AI Ethics, Impossibility Theorems and Tradeoffs

Chris Stucchio
Director of Data Science, Simpl
https://chrisstucchio.com
@stucchio

Simplest example

Supermarket theft prevention algorithm:

1. Make a spreadsheet of item SKU, shrinkage (theft) rate and price
2. Sort list by shrinkage*price.
3. Put anti-theft devices on the SKUs with the highest rates of shrinkage.

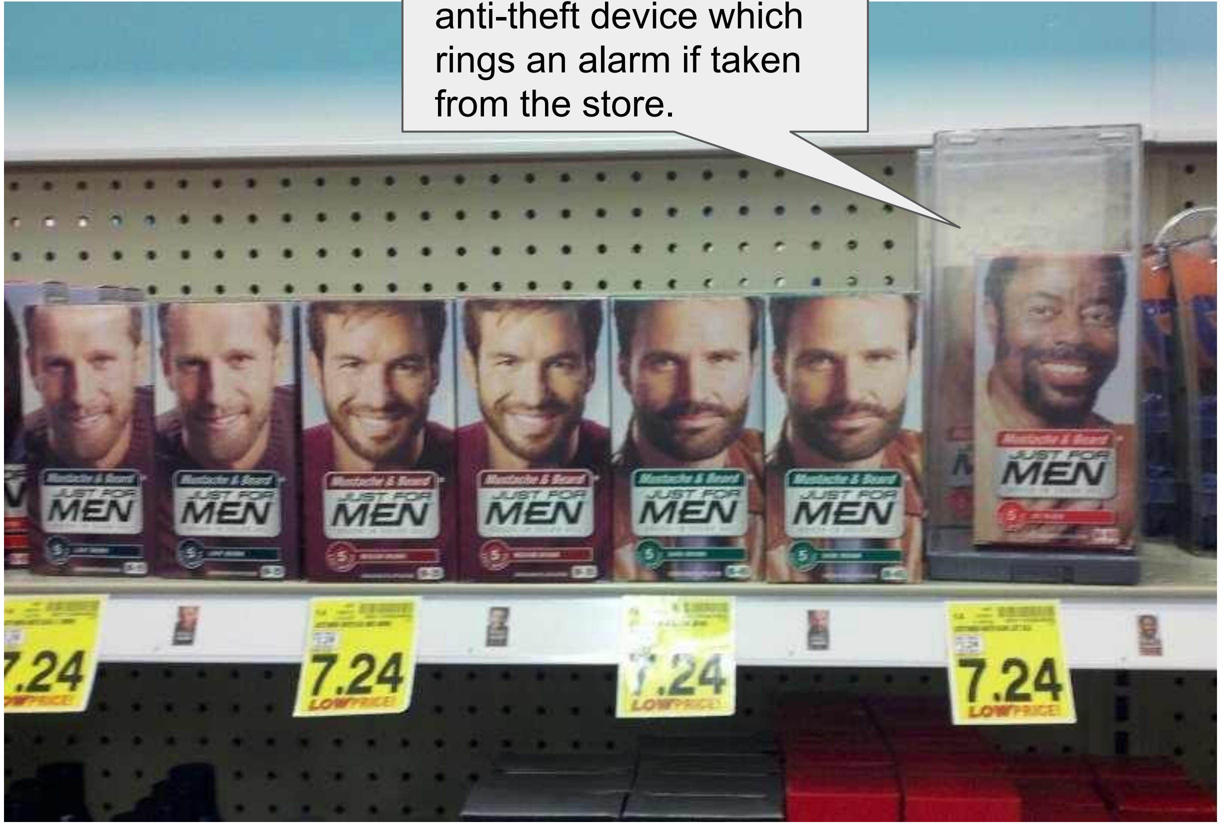
sku	shrinkage	price	=b*c
abc123	0.17	$7.24	1.23
def456	0.06	$12.53	0.752
ghi789	0.08	$8.29	0.66
jkl012	0.09	$4.50	0.40
mno234	0.16	$0.99	0.16

Simplest example

Supermarket theft prevention algorithm:

1. Make a spreadsheet of item SKU, shrinkage (theft) rate and price
2. Sort list by shrinkage*price.
3. Put anti-theft devices on the SKUs with the highest rates of shrinkage.

