

ECONOMICS & COMPUTATION

Spring 2026 | Lecture 20

Fairness in Machine Learning

Ariel Procaccia | Harvard University

UNFAIRNESS

- AI algorithms are supposedly unbiased
- But they are trained based on data that encodes societal biases, and may exacerbate those biases
- There is a significant body of work that alleges discrimination by AI algorithms

EXAMPLE: AD DELIVERY

Title	URL	Coefficient	appears in agents		total appearances	
			female	male	female	male
Top ads for identifying the simulated female group						
Jobs (Hiring Now)	www.jobsinyourarea.co	0.34	6	3	45	8
4Runner Parts Service	www.westernpatoyotaservice.com	0.281	6	2	36	5
Criminal Justice Program	www3.mc3.edu/Criminal+Justice	0.247	5	1	29	1
Goodwill - Hiring	goodwill.careerboutique.com	0.22	45	15	121	39
UMUC Cyber Training	www.umuc.edu/cybersecuritytraining	0.199	19	17	38	30
Top ads for identifying agents in the simulated male group						
\$200k+ Jobs - Execs Only	careerchange.com	-0.704	60	402	311	1816
Find Next \$200k+ Job	careerchange.com	-0.262	2	11	7	36
Become a Youth Counselor	www.youthcounseling.degreeleap.com	-0.253	0	45	0	310
CDL-A OTR Trucking Jobs	www.tadivers.com/OTRJobs	-0.149	0	1	0	8
Free Resume Templates	resume-templates.resume-now.com	-0.149	3	1	8	10

[Datta et al. 2015]

EXAMPLE: CRIMINAL JUSTICE



Bernard Parker, left, was rated high risk, Dylan Pugett was rated low risk. (Josh Ritchie for ProPublica)

Machine Bias

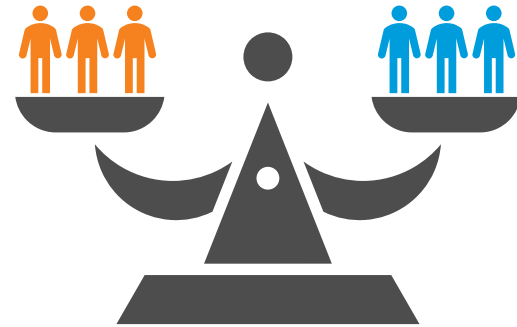
There's software used across the country to predict future criminals. And it's biased against blacks.

*by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016*

TWO TYPES OF FAIRNESS



Individual fairness



Group fairness



Cynthia Dwork

1958–

Professor of Computer Science at Harvard. In the last 20 years, played a pivotal role in the formation of differential privacy and fair AI.

