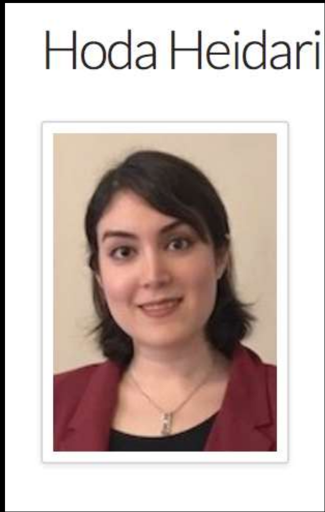
Digital Society Initiative, University of Zurich

Algorithmic fairness, discrimination, and equality of opportunity

Michele Loi

Fourteenth Winter School
Inequality and Social Welfare Theory
Canazei
University of Verona



A Moral Framework for Understanding Fair ML through Economic Models of Equality of Opportunity

Hoda Heidari
ETH Zürich
hheidari@inf.ethz.ch

Michele Loi
University of Zürich
michele.loi@uzh.ch

Krishna P. Gummadi
MPI-SWS
gummadi@mpi-sws.org

Andreas Krause
ETH Zürich
krausea@ethz.ch

1. Machine learning

"An agent is **learning** if it improves its performance on future tasks after making observations about the world."

From: Russell, Staurt J., and Peter Norvig. 2010. *Artificial Intelligence (A Modern Approach)*. 3rd ed. Upper Saddler River, NJ: Prentice Hall.

Machine learning

"An agent is **learning** if it improves its performance on future tasks after making observations about the world."

unsupervised

"In **unsupervised learning** the agent learns patterns in the input even though no explicit feedback is supplied. The most common unsupervised learning task is **clustering**"

supervised

"In **supervised learning** the agent observes some example input–output pairs and learns a function that maps from input to output. […] [e.g.] the inputs are camera images and the outputs again come from a teacher who says "that's a bus.""

From: Russell, Staurt J., and Peter Norvig. 2010. *Artificial Intelligence (A Modern Approach)*. 3rd ed. Upper Saddler River, NJ: Prentice Hall.

input

output

dog
.9

not dog
.1

Algorithms for machine learning

E.g. Decision trees, linear regression, logistic regression

From: Russell, Staurt J., and Peter Norvig. 2010. *Artificial Intelligence (A Modern Approach)*. 3rd ed. Upper Saddler River, NJ: Prentice Hall.

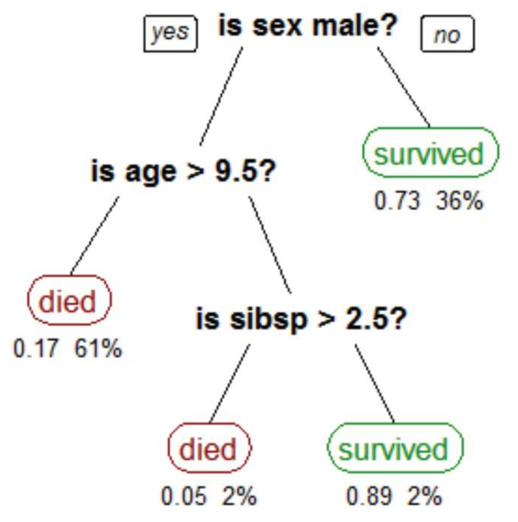


Image taken from wikipedia

```
function DECISION-TREE-LEARNING(examples, attributes, parent_examples) returns
a tree

  if examples is empty then return PLURALITY-VALUE(parent_examples)
  else if all examples have the same classification then return the classification
  else if attributes is empty then return PLURALITY-VALUE(examples)
  else
     A ← argmax_{a ∈ attributes} IMPORTANCE(a, examples)
     tree ← a new decision tree with root test A
     for each value v_k of A do
        exs ← {e : e ∈ examples and e.A = v_k}
        subtree ← DECISION-TREE-LEARNING(exs, attributes − A, examples)
        add a branch to tree with label (A = v_k) and subtree subtree
     return tree
```

Algorithms for machine learning

Artificial neural networks

(with many layers: deep learning)

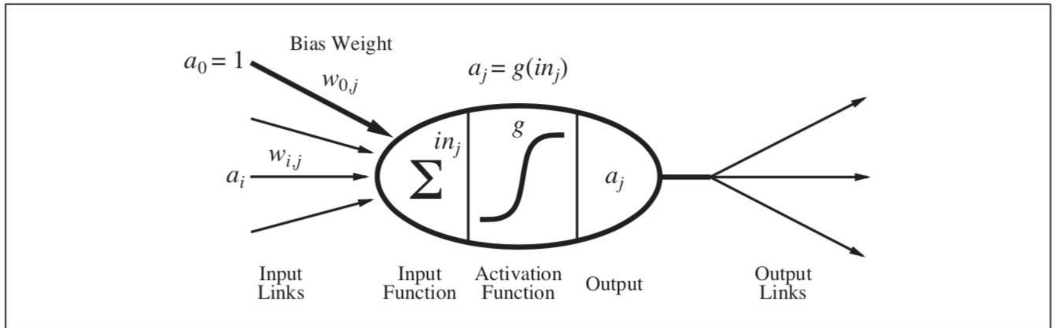
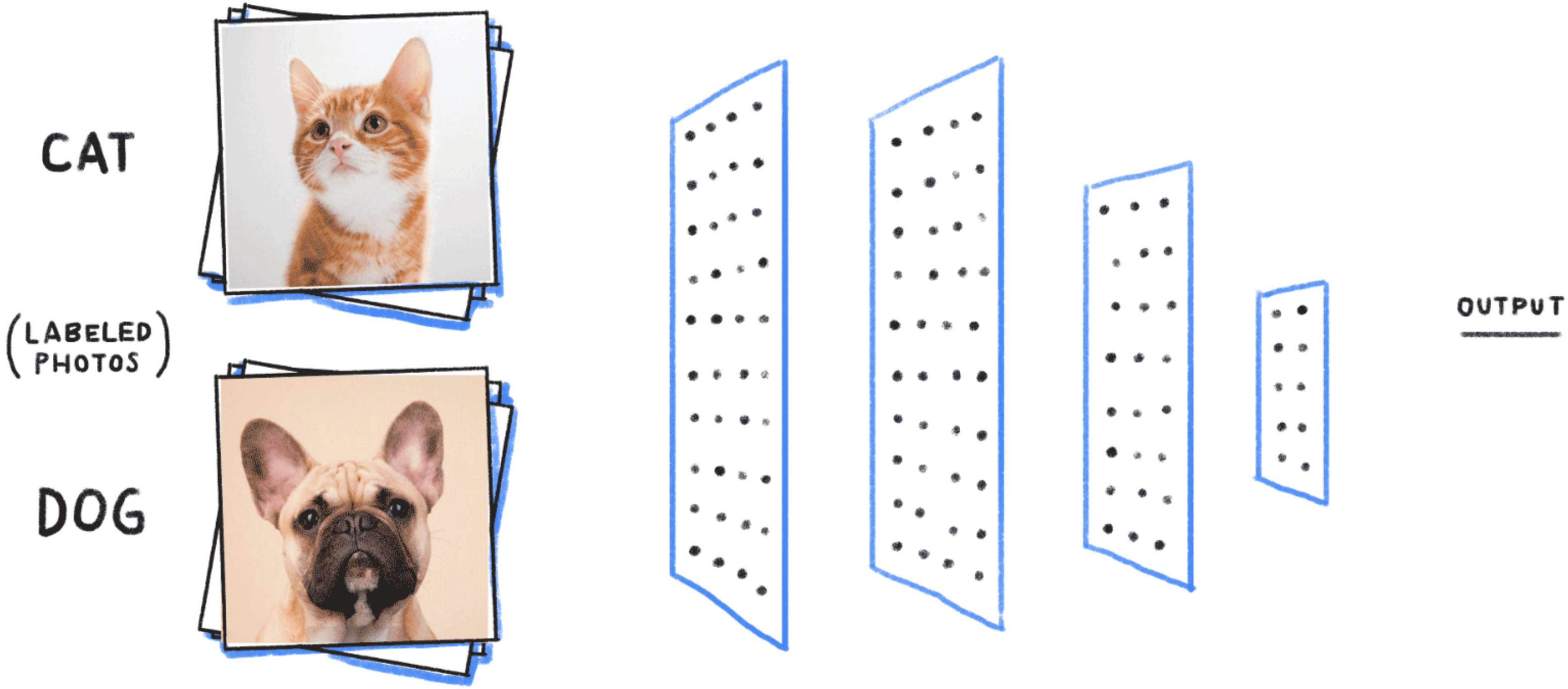


Figure 18.19 A simple mathematical model for a neuron. The unit's output activation is $a_j = g(\sum_{i=0}^{n} w_{i,j} a_i)$, where a_i is the output activation of unit i and $w_{i,j}$ is the weight on the link from unit i to this unit.

