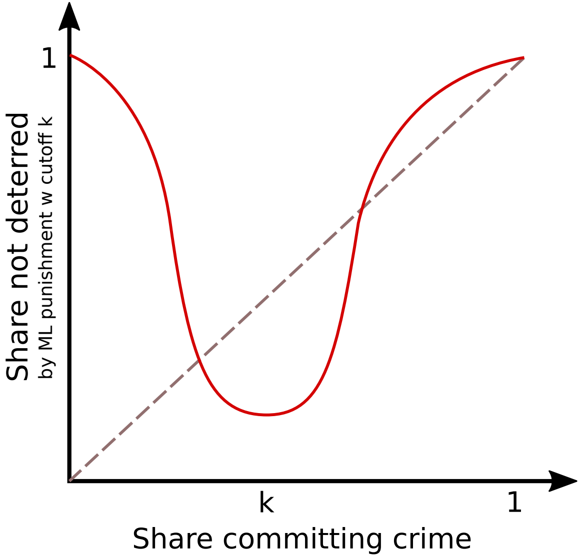
Proof:

- All of type f behave in same way
- $\Rightarrow$ fixed characteristics perfect predictor of crime
- $\Rightarrow$ ML punishes based on fixed characteristics
- $\Rightarrow$ No incentives

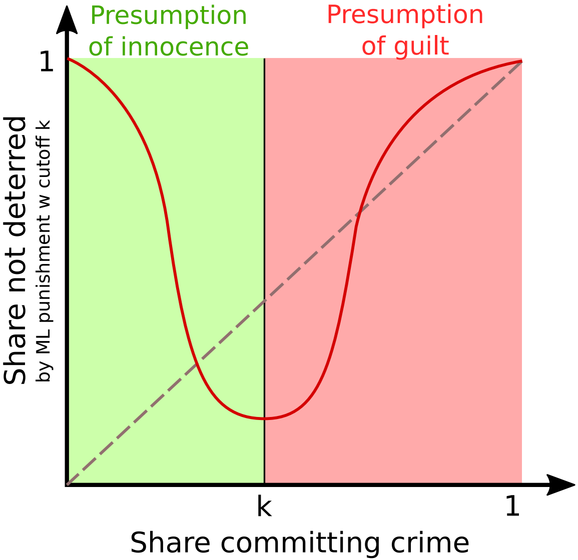
Robot Judges—the Long Run Effect

Consider all agents with fixed characteristics f

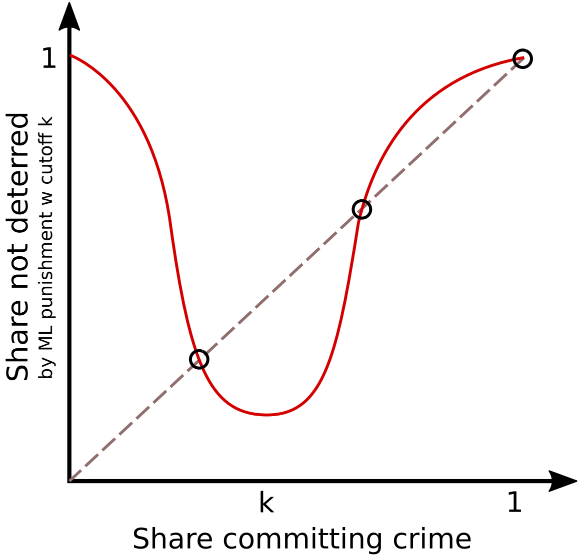