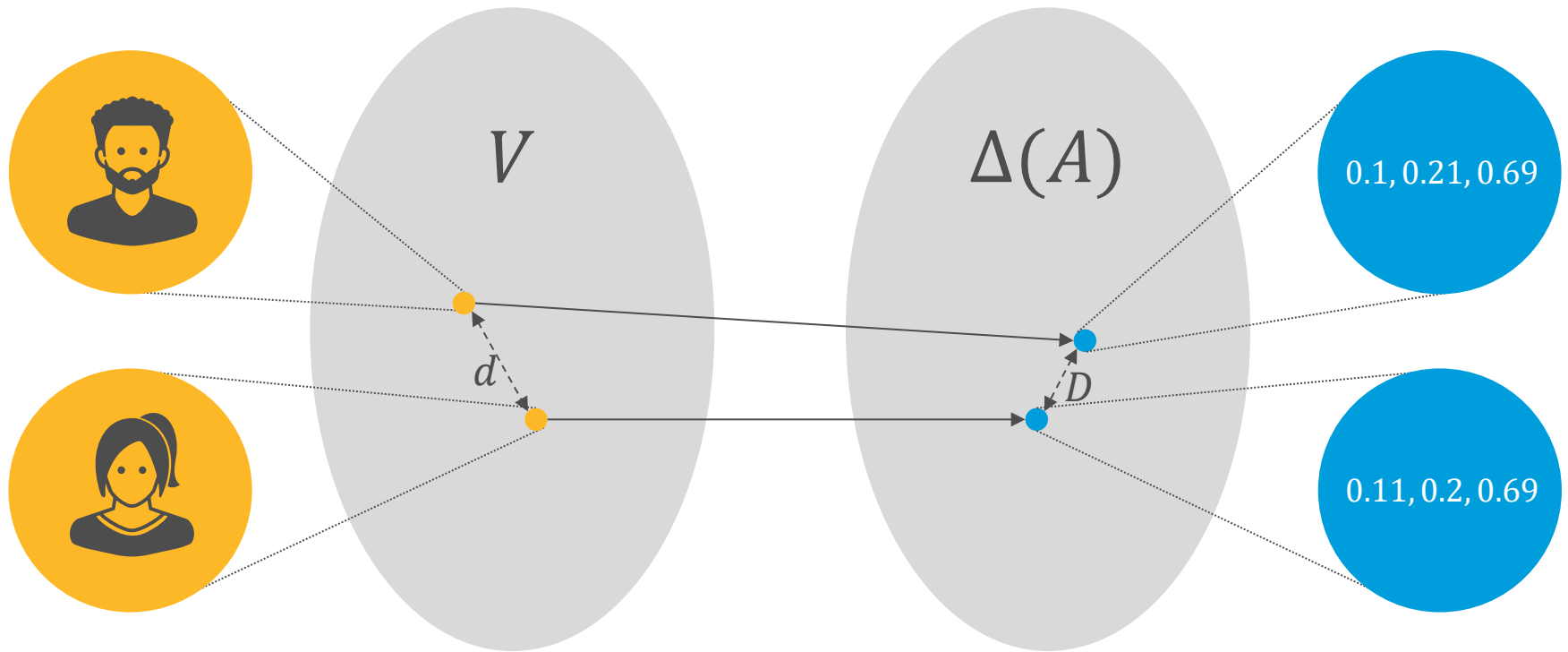


SIMILARITY-BASED FAIRNESS

- Set of individuals V and outcomes A
- Randomized classifier $M: V \rightarrow \Delta(A)$ where $\Delta(A)$ is distributions over outcomes
- Metric on individuals $d: V \times V \rightarrow \mathbb{R}^+$
- Metric D on distributions over outcomes
- M satisfies the **Lipschitz property** if for all $x, y \in V$,

$$D(M(x), M(y)) \leq d(x, y)$$

SIMILARITY-BASED FAIRNESS



SIMILARITY-BASED FAIRNESS

- We can get a Lipschitz classifier by setting $M(x) = M(y)$ for all $x, y \in V$
- But we want to minimize a **loss function**
$$L: V \times A \rightarrow \mathbb{R}^+$$
- This leads to the optimization problem

$$\begin{aligned} \min & \sum_{x \in V} \sum_{a \in A} \mu_x(a) \cdot L(x, a) \\ \text{s.t.} & \forall x, y \in V, D(\mu_x, \mu_y) \leq d(x, y) \\ & \forall x \in V, \mu_x \in \Delta(A) \end{aligned}$$

SIMILARITY-BASED FAIRNESS

- Various options for the metric D
- Example: **total variation**, defined for distributions P and Q as

$$D_{tv}(P, Q) = \frac{1}{2} \sum_{a \in A} |P(a) - Q(a)|$$

- When $D = D_{tv}$, the optimization problem is a linear program

Poll 1

Where would the similarity metric come from?



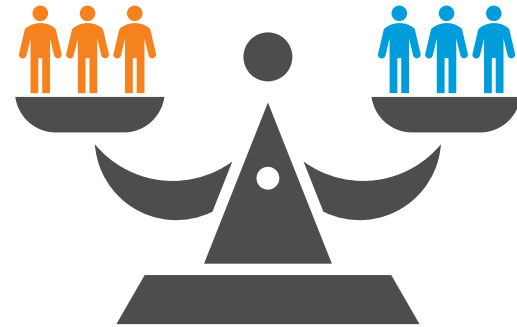
ENVY-FREENESS

- Each $x \in V$ has a utility u_{xa} for each outcome $a \in A$
- A randomized classifier M is **envy free** if and only if for all $x, y \in V$,
$$\mathbb{E}_{a \sim M(x)}[u_{xa}] \geq \mathbb{E}_{a \sim M(y)}[u_{xa}]$$
- This gives a completely different way of thinking about individual fairness
- But envy-freeness isn't useful in situations where there is a desirable and an undesirable outcome, like bail and loans