From: Russell, Staurt J., and Peter Norvig. 2010. *Artificial Intelligence (A Modern Approach)*. 3rd ed. Upper Saddler River, NJ: Prentice Hall.



from : Becoming Human: Artificial Intelligence Magazine

models from machine learning can be ‘racist’

Google personalised ad for public records

Trevor John **Trevor John, Arrested?**

models from machine learning can be ‘racist’

Computing
The Observer

Interview

'A white mask worked better': why algorithms are not colour blind

By Ian Tucker

Sun 28 May 2017 08.27 EDT

3,817 498

When Joy Buolamwini found that a robot recognised her face better when she wore a white mask, she knew a problem needed fixing

▲ Joy Buolamwini gives her TED talk on the bias of algorithms Photograph: TED

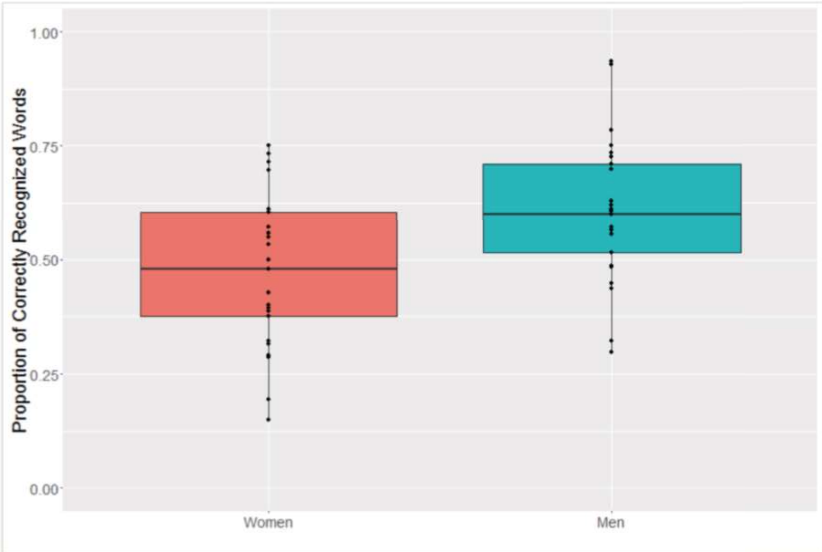
algorithms can be ‘sexist’

GOOGLE’S SPEECH RECOGNITION HAS A GENDER BIAS

Posted by **Rachael Tatman** in **Uncategorized** and tagged with **computational linguistics**, **gender**, **linguistics**, **sociolinguistics**, **speech recognition**, **speech signal**, **speech technology**

In my last post, I looked at how Google’s automatic speech recognition worked with different dialects. To get this data, I hand-checked annotations more than 1500 words from fifty different accent tag videos .

Now, because I’m a sociolinguist and I know that it’s important to stratify your samples, I made sure I had an equal number of male and female speakers for each dialect. And when I compared performance on male and female talkers, I found something deeply disturbing: YouTube’s auto captions consistently performed better on male voices than female voice ($t(47) = -2.7$, $p < 0.01$.) . (You can see my data and analysis here.)



On average, for each female speaker less than half (47%) her words were captioned correctly. The average male speaker, on the other hand, was captioned correctly 60% of the time.

algorithms can be ‘sexist’

CNBC

MENU MARKETS BUSINESS INVESTING TECH POLITICS CNBC TV

RETAIL

APPAREL | DISCOUNTERS | DEPARTMENT STORES | E-COMMERCE | FOOD AND BEV

Amazon scraps a secret A.I. recruiting tool that showed bias against women

- Amazon.com's machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.
- The team had been building computer programs since 2014 to review job applicants' resumes with the aim of mechanizing the search for top talent, five people familiar with the effort told Reuters.
- The company's experimental hiring tool used artificial intelligence to give job candidates scores ranging from one to five stars — much like shoppers rate products on Amazon, some of the people said.

Published 6:15 AM ET Wed, 10 Oct 2018 | Updated 2:25 PM ET Thu, 11 Oct 2018

REUTERS

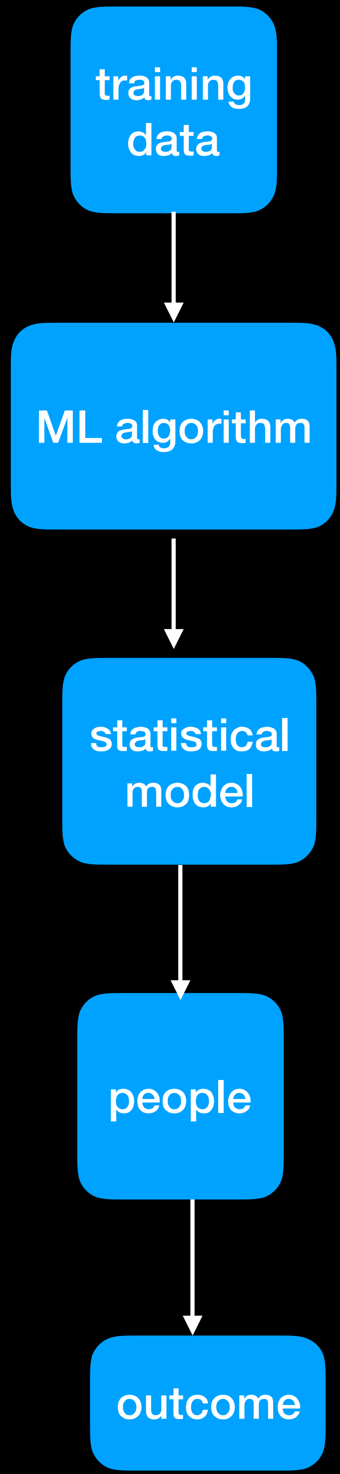
FAT / ML 2018 2017 2016 2015 2014 Organization Resources Mailing list

Fairness, Accountability, and Transparency in Machine Learning

Bringing together a growing community of researchers and practitioners concerned with fairness, accountability, and transparency in machine learning

2. learning fair predictors

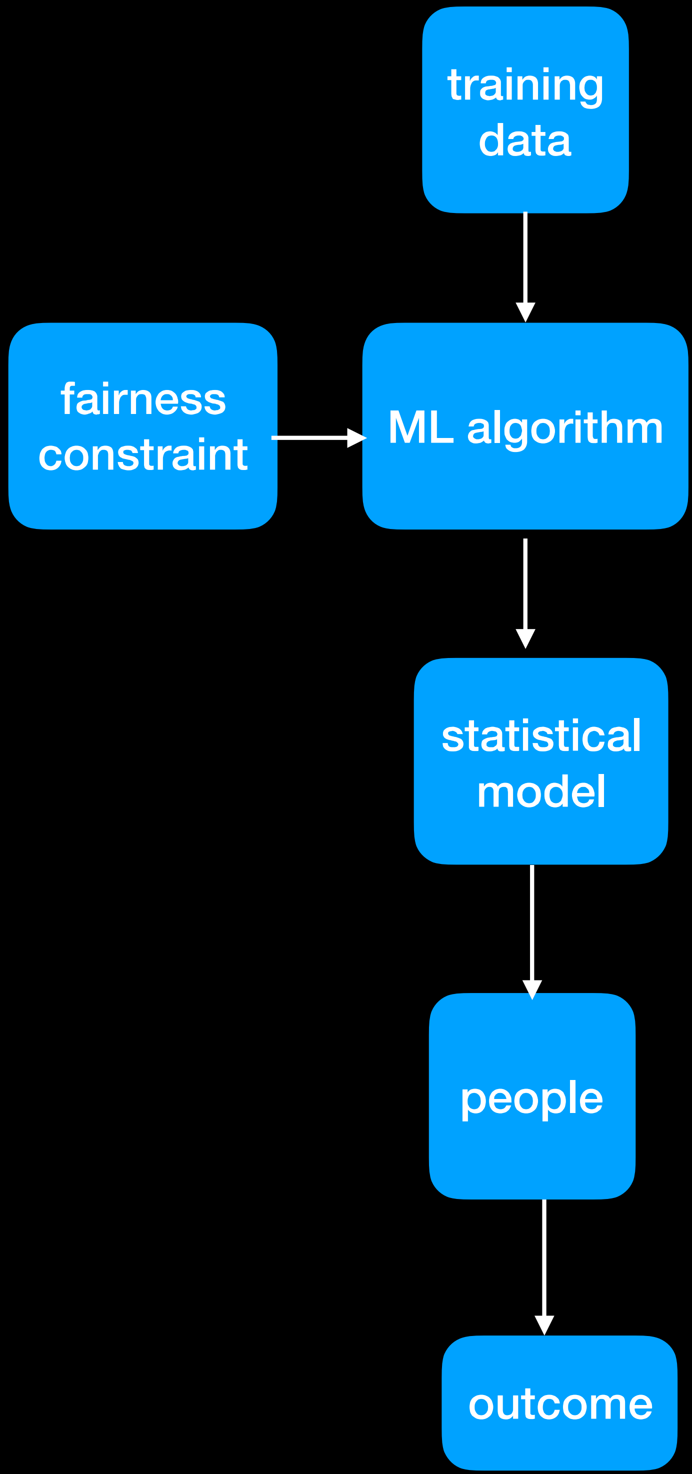
Q. How do computer scientists achieve 'fair' ML predictors?