Robot Judges—the Long Run Effect



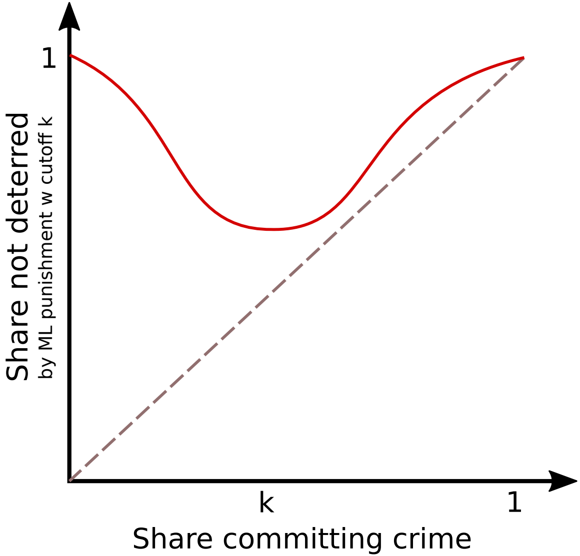
Robot Judges—the Long Run Effect



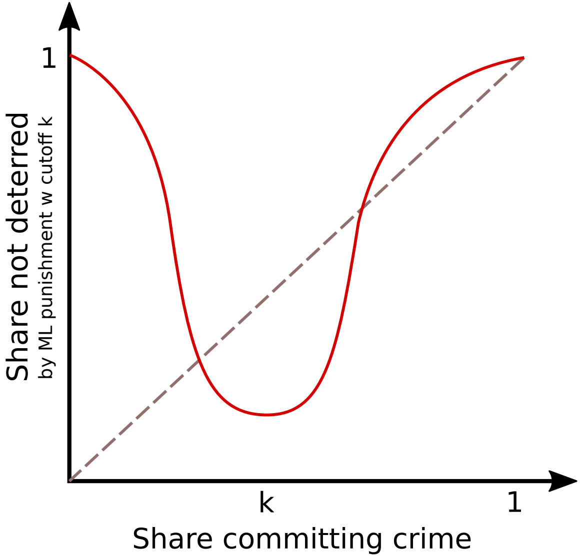
Three Equilibria



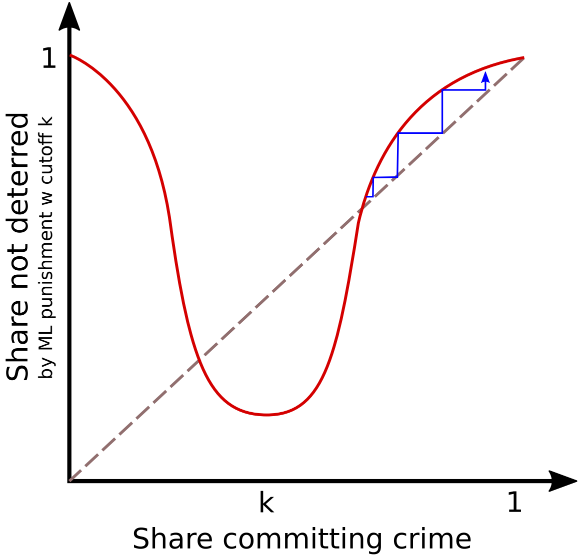
Example with One Equilibrium



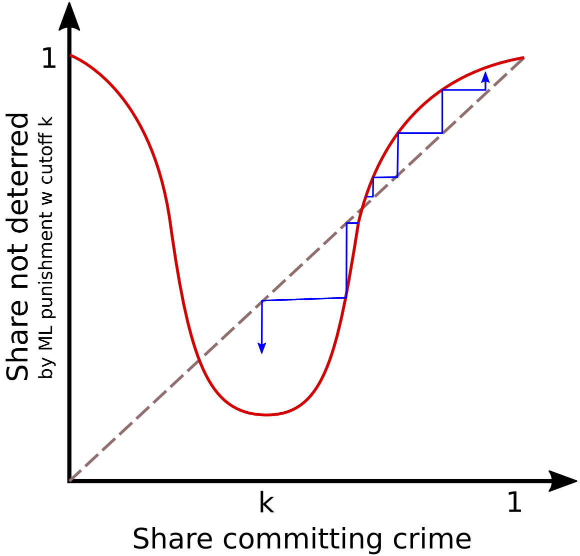
Equilibrium Selection



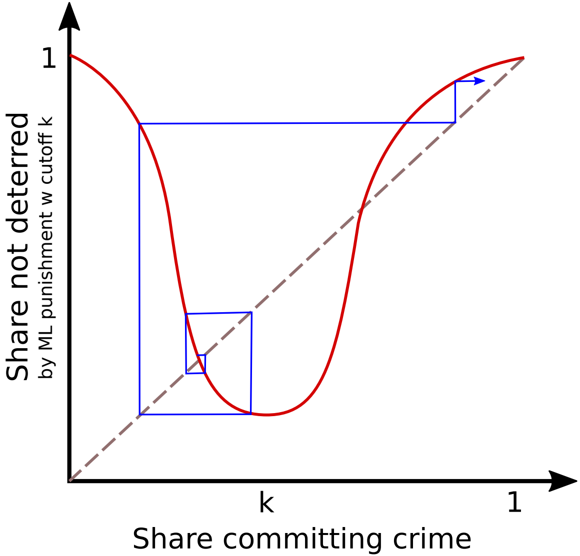
Bad Equilibrium Stable



Middle Equilibrium Unstable



Good Equilibrium Stable?