TWO TYPES OF FAIRNESS



Individual fairness

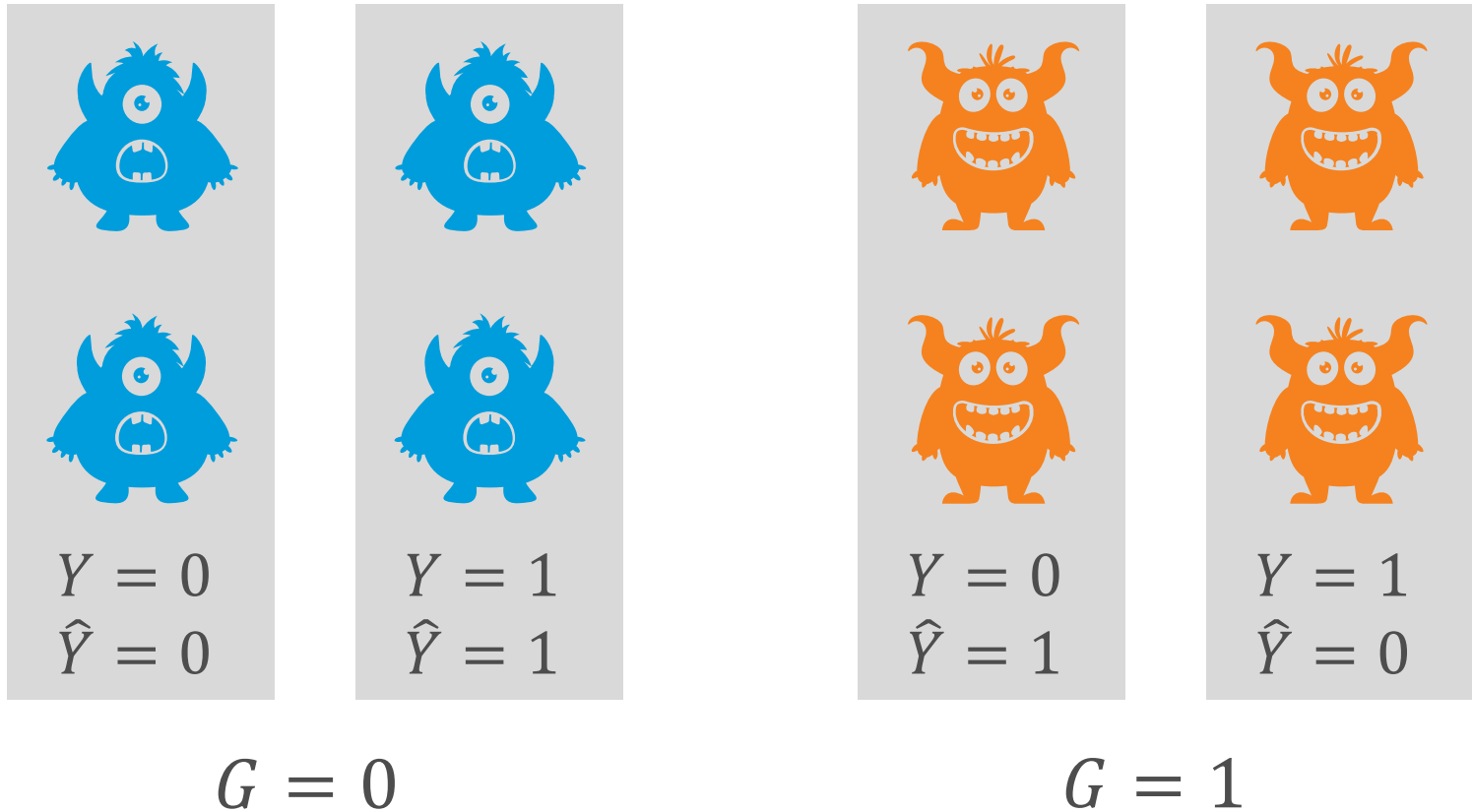


Group fairness

DEMOGRAPHIC PARITY

- Assume we are making a binary decision $\hat{Y} \in \{0,1\}$, and there is a legally protected attribute $G \in \{0,1\}$
- **Demographic parity:**
$$\Pr[\hat{Y} = 1 \mid G = 0] = \Pr[\hat{Y} = 1 \mid G = 1]$$
- May accept unqualified individuals when $G = 0$, and qualified individuals when $G = 1$!

DEMOGRAPHIC PARITY



This classifier satisfies demographic parity!

EQUALIZED ODDS

- \hat{Y} satisfies **equalized odds** with respect to protected attribute G if the groups have equal false positive and false negative rates
- That is, for all $y, \hat{y} \in \{0,1\}$,
$$\begin{aligned}\Pr[\hat{Y} = \hat{y} \mid G = 0, Y = y] \\ = \Pr[\hat{Y} = \hat{y} \mid G = 1, Y = y]\end{aligned}$$

RELATION BETWEEN PROPERTIES

- **Demographic parity:**

$$\Pr[\hat{Y} = 1 \mid G = 0] = \Pr[\hat{Y} = 1 \mid G = 1]$$

- **Equalized odds:** For all $y, \hat{y} \in \{0,1\}$,

$$\begin{aligned} \Pr[\hat{Y} = \hat{y} \mid G = 0, Y = y] \\ = \Pr[\hat{Y} = \hat{y} \mid G = 1, Y = y] \end{aligned}$$

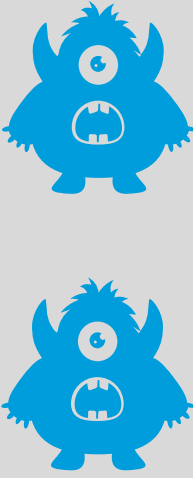
Poll 2

What is the relation between demographic parity and equalized odds?