2. learning fair predictors

Q. How do computer scientists achieve 'fair' ML predictors?

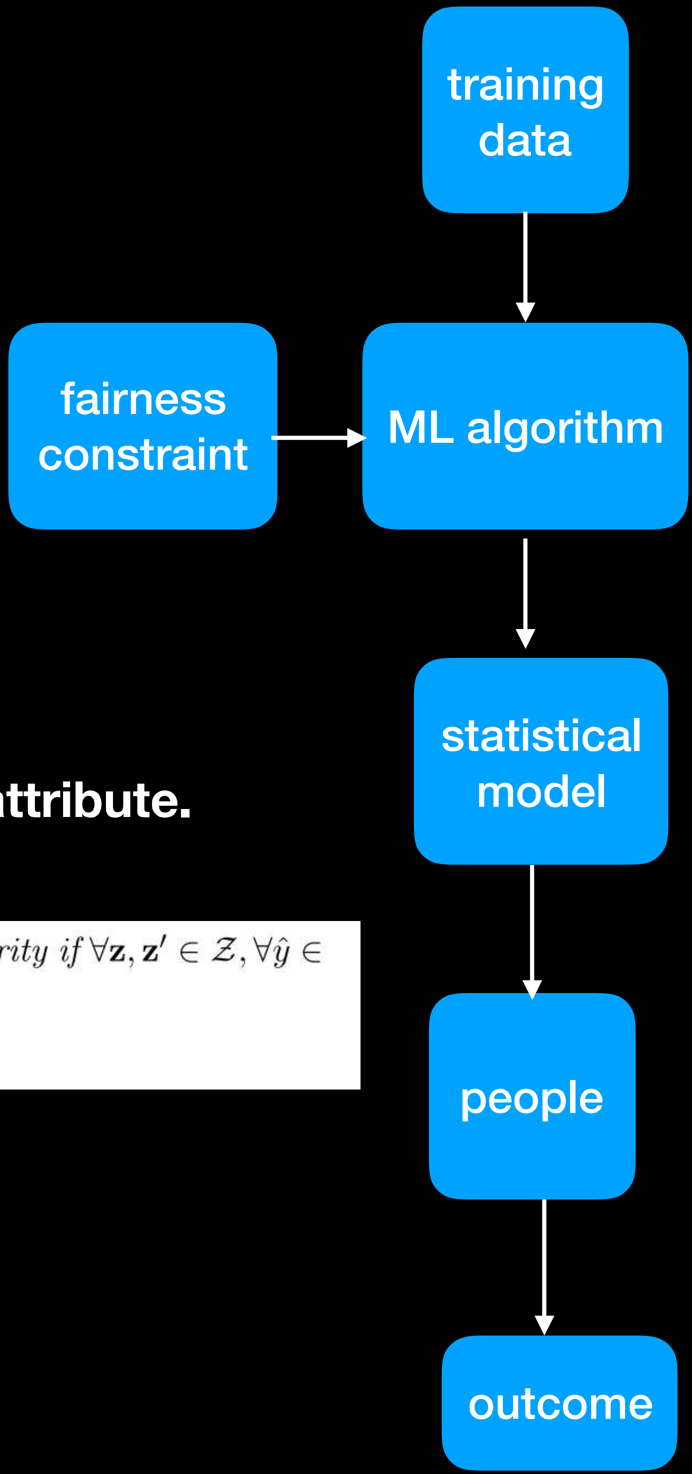
A. By requiring predictive models to satisfy mathematically defined fairness constraints.



2. learning fair predictors

Q. How do computer scientists achieve ‘fair’ ML predictors?

A. By requiring predictive models to satisfy mathematically defined fairness constraints.



E.g. statistical parity: the output of the prediction/classification does not depend on the ‘sensitive’ attribute.

Definition 3 (Statistical Parity) *A predictive model h satisfies statistical parity if* $\forall \mathbf{z}, \mathbf{z}' \in \mathcal{Z}, \forall \hat{y} \in \mathcal{Y}$:

$$\mathbb{P}_{(\mathbf{X},Y)\sim F}[h(\mathbf{X}) = \hat{y} | \mathbf{Z} = \mathbf{z}] = \mathbb{P}_{(\mathbf{X},Y)\sim F}[h(\mathbf{X}) = \hat{y} | \mathbf{Z} = \mathbf{z}'].$$

e.g.

X = CV data
Y^ = predicted to be an excellent hire
Z = [male, female]

Do you believe statistical parity is fair?

Suppose you are developing a statistical model to help judges decide if a person towards the end of his or her jail sentence should be released on parole

questionnaire based

	Predictive Feature	Example Question
1.	Current Charges	Are you currently charged with a misdemeanor, non-violent felony or violent felony?
2.	Criminal History: self	How many times have you violated your parole?
3.	Substance Abuse	Did you use heroin, cocaine, crack or meth as a juvenile?
4.	Stability of Employment & Living Situation	How often do you have trouble paying bills?
5.	Personality	Do you have the ability to "sweet talk" people into getting what you want?
6.	Criminal Attitudes	Do you think that a hungry person has a right to steal?
7.	Neighborhood Safety	Is there much crime in your neighborhood?
8.	Criminal History: family and friends	How many of your friends have ever been arrested?
9.	Quality of Social Life & Free Time	Do you often feel left out of things?
10.	Education & School Behavior	What were your usual grades in high school?

Table 1: The ten features assessed in our survey and the questions provided as examples in the scenario. The features and questions are drawn from the COMPAS questionnaire.

From: Grgić-Hlača, Nina, Elissa M. Redmiles, Krishna P. Gummadi, and Adrian Weller. 2018. "Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction." doi:10.1145/3178876.3186138.

Statistical parity: the output of the classifier does not depend on the ‘sensitive’ attribute.

Definition 3 (Statistical Parity) *A predictive model h satisfies statistical parity if* $\forall \mathbf{z}, \mathbf{z}' \in \mathcal{Z}, \forall \hat{y} \in \mathcal{Y}$:

$$\mathbb{P}_{(\mathbf{X},Y)\sim F}[h(\mathbf{X}) = \hat{y} | \mathbf{Z} = \mathbf{z}] = \mathbb{P}_{(\mathbf{X},Y)\sim F}[h(\mathbf{X}) = \hat{y} | \mathbf{Z} = \mathbf{z}'].$$

e.g.

X = questionnaire data

Y^ = predicted to not reoffend while on Parole

Z = [male, female]

2. learning fair predictors

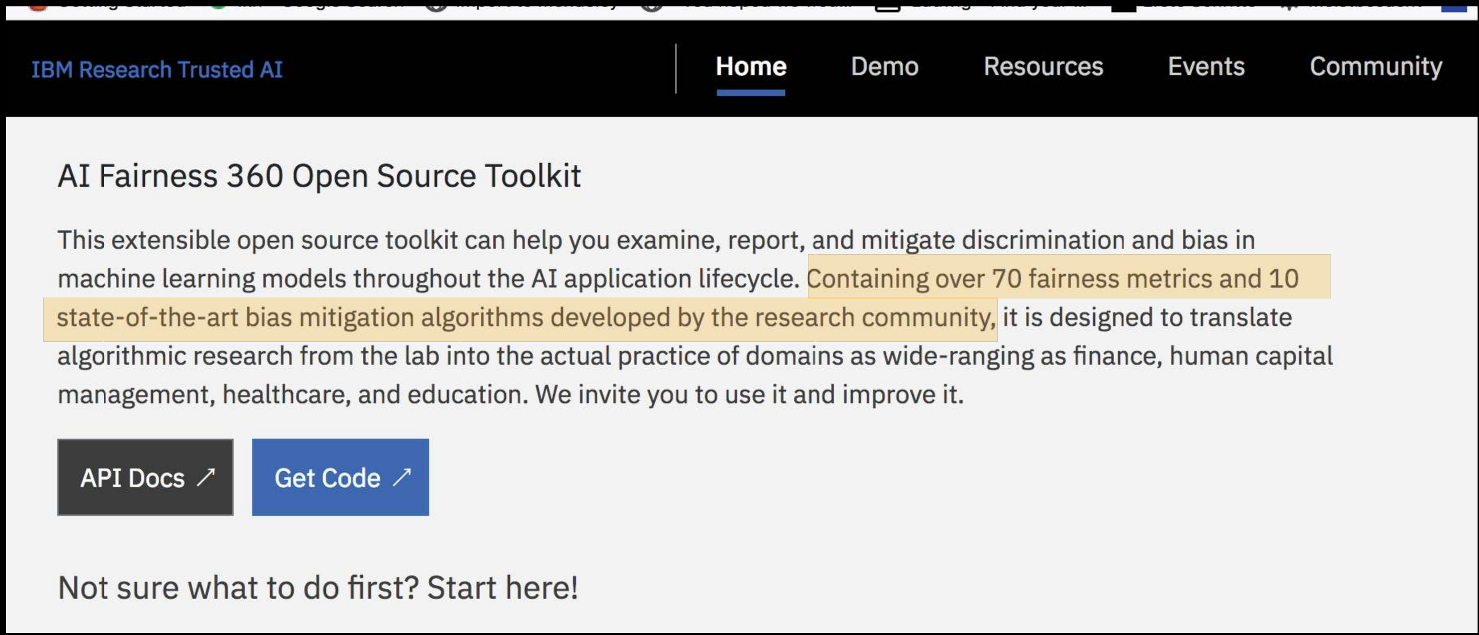
Q. How do computer scientists achieve 'fair' ML predictors?