Summary

No equilibrium with zero crime rate

Equilibrium particularly bad when:

- Fixed characteristics highly predictive of Π
- Evidence is imprecise

Effect of punishment threshold k ambiguous

Fixes

Using only evidence?

- Punish iff $\mathrm{E}[A \mid Z = z] > k$?

Using only evidence?

- Punish iff $\mathrm{E}[A \mid Z = z] > k$? ✗

Using only evidence?

- Punish iff $\mathrm{E}[A \mid Z = z] > k$? ✗

- Optimal: Punish if $\frac{\Pr[A=1|Z=z]}{\Pr[A=0|Z=z]} \Big/ \frac{\Pr[A=1]}{\Pr[A=0]} > k$

Distinguishing Evidence from Fixed Characteristics

Assume

$$x = \{x_1, x_2, \ldots, x_n\}$$

Is x_2 a piece of evidence or a fixed characteristic?

Potential Solutions

Solution 1: Exclude known fixed characteristics

Potential Solutions

Solution 1: Exclude known fixed characteristics ✗

Potential Solutions

Solution 1: Exclude known fixed characteristics ✗

Solution 2: Debiasing ex-post ✓

Solution 2: Debiasing Ex Post

Assume observe only subset $G = h(F)$ of fixed characteristics.

Proposition

If $\Pi \perp F \mid G$, optimal non-discriminatory punishment punishes iff

$$\frac{\Pr[A=1 \mid X=x]}{\Pr[A=0 \mid X=x]} \Big/ \frac{\Pr[A=1 \mid G=g]}{\Pr[A=0 \mid G=g]} > k$$

for a constant k.

- Equalizes error rates across groups as in Hardt et al (2016)
- Does not respond to changes in the overall crime rate

Example

- Assume the benefit of crime is independent of other fixed characteristics conditional on income Y
- Then the strength of evidence of an agent with income y is

$$\frac{\Pr\left[A=1 \mid X=x\right]}{\Pr\left[A=0 \mid X=x\right]} \Big/ \frac{\Pr\left[A=1 \mid Y=y\right]}{\Pr\left[A=0 \mid Y=y\right]}$$

Empirical Application

Brazilian Labor Courts

Conciliation hearing in Brazilian labor court.

Collaboration with Legal Tech Firm

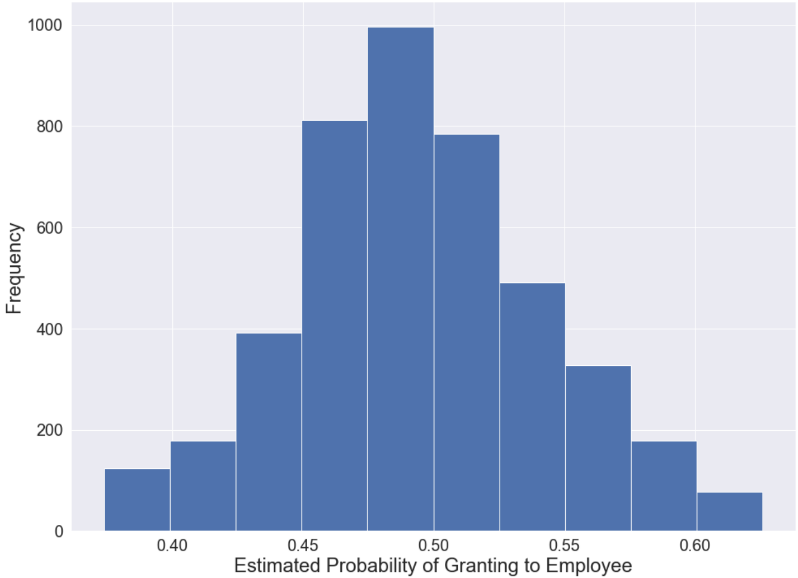
- 14 million labor court cases (currently training on 44,000)
- Includes litigant's arguments