Whoops!



The plastic box is an anti-theft device which rings an alarm if taken from the store.

Simplest example

Why this is bad

- Likely makes black customers feel offended.
- The inconvenience of a slower checkout has a disparate impact (i.e. black customers, who are mostly not stealing, face the inconvenience more)
- Perpetuates racist stereotypes (which the data suggests have an element of truth).

Why this is good

- Reducing theft lowers prices for all customers.
- Without effective anti-theft measures, shops may stop carrying frequently stolen products.
- Resources (anti-theft devices, checkout time) are limited and must be allocated wisely.
- Better to inconvenience 10% of customers than 100%.

Fundamental conflict in AI Ethics

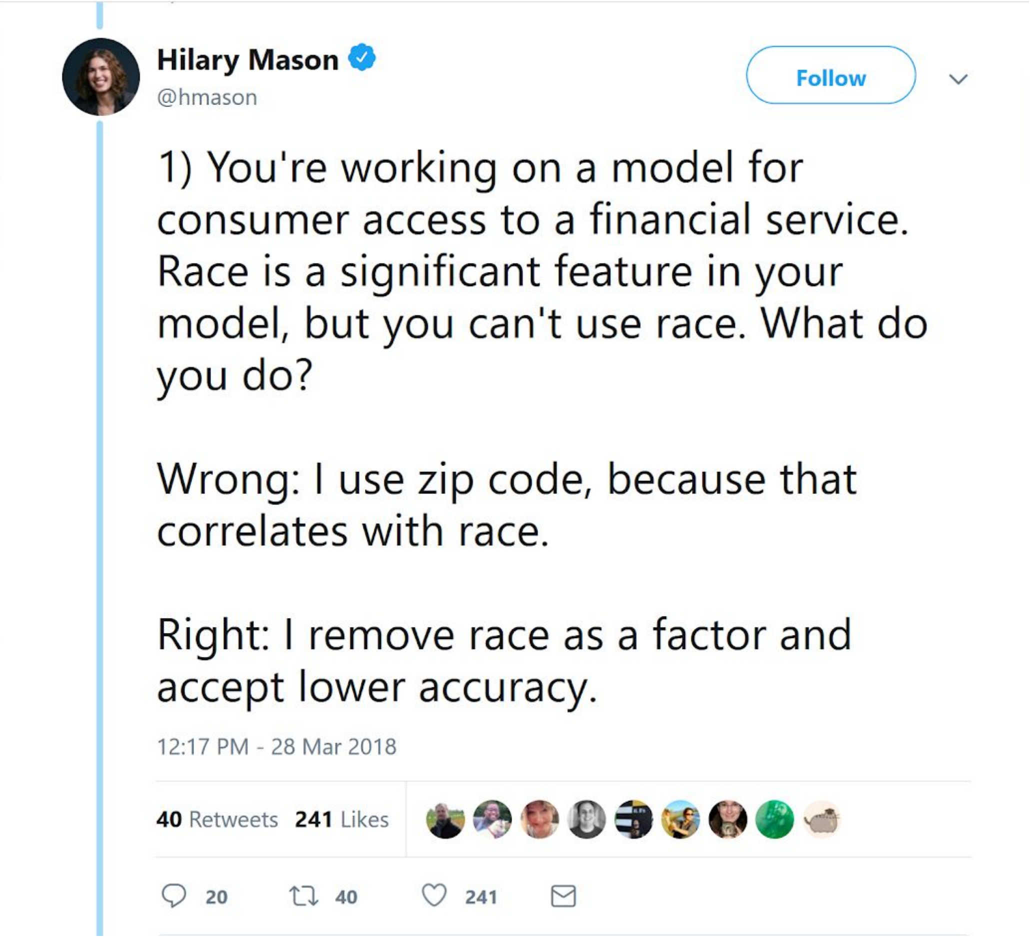
This talk is NOT about…

Cheerleeding

Lots of people in Silicon Valley think there’s a single clear answer.