A. By requiring output predictive models to satisfy mathematically defined fairness constraints.

Statistical parity ≠ Predictive value parity ≠ Calibration etc… !

There are different plausible constraints

3. which constraints?



Which one of the 70 fairness metrics fits my (moral) needs?

The equality of opportunity idea.

1. Moritz Hardt

Hardt, Moritz, Eric Price, and Nathan Srebro. 2016. "Equality of Opportunity in Supervised Learning." *ArXiv:1610.02413 [Cs]*, October. http://arxiv.org/abs/1610.02413.

The equality of opportunity idea.

1. Moritz Hardt

Hardt, Moritz, Eric Price, and Nathan Srebro. 2016. “Equality of Opportunity in Supervised Learning.” *ArXiv:1610.02413 [Cs]*, October. http://arxiv.org/abs/1610.02413.

Equal odds for binary classifiers:

$$\Pr\left\{\widehat{Y} = 1 \mid A = 0, Y = y\right\} = \Pr\left\{\widehat{Y} = 1 \mid A = 1, Y = y\right\}, \quad y \in \{0,1\}$$

Prediction and protected variable (A) are independent conditional on Y (actual label)

The equality of opportunity idea.

1. Moritz Hardt

Hardt, Moritz, Eric Price, and Nathan Srebro. 2016. "Equality of Opportunity in Supervised Learning." *ArXiv:1610.02413 [Cs]*, October. http://arxiv.org/abs/1610.02413.

Equal odds for binary classifiers:

$$\Pr\left\{\widehat{Y}=1 \mid A=0, Y=y\right\}=\Pr\left\{\widehat{Y}=1 \mid A=1, Y=y\right\}, \quad y \in \{0,1\}$$

Prediction and protected variable (A) are independent conditional on Y (actual label)

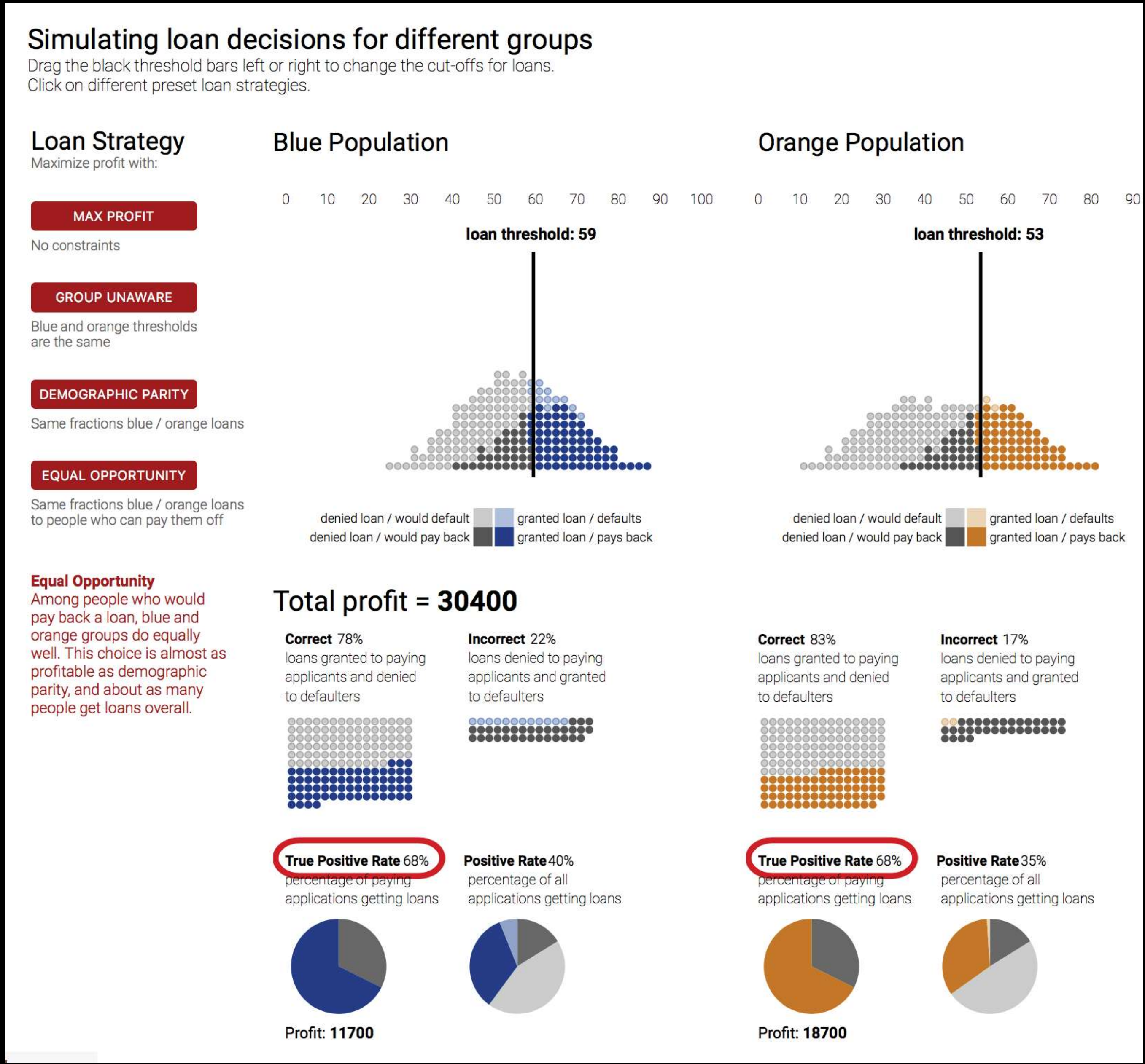
Equality of opportunity: only for the 'beneficial' outcome

$$\Pr\left\{\widehat{Y}=1 \mid A=0, Y=1\right\}=\Pr\left\{\widehat{Y}=1 \mid A=1, Y=1\right\}.$$

The equality of opportunity idea.

1. Moritz Hardt

$$\Pr\left\{\widehat{Y}=1 \mid A=0, Y=1\right\}=\Pr\left\{\widehat{Y}=1 \mid A=1, Y=1\right\}.$$



Responsible AI Practices (no date). *Google AI*. Available from https://ai.google/education/responsible-ai-practices/ [Accessed 7 June 2018].

source: Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2017. "Fairness in Criminal Justice Risk Assessments: The State of the Art." *ArXiv: 1703.09207 [Stat]*, March. http://arxiv.org/abs/1703.09207.

Y^ Y^

Y=
Y=

	Failure Predicted	Success Predicted	Conditional Procedure Error
Failure – A Positive	a True Positives	b False Negatives	$b/(a+b)$ False Negative Rate
Success – A Negative	c False Positives	d True Negatives	$c/(c+d)$ False Positive Rate
Conditional Use Error	$c/(a+c)$ Failure Prediction Error	$b/(b+d)$ Success Prediction Error	$\frac{(c+b)}{(a+b+c+d)}$ Overall Procedure Error

Conditional Procedure Error – The proportion of cases incorrectly classified conditional on one of the two *actual* outcomes: $b/(a+b)$, which is the *false negative rate*, and $c/(c+d)$, which is the *false positive rate*.

Conditional procedure equality: (Pro-publica fairness)

E.g. a/(a+b) and d/(c + d) is the same for men and women.

The equality of opportunity idea.

1. Moritz Hardt

$$\Pr\left\{\widehat{Y} = 1 \mid A = 0, Y = 1\right\} = \Pr\left\{\widehat{Y} = 1 \mid A = 1, Y = 1\right\}.$$

Q: who are the people whose opportunities are equal?