ML models

Model	Features	F1
1	Litigant's arguments	0.75
2	Fixed characteristics of firm	0.51
3	Both	(in progress)

Fixed characteristics = sector and past cases

Predicted “Guilt” Based on Fixed Characteristics



Testing Statistical Discrimination

Optimal non-discriminatory rule:

- $\frac{\text{Type I errors}}{\text{Type II errors}}$ constant across $f \in \mathcal{F}$

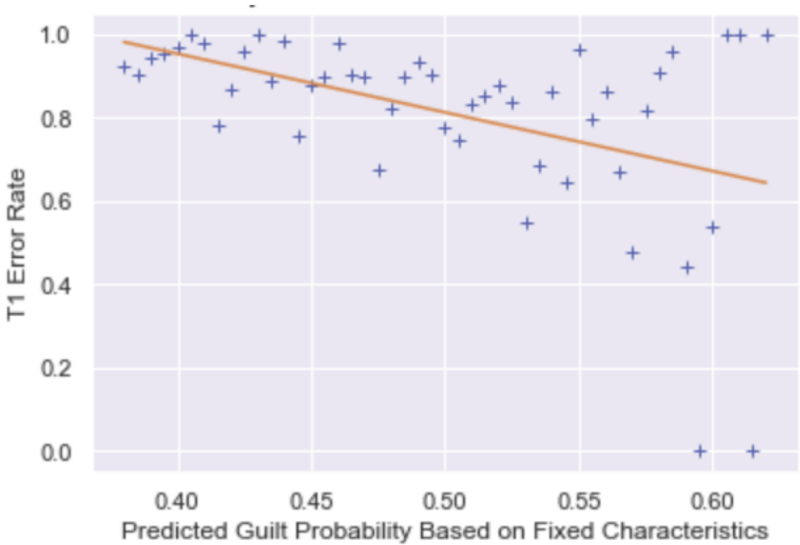
Naive machine learning rule:

- $\frac{\text{Type I errors}}{\text{Type II errors}}$ increasing in $\mathrm{E}[A \mid F = f]$.

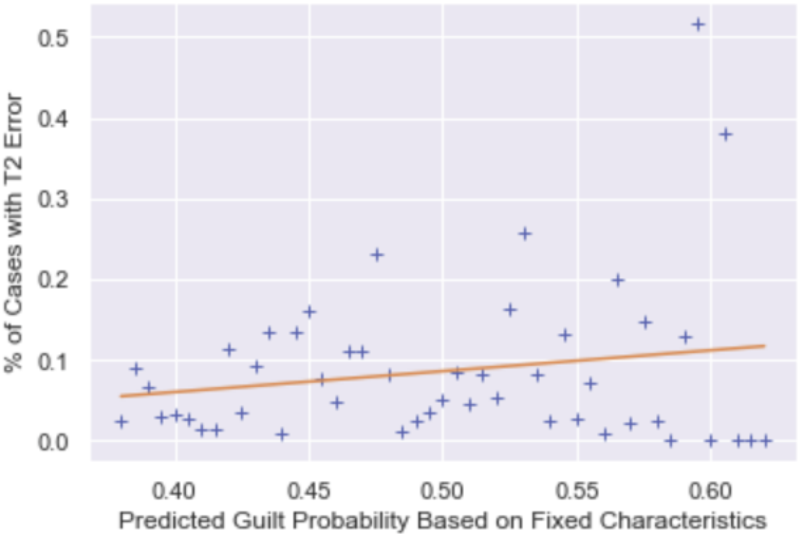
Using this, we can test:

1. how bad the naive ML rule is in practice
2. whether Solution 1-3 works

Type I Errors



Type II Errors



Conclusions

- Machine learning can lead to incentive problems
- Self-fulfilling prophecies
  - Full eradication of undesired behavior impossible
- Especially when:
  - fixed characteristics highly predictive of behavior
  - actions are imprecisely observed
- Debiasing might work

Discussion