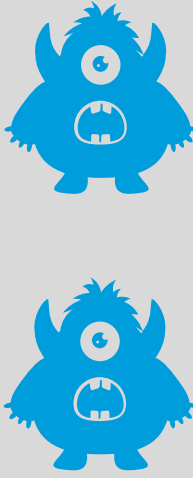
- DP \Rightarrow EO
- DP \Leftrightarrow EO
- EO \Rightarrow DP
- Incomparable



RELATION BETWEEN PROPERTIES

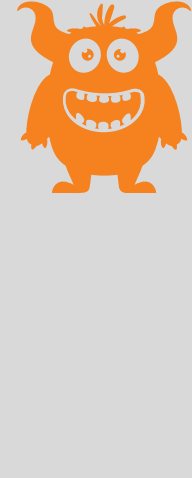


$Y = 0$
 $\hat{Y} = 0$



$Y = 1$
 $\hat{Y} = 1$

$G = 0$



$Y = 0$
 $\hat{Y} = 0$

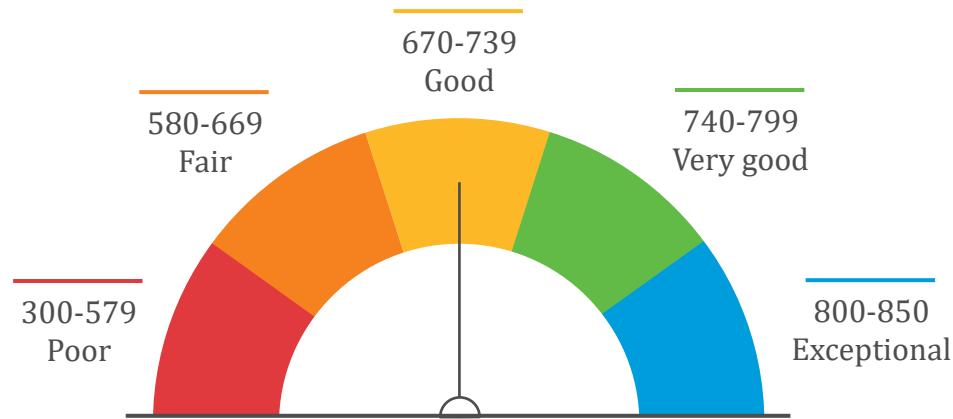


$Y = 1$
 $\hat{Y} = 1$

$G = 1$

$\hat{Y} = Y$ may not satisfy demographic parity!

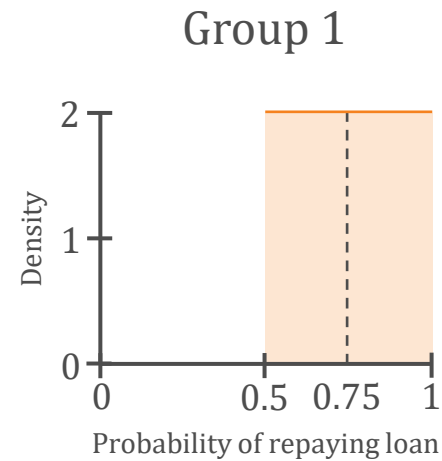
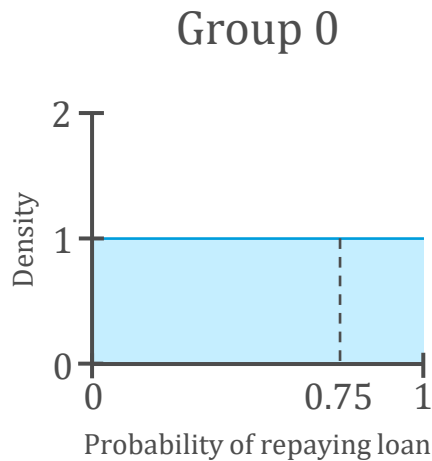
EQUALIZED ODDS: RISK SCORES



- FICO scores are a proprietary classifier widely used in the United States to predict credit worthiness
- Range from 300 to 850, where cutoff of 620 is commonly used for prime-rate loans, which corresponds to a default rate of 18%