4. Fairness trade-offs

4. Fairness trade-offs

The COMPAS/Pro-publica case

2016

Bernard Parker, left, was rated high risk; Dylan Fugett 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

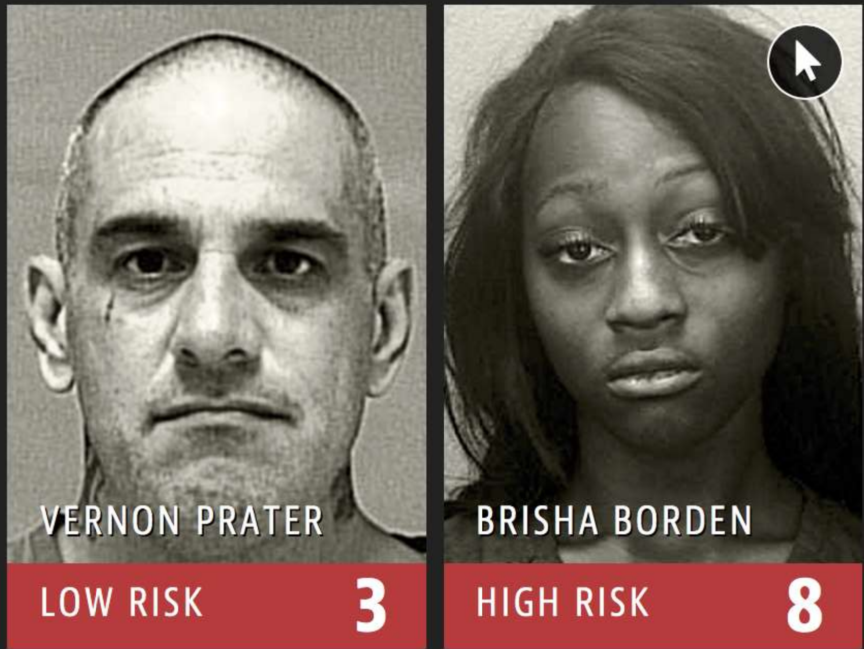
May 23, 2016

Our analysis of Northpointe's tool, called COMPAS (which stands for Correctional Offender Management Profiling for Alternative Sanctions), found that black defendants were far more likely than white defendants to be incorrectly judged to be at a higher risk of recidivism, while white defendants were more likely than black defendants to be incorrectly flagged as low risk.

recidivism:

For most of our analysis, we defined recidivism as a new arrest within two years.

Two Petty Theft Arrests



Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

Pro-publica's claims:

We looked at more than 10,000 criminal defendants in Broward County, Florida, and compared their predicted recidivism rates with the rate that actually occurred over a two-year period.

- Black defendants were often predicted to be at a higher risk of recidivism than they actually were. Our analysis found that black defendants who did not recidivate over a two-year period were nearly twice as likely to be misclassified as higher risk compared to their white counterparts (45 percent vs. 23 percent).
- White defendants were often predicted to be less risky than they were. Our analysis found that white defendants who re-offended within the next two years were mistakenly labeled low risk almost twice as often as black re-offenders (48 percent vs. 28 percent).

Black defendants were also twice as likely as white defendants to be misclassified as being a higher risk of violent recidivism. And white violent recidivists were 63 percent more likely to have been misclassified as a low risk of violent recidivism, compared with black violent recidivists.

two uses of COMPAS scores:

bail decisions:
should the person be released from prison before her trial?

→ provides COMPAS scores of individuals classified as 'high risk' that may not have been put in jail

parole decisions:
should the person be released from jail before the completion of the maximum jail sentence?

→ this is the practice where risk scores are used, judged to be unfairly discriminatory

Pro-publica: obtaining information about the 'false positives' (high risk labels who do not reoffend)

Through a public records request, ProPublica obtained two years worth of COMPAS scores from the Broward County Sheriff's Office in Florida. We received data for all 18,610 people who were scored in 2013 and 2014.

Starting with the database of COMPAS scores, we built a profile of each person's criminal history, both before and after they were scored.

We removed people from our data set for whom we had less than two years of recidivism information.

We removed people from the risk set while they were incarcerated.

We marked scores other than "low" as higher risk.

Contingency tables

All Defendants	Low	High
Survived	2681	1282
Recidivated	1216	2035

FP rate: 32.35

FN rate: 37.40

PPV: 0.61

NPV: 0.69

LR+: 1.94

LR-: 0.55

Contingency tables

All Defendants	Low	High
Survived	2681	1282
Recidivated	1216	2035

FP rate: 32.35

FN rate: 37.40

PPV: 0.61

NPV: 0.69

LR+: 1.94

LR-: 0.55

Black Defendants	Low	High
Survived	990	805
Recidivated	532	1369

FP rate: 44.85

FN rate: 27.99

PPV: 0.63

NPV: 0.65

LR+: 1.61

LR-: 0.51

White Defendants	Low	High
Survived	1139	349
Recidivated	461	505

FP rate: 23.45

FN rate: 47.72

PPV: 0.59

NPV: 0.71

LR+: 2.23

LR-: 0.62

source: Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2017. "Fairness in Criminal Justice Risk Assessments: The State of the Art." ***ArXiv: 1703.09207 [Stat]*****, March. http://arxiv.org/abs/1703.09207.**

	Failure Predicted	Success Predicted	Conditional Procedure Error
Failure – A Positive	a True Positives	b False Negatives	$b/(a+b)$ False Negative Rate
Success – A Negative	c False Positives	d True Negatives	$c/(c+d)$ False Positive Rate
Conditional Use Error	$c/(a+c)$ Failure Prediction Error	$b/(b+d)$ Success Prediction Error	$\frac{(c+b)}{(a+b+c+d)}$ Overall Procedure Error

N.B.

Failure of parole = the risk you try to prevent

E.g. 'positive' = arrest for violent crime

Pro-publica claims:

We looked at more than 10,000 criminal defendants in Broward County, Florida, and compared their predicted recidivism rates with the rate that actually occurred over a two-year period.

- Black defendants were often predicted to be at a higher risk of recidivism than they actually were. Our analysis found that black defendants who did not recidivate over a two-year period were nearly twice as likely to be misclassified as higher risk compared to their white counterparts (45 percent vs. 23 percent).
- White defendants were often predicted to be less risky than they were. Our analysis found that white defendants who re-offended within the next two years were mistakenly labeled low risk almost twice as often as black re-offenders (48 percent vs. 28 percent).

Black defendants were also twice as likely as white defendants to be misclassified as being a higher risk of violent recidivism. And white violent recidivists were 63 percent more likely to have been misclassified as a low risk of violent recidivism, compared with black violent recidivists.

source: Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2017. “Fairness in Criminal Justice Risk Assessments: The State of the Art.” *ArXiv: 1703.09207 [Stat]*, March. http://arxiv.org/abs/1703.09207.

		Failure Predicted	Success Predicted	Conditional Procedure Error
Y	Failure – A Positive	a True Positives	b False Negatives	$b/(a+b)$ False Negative Rate
Y	Success – A Negative	c False Positives	d True Negatives	$c/(c+d)$ False Positive Rate
	Conditional Use Error	$c/(a+c)$ Failure Prediction Error	$b/(b+d)$ Success Prediction Error	$\frac{(c+b)}{(a+b+c+d)}$ Overall Procedure Error

Conditional Procedure Error – The proportion of cases incorrectly classified conditional on one of the two *actual* outcomes: $b/(a+b)$, which is the *false negative rate*, and $c/(c+d)$, which is the *false positive rate*.

Pro-publica fairness

source: Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2017. "Fairness in Criminal Justice Risk Assessments: The State of the Art." ***ArXiv: 1703.09207 [Stat]*****, March. http://arxiv.org/abs/1703.09207.**

	Failure Predicted	Success Predicted	Conditional Procedure Error
Failure – A Positive	a True Positives	b False Negatives	$b/(a+b)$ False Negative Rate
Success – A Negative	c False Positives	d True Negatives	$c/(c+d)$ False Positive Rate
Conditional Use Error	$c/(a+c)$ Failure Prediction Error	$b/(b+d)$ Success Prediction Error	$\frac{(c+b)}{(a+b+c+d)}$ Overall Procedure Error

Conditional Procedure Error – The proportion of cases incorrectly classified conditional on one of the two *actual* outcomes: $b/(a+b)$, which is the *false negative rate*, and $c/(c+d)$, which is the *false positive rate*.

'Pro-publica' fairness =

Conditional procedure equality:

E.g. a/(a+b) and d/(c + d) is the same for white and black prisoners.

source: Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2017. "Fairness in Criminal Justice Risk Assessments: The State of the Art." ***ArXiv: 1703.09207 [Stat]*****, March. http://arxiv.org/abs/1703.09207.**

	Failure Predicted	Success Predicted	Conditional Procedure Error
Failure – A Positive	a True Positives	b False Negatives	$b/(a+b)$ False Negative Rate
Success – A Negative	c False Positives	d True Negatives	$c/(c+d)$ False Positive Rate
Conditional Use Error	$c/(a+c)$ Failure Prediction Error	$b/(b+d)$ Success Prediction Error	$\frac{(c+b)}{(a+b+c+d)}$ Overall Procedure Error

Conditional Use Error – The proportion of cases incorrectly predicted conditional on one of the two *predicted* outcomes: $c/(a+c)$, which is the proportion of incorrect failure predictions, and $b/(b+d)$, which is the proportion of incorrect success predictions.[5]

source: Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2017. "Fairness in Criminal Justice Risk Assessments: The State of the Art." *ArXiv: 1703.09207 [Stat]*, March. http://arxiv.org/abs/1703.09207.

	Failure Predicted	Success Predicted	Conditional Procedure Error
Failure – A Positive	a True Positives	b False Negatives	$b/(a+b)$ False Negative Rate
Success – A Negative	c False Positives	d True Negatives	$c/(c+d)$ False Positive Rate
Conditional Use Error	$c/(a+c)$ Failure Prediction Error	$b/(b+d)$ Success Prediction Error	$\frac{(c+b)}{(a+b+c+d)}$ Overall Procedure Error

Conditional Use Error – The proportion of cases incorrectly predicted conditional on one of the two *predicted* outcomes: $c/(a+c)$, which is the proportion of incorrect failure predictions, and $b/(b+d)$, which is the proportion of incorrect success predictions.[5]

Conditional use accuracy equality is achieved by $\hat{f}(L, S)$ when conditional use accuracy is the same for both protected group categories (Berk., 2016b). One is conditioning on the algorithm's *predicted* outcome not the actual outcome. That is, $a/(a+c)$ is the same for men and women, and $d/(b+d)$ is the same for men and women.

source: Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2017. "Fairness in Criminal Justice Risk Assessments: The State of the Art." *ArXiv: 1703.09207 [Stat]*, March. http://arxiv.org/abs/1703.09207.

	Failure Predicted	Success Predicted	Conditional Procedure Error
Failure – A Positive	a True Positives	b False Negatives	$b/(a+b)$ False Negative Rate
Success – A Negative	c False Positives	d True Negatives	$c/(c+d)$ False Positive Rate
Conditional Use Error	$c/(a+c)$ Failure Prediction Error	$b/(b+d)$ Success Prediction Error	$\frac{(c+b)}{(a+b+c+d)}$ Overall Procedure Error

Should a/(a+c) and d/(b+d) be the same for the white and black population?

Definition 6 (Predictive Value Parity) *A predictive model h satisfies predictive value parity if* $\forall \mathbf{z}, \mathbf{z}' \in \mathcal{Z}, \forall y, \hat{y} \in \mathcal{Y}$:

$$\mathbb{P}_{(\mathbf{X},Y)\sim F}[Y = y|\mathbf{Z} = \mathbf{z}, \hat{Y} = \hat{y}] = \mathbb{P}_{(\mathbf{X},Y)\sim F}[Y = y|\mathbf{Z} = \mathbf{z}', \hat{Y} = \hat{y}].$$

COMPAS' POSSIBLE LINE OF DEFENCE

In this paper we show that the differences in false positive and false negative rates cited as evidence of racial bias in the ProPublica article are a direct consequence of applying an instrument that is free from predictive bias[1] to a population in which recidivism prevalence differs across groups.

Source: Chouldechova, Alexandra. 2016. "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments." *ArXiv: 1610.07524 [Cs, Stat]*, October. http://arxiv.org/abs/1610.07524.

source: Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2017. “Fairness in Criminal Justice Risk Assessments: The State of the Art.” *ArXiv: 1703.09207 [Stat]*, March. http://arxiv.org/abs/1703.09207.

	Failure Predicted	Success Predicted	Conditional Procedure Error
Failure – A Positive	a True Positives	b False Negatives	$b/(a+b)$ False Negative Rate
Success – A Negative	c False Positives	d True Negatives	$c/(c+d)$ False Positive Rate
Conditional Use Error	$c/(a+c)$ Failure Prediction Error	$b/(b+d)$ Success Prediction Error	$\frac{(c+b)}{(a+b+c+d)}$ Overall Procedure Error

COMPAS’ fairness ->

Conditional use accuracy equality is achieved by $\hat{f}(L, S)$ when conditional use accuracy is the same for both protected group categories (Berk., 2016b). One is conditioning on the algorithm’s *predicted* outcome not the actual outcome. That is, $a/(a+c)$ is the same for men and women, and $d/(b+d)$ is the same for men and women.

Problem

except in degenerate cases, you cannot have both forms of equality

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. 2016. "Inherent Trade-Offs in the Fair Determination of Risk Scores." *ArXiv: 1609.05807 [Cs, Stat]*, September. http://arxiv.org/abs/1609.05807.

Chouldechova, Alexandra. 2016. "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments." *ArXiv: 1610.07524 [Cs, Stat]*, October. http://arxiv.org/abs/1610.07524.

Problem

except in degenerate cases, you cannot have both forms of equality

Degenerate cases:

- *Perfect prediction*. Suppose that for each feature vector σ, we have either $p_\sigma = 0$ or $p_\sigma = 1$. This means that we can achieve perfect prediction, since we know each person's class label (positive or negative) for certain. In this case, we can assign all feature vectors σ with $p_\sigma = 0$ to a bin b with score $v_b = 0$, and all σ with $p_\sigma = 1$ to a bin b' with score $v_{b'} = 1$. It is easy to check that all three of the conditions (A), (B), and (C) are satisfied by this risk assignment.
- *Equal base rates*. Suppose, alternately, that the two groups have the same fraction of members in the positive class; that is, the average value of p_σ is the same for the members of group 1 and group 2. (We can refer to this as the *base rate* of the group with respect to the classification problem.) In this case, we can create a single bin b with score equal to this average value of p_σ, and we can assign everyone to bin b. While this is not a particularly informative risk assignment, it is again easy to check that it satisfies fairness conditions (A), (B), and (C).

Source: Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. 2016. "Inherent Trade-Offs in the Fair Determination of Risk Scores." *ArXiv: 1609.05807 [Cs, Stat]*, September. http://arxiv.org/abs/1609.05807.

For risk scores:

(B) *Balance for the negative class* requires that the average score assigned to people of group 1 who belong to the negative class should be the same as the average score assigned to people of group 2 who belong to the negative class. In other words, the assignment of scores shouldn't be systematically more inaccurate for negative instances in one group than the other.

(C) *Balance for the positive class* symmetrically requires that the average score assigned to people of group 1 who belong to the positive class should be the same as the average score assigned to people of group 2 who belong to the positive class.

VS.

Definition 2.1 (Test fairness)**.** A score $S = S(x)$ is *test-fair* (well-calibrated)[2] if it reflects the same likelihood of recidivism irrespective of the individual's group membership, R. That is, if for all values of s,

$$\mathbb{P}(Y = 1 \mid S = s, R = b) = \mathbb{P}(Y = 1 \mid S = s, R = w). \quad (2.1)$$

CALIBRATION

Sources:

1. Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. 2016. "Inherent Trade-Offs in the Fair Determination of Risk Scores." *ArXiv: 1609.05807 [Cs, Stat]*, September. http://arxiv.org/abs/1609.05807.
2. Chouldechova, op. cit.

5. Fair predictions and economics

Is ‘accurate prediction’ an end in itself?

The cost of fairness

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of KDD '17, Halifax, NS, Canada, August 13-17, 2017*, 10 pages.
DOI: 10.1145/3097983.3098095

Definition 3.1 (Immediate utility). For c a constant such that $0 < c < 1$, the immediate utility of a decision rule d is

$$u(d, c) = \mathbb{E}\left[Yd(X) - cd(X)\right] = \mathbb{E}\left[Yd(X)\right] - c\mathbb{E}\left[d(X)\right]. \quad (5)$$

benefit

proportional to violent crimes prevented

cost

proportional to n. people detained

Corbett-Davies asks what maximises immediate utility for release decisions, comparing optimisation with and without parity constraints.

The unconstrained algorithm uses a single threshold and achieves a higher utility than constrained (i.e. fair) ones.

$$p_{Y|X} > c,$$

Corbett-Davies' unconstrained model optimizes immediate utility for release decisions, comparing optimisation with and without parity constraints.