EQUALIZED ODDS: RISK SCORES

Suppose a bank gives a loan ($\hat{Y} = 1$) if and only if the estimated probability of repayment is at least 0.75

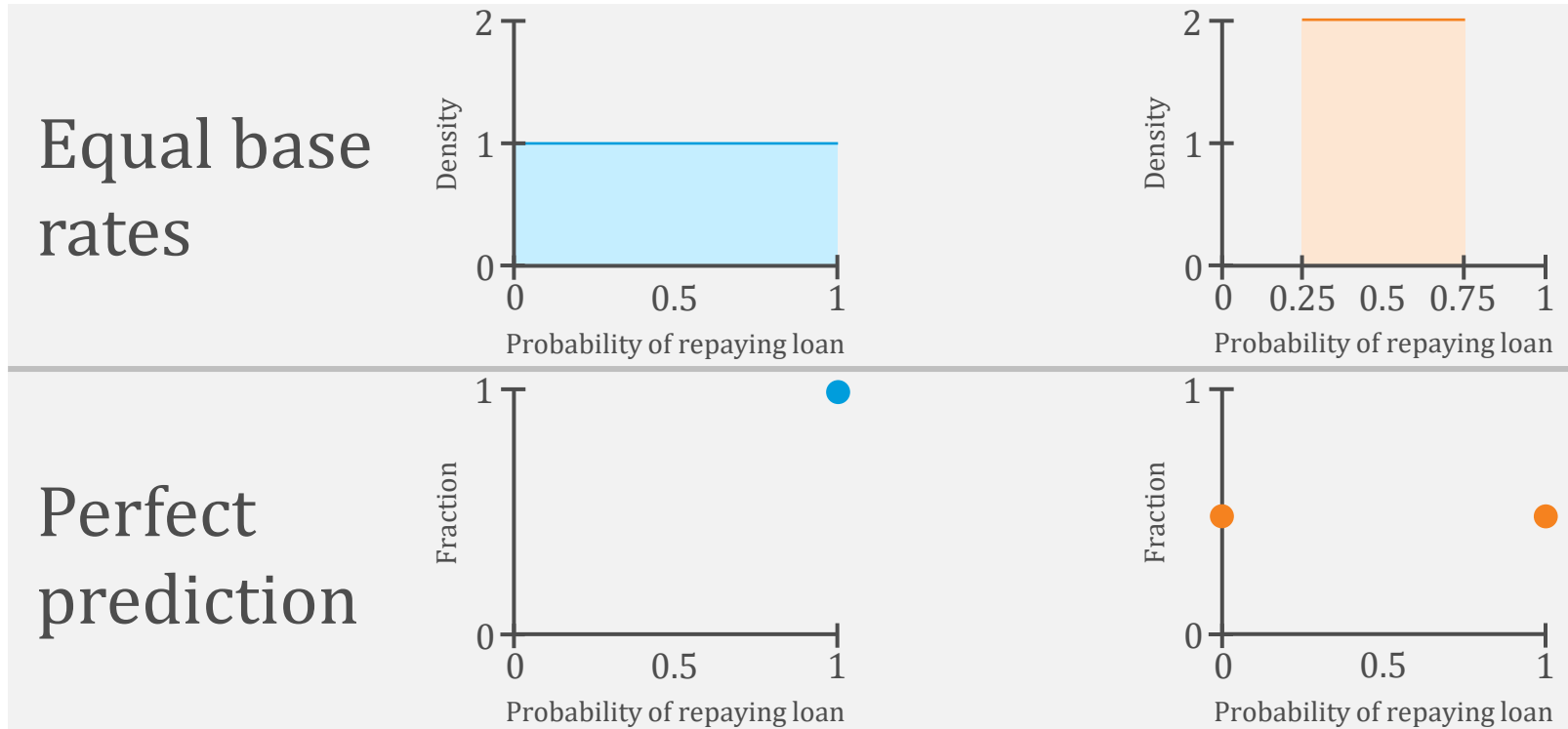


$$\Pr[\hat{Y} = 0 | G = 0, Y = 1] = \frac{0.75 \cdot 0.375}{0.5} = 0.56$$

$$\Pr[\hat{Y} = 0 | G = 1, Y = 1] = \frac{0.5 \cdot 0.625}{0.75} = 0.41$$

The risk threshold classifier violates equalized odds even if predictions are **calibrated**

EQUALIZED ODDS: RISK SCORES



Theorem (informal): If a risk assignment satisfies calibration and equalized odds, the instance must allow for perfect prediction or have equal base rates