The unconstrained algorithm uses a single threshold and achieves a higher utility than constrained ones.

$$p_{Y|X} > c,$$

N.B. 'optimizing' here DOES NOT mean achieving the highest accuracy

Definition 3.1 (Immediate utility). For c a constant such that $0 < c < 1$, the immediate utility of a decision rule d is

$$\begin{aligned} u(d, c) &= \mathbb{E}\left[Yd(X) - cd(X)\right] \\ &= \mathbb{E}\left[Yd(X)\right] - c\mathbb{E}\left[d(X)\right]. \end{aligned} \quad (5)$$

6. Equality of opportunity *theory*

A Moral Framework for Understanding of Fair ML through Economic Models of Equality of Opportunity

Hoda Heidari
ETH Zürich
hheidari@inf.ethz.ch

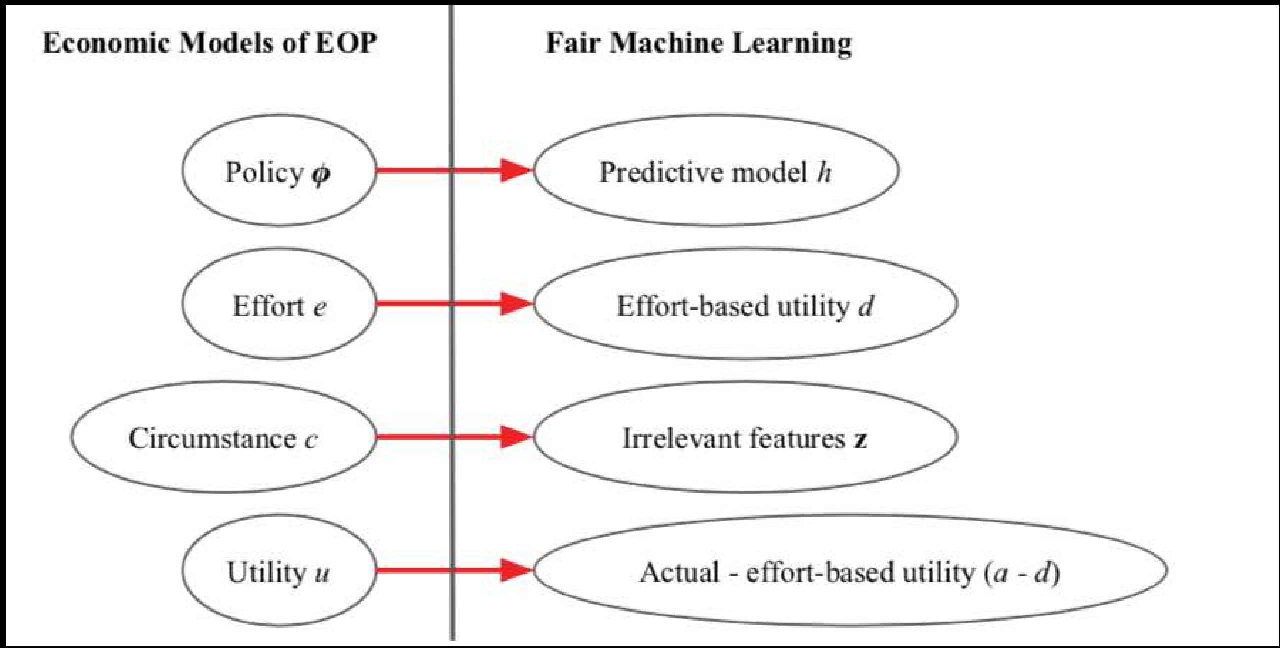
Michele Loi
University of Zürich
michele.loi@uzh.ch

Krishna P. Gummadi
MPI-SWS
gummadi@mpi-sws.org

Andreas Krause
ETH Zürich
krausea@ethz.ch

Fair distribution of the (dis) advantages of statistical prediction

A fair predictor distributes (advantage) utility fairly to individuals subject to decision making.



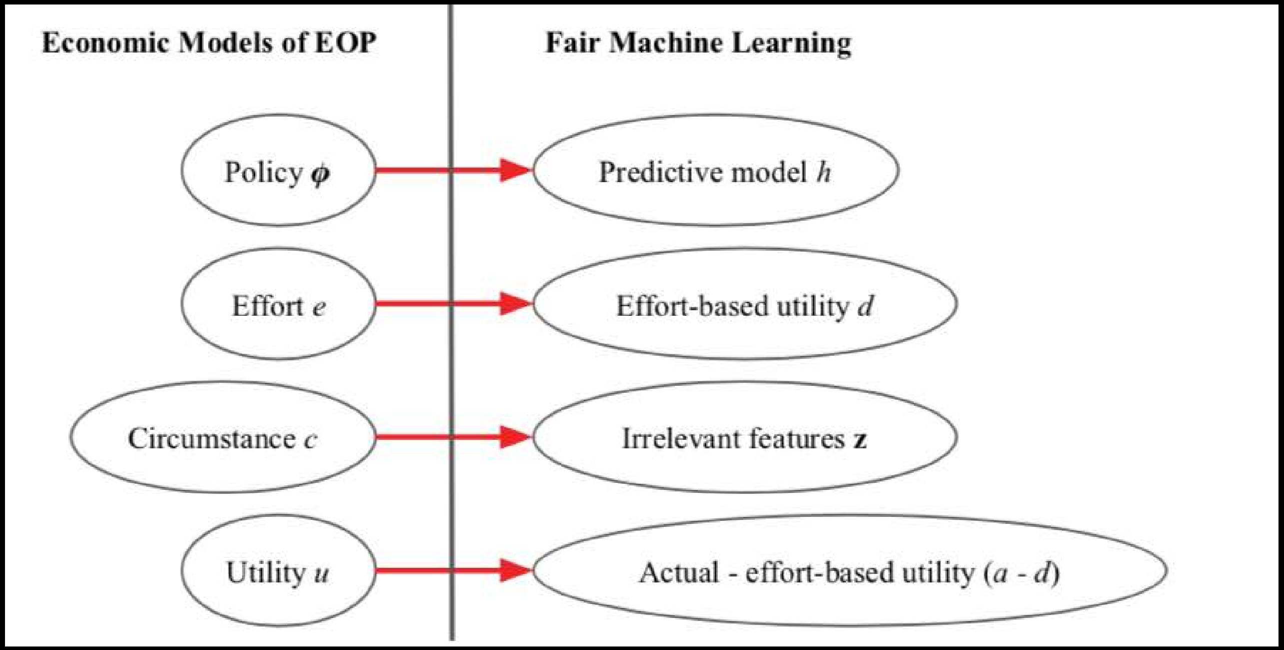
Actual utility = utility as a result of the decision, following the prediction

Effort-based utility = utility that corresponds to effort

(Advantage) utility = actual - effort-based utility

Fair distribution of the (dis) advantages of statistical prediction

A fair predictor distributes (advantage) utility fairly to individuals subject to decision making.



Actual utility = utility as a result of the decision, following the prediction

Effort-based utility = utility that corresponds to effort

(Advantage) utility = actual - effort-based utility ⟷ **distribuendum**

$$d = g(\mathbf{x}, y, h).$$

predictive model

$h : \mathcal{X} \to \mathcal{Y}$

actual value, label

effort-based (‘deserved’) utility

observable attributes

$\mathbf{x} = \langle \mathbf{z}, \mathbf{w} \rangle,$

not accountable

accountable

$$a = f(\mathbf{x}, y, h)$$

actual utility

Rawlsian equality of opportunity

Definition 1 (Rawlsian Equality of Opportunity (R-EOP)) *A policy ϕ satisfies Rawlsian EOP if for all circumstances c, c' and all effort levels e,*

$$F^{\phi}(.|c,e) = F^{\phi}(.|c',e).$$

cumulative distribution of utility under policy ϕ at a fixed effort level e and circumstance c

$$F^{h}(.|\mathbf{Z}=\mathbf{z}, D=d) = F^{h}(.|\mathbf{Z}=\mathbf{z}', D=d).$$

for predictions

Let Fh(.) specify the distribution of utility across individuals under predictive model h.

Rawlsian equality of opportunity

Definition 1 (Rawlsian Equality of Opportunity (R-EOP)) *A policy ϕ satisfies Rawlsian EOP if for all circumstances c, c' and all effort levels e,*

$$F^{\phi}(.|c, e) = F^{\phi}(.|c', e).$$

cumulative distribution of utility under policy ϕ at a fixed effort level e and circumstance c

$$d = g(\mathbf{x}, y, h).$$

$$F^{h}(.|\mathbf{Z} = \mathbf{z}, D = d) = F^{h}(.|\mathbf{Z} = \mathbf{z}', D = d).$$

for predictions

$$d = g(\mathbf{x}, y, \cancel{h}).$$

Let Fh(.) specify the distribution of utility across individuals under predictive model h.

Note that this conception of EOP takes an *absolutist* view of effort: it assumes e is a scalar whose absolute value is meaningful and can be compared across individuals. This view requires effort e to be inherent to individuals and not itself impacted by the circumstance c or the policy ϕ.

Luck-egalitarian

let $F_E^{c,\phi}$ be the effort distribution of type c under policy ϕ.

Definition 2 (Luck Egalitarian Equality of Opportunity (e-EOP)) *A policy ϕ satisfies Luck Egalitarian EOP if for all $\pi \in [0,1]$ and any two circumstances c, c':*

$$F^{\phi}(.|c,\pi) = F^{\phi}(.|c',\pi).$$

$F^{\phi}(.|c,\pi)$ specify the distribution of utility for individuals of type c at the πth quantile $(0 \leq \pi \leq 1)$ of $F_E^{c,\phi}$.

we have shown that:

Some existing fairness conceptions correspond to different 'interpretations' of EoP

Notion of fairness	Effort-based utility D	Actual utility A	Notion of EOP
Accuracy Parity	constant (e.g. 0)	$(\hat{Y} - Y)^2$	Rawlsian
Statistical Parity	constant (e.g. 1)	$\hat{Y}$	Rawlsian
Equality of Odds	Y	$\hat{Y}$	Rawlsian
Predictive Value Parity	$\hat{Y}$	Y	egalitarian

Table 1: Interpretation of existing notions of algorithmic fairness for binary classification as special instances of EOP.

E.g. equality of odds

remember?

	Failure Predicted	Success Predicted	Conditional Procedure Error
Failure – A Positive	a True Positives	b False Negatives	$b/(a+b)$ False Negative Rate
Success – A Negative	c False Positives	d True Negatives	$c/(c+d)$ False Positive Rate
Conditional Use Error	$c/(a+c)$ Failure Prediction Error	$b/(b+d)$ Success Prediction Error	$\frac{(c+b)}{(a+b+c+d)}$ Overall Procedure Error

Pro-publica fairness

Conditional Procedure Error – The proportion of cases incorrectly classified conditional on one of the two *actual* outcomes: $b/(a+b)$, which is the *false negative rate*, and $c/(c+d)$, which is the *false positive rate*.

Conditional procedure accuracy equality is achieved by $\hat{f}(L,S)$ when conditional procedure accuracy is the same for both protected group categories (Berk, 2016b). In our notation, $a/(a+b)$ is the same

Definition 4 (Equality of Odds) *A predictive model* h *satisfies equality of odds if* $\forall \mathbf{z}, \mathbf{z}' \in \mathcal{Z}, \forall y, \hat{y} \in \mathcal{Y}$:

$$\mathbb{P}_{(\mathbf{X},Y)\sim F}[\hat{Y} = \hat{y} | \mathbf{Z} = \mathbf{z}, Y = y] = \mathbb{P}_{(\mathbf{X},Y)\sim F}[\hat{Y} = \hat{y} | \mathbf{Z} = \mathbf{z}', Y = y].$$

e.g. P (Y^ = 1 | W, Y=0) = P (Y^= 1 | B, Y=0) & P (Y^ = 0 | W, Y=1) = P (Y^= 0 | B, Y=1)

Assume:

$A = h(\mathbf{X}) = \hat{Y}$ *(i.e., the actual utility is equal to the predicted label)*

$D = g(\mathbf{W}, Y)$ *where* $g(\mathbf{W}, Y) = Y$

i.e., effort-based utility of an individual is assumed to be the same as their true label

Rawlsian EoP is equivalent to equality of odds

This mode of analysis highlights a crucial moral assumption of equal odds in an EoP perspective

Equal odds assumes that all individuals with the same value of Y have the same D (effort-based utility)

→

Prisoners released on parole are equivalent in their effort-based utility

Is this always reasonable?

Luck-egalitarian

Definition 2 (Luck Egalitarian Equality of Opportunity (e-EOP)) *A policy ϕ satisfies Luck Egalitarian EOP if for all $\pi \in [0,1]$ and any two circumstances c, c':*

$$F^{\phi}(.|c,\pi) = F^{\phi}(.|c',\pi).$$

Definition 8 (e-EOP for supervised learning) *Suppose $d = f(\mathbf{x}, y, h)$. Predictive model h satisfies egalitarian EOP if for all $\pi \in [0,1]$ and $\mathbf{z}, \mathbf{z}' \in \mathcal{Z}$,*

$$F^{h}(.|\mathbf{Z} = \mathbf{z}, \Pi = \pi) = F^{h}(.|\mathbf{Z} = \mathbf{z}', \Pi = \pi).$$

source: Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2017. "Fairness in Criminal Justice Risk Assessments: The State of the Art." *ArXiv: 1703.09207 [Stat]*, March. http://arxiv.org/abs/1703.09207.

	Failure Predicted	Success Predicted	Conditional Procedure Error
Failure – A Positive	a True Positives	b False Negatives	$b/(a+b)$ False Negative Rate
Success – A Negative	c False Positives	d True Negatives	$c/(c+d)$ False Positive Rate
Conditional Use Error	$c/(a+c)$ Failure Prediction Error	$b/(b+d)$ Success Prediction Error	$\frac{(c+b)}{(a+b+c+d)}$ Overall Procedure Error

COMPAS fairness

Conditional Use Error – The proportion of cases incorrectly predicted conditional on one of the two *predicted* outcomes: $c/(a+c)$, which is the proportion of incorrect failure predictions, and $b/(b+d)$, which is the proportion of incorrect success predictions.[5]

Conditional use accuracy equality is achieved by $\hat{f}(L, S)$ when conditional use accuracy is the same for both protected group categories (Berk., 2016b). One is conditioning on the algorithm's *predicted* outcome not the actual outcome. That is, $a/(a+c)$ is the same for men and women, and $d/(b+d)$ is the same for men and women.

Predictive value parity

Definition 6 (Predictive Value Parity) *A predictive model h satisfies predictive value parity if* $\forall \mathbf{z}, \mathbf{z}' \in \mathcal{Z}, \forall y, \hat{y} \in \mathcal{Y}$:

$$\mathbb{P}_{(\mathbf{X},Y)\sim F}[Y = y | \mathbf{Z} = \mathbf{z}, \hat{Y} = \hat{y}] = \mathbb{P}_{(\mathbf{X},Y)\sim F}[Y = y | \mathbf{Z} = \mathbf{z}', \hat{Y} = \hat{y}].$$

e.g. P (Y = 1 | W, Y^=0) = P (Y= 1 | B, Y^=0) & P (Y = 0 | W, Y=1) = P (Y= 0 | B, Y=1)

Assumptions

$$A = Y$$

$$D = g(\mathbf{X}, Y, h)$$

$$g(\mathbf{X}, Y, h) = h(\mathbf{X}) = \hat{Y}$$

E.g.

Benefit = non reoffending

Accountability -> prediction !

(as calculated by the model used)

Contexts

- in which we consider people accountable for our predictions about them:
- in which the actual outcome (Y) is the most significant harm/benefit at stake

E.g. preventing possibly drunken drivers from driving, also for their own good

D = g(X,Y, h)=h(X) alcohol level (predictive of a car accident)

A = avoiding an accident

Luck-Egalitarian EoP is equivalent to predictive value parity

Some existing fairness conceptions correspond to different ‘interpretations’ of EoP

Notion of fairness	Effort-based utility D	Actual utility A	Notion of EOP
Accuracy Parity	constant (e.g. 0)	$(\hat{Y} - Y)^2$	Rawlsian
Statistical Parity	constant (e.g. 1)	$\hat{Y}$	Rawlsian
Equality of Odds	Y	$\hat{Y}$	Rawlsian
Predictive Value Parity	$\hat{Y}$	Y	egalitarian

Table 1: Interpretation of existing notions of algorithmic fairness for binary classification as special instances of EOP.

A new fairness metric

$$h^\pi \in \arg\max_{h \in \mathcal{H}} \min_{\mathbf{z} \in \mathcal{Z}} v^{\mathbf{z}}(\pi, h).$$

$$h^* \in \arg\max_{h \in \mathcal{H}} \min_{\mathbf{z} \in \mathcal{Z}} \int_0^1 v^{\mathbf{z}}(\pi, h) d\pi.$$

Roemer

$$\mathcal{F}(h, T) = \min_{\mathbf{z} \in \mathcal{Z}} \frac{1}{n_{\mathbf{z}}} \sum_{i \in T : \mathbf{z}_i = \mathbf{z}} u(\mathbf{x}_i, y_i, h)$$

Heidari et al

Example:

Y = "per capita number of violent crimes"

→ law enforcement resources

→ values of properties

→ attraction of investment

- For a majority-Caucasian neighborhood,

$$u(0, y, \hat{y}) = (1 + 0.5\hat{y}y) - (0.5\hat{y}).$$

- For a minority-Caucasian neighborhood,

$$u(1, y, \hat{y}) = (1 + 3\hat{y}y + 2\hat{y}) - (y).$$

Utility assumptions

Conclusions:

Determining accountability features and effort-based utility is arguably outside the expertise of computer scientists, and has to be resolved through the appropriate process with input from stakeholders and domain experts.

In any given application domain, reasonable people may disagree on what constitutes factors that peo- ple should be considered morally accountable for, and there will rarely be a consensus on the most suitable notion of fairness.

This, however, does not imply that in a given context all existing notions of algorithmic fairness are equally acceptable from a moral standpoint.