Many talks are little more than telling the audience this single clear answer.

This talk takes **no ethical position** - it just tells you which ethical positions you **cannot simultaneously take**.



Hilary Mason
@hmason

1) You're working on a model for consumer access to a financial service. Race is a significant feature in your model, but you can't use race. What do you do?

Wrong: I use zip code, because that correlates with race.

Right: I remove race as a factor and accept lower accuracy.

12:17 PM - 28 Mar 2018

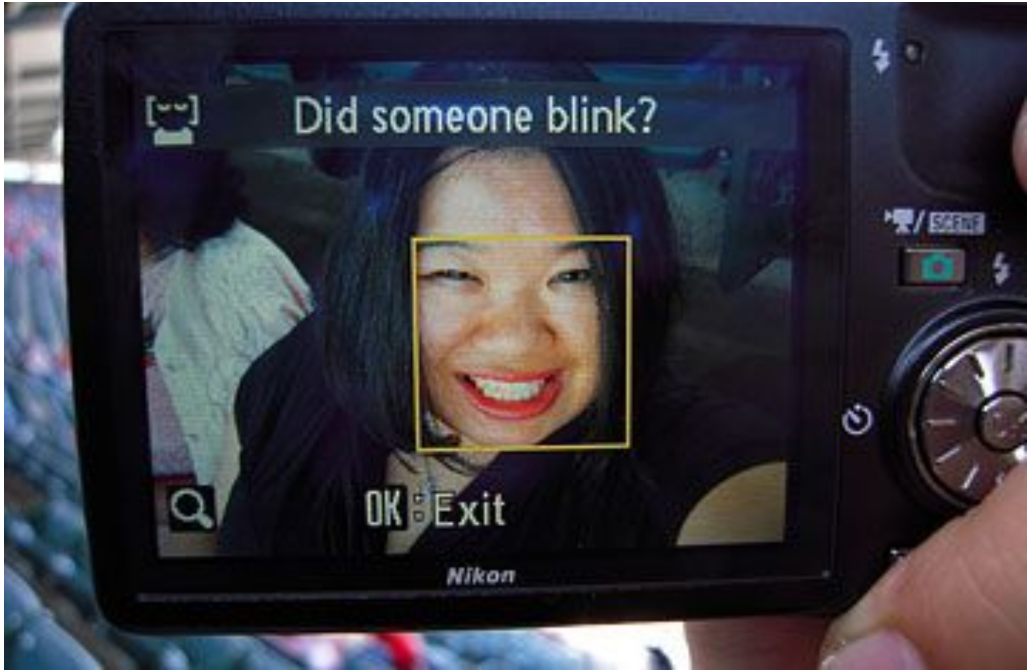
40 Retweets 241 Likes

Errors

No algorithm is 100% accurate.

If you can improve accuracy, you should. **There is no ethical question here** - only a hard problem in image processing.

Fixing these problems = making more money.



"Racist Camera! No, I did not blink... I'm just Asian!" - jozjozjoz

Europe

I'm an American who lives/works in India.

My knowledge of Europe:

- GDPR is an incoherent and underspecified mess.
- You can force people to forget true facts about you.
- Too many regulatory regimes.
- Delicious cheese.

Sorry!

tasteatlas
EUROPEAN
STREET FOOD MAP
www.tasteatlas.com/europe
PORILAINEN
TUNNBRÖDSRULLE
PIRUKAS
PIROZHKI
PØLSER
BREAKFAST ROLL
FISH AND CHIPS
KAPSALON
CURRYWURST
ZAPIEKANKA
PEREPICHKA
PÁREK V ROHLÍKU
JAMBON-BEURRE
LÁNGOS
COVRIGI
FARINATA DI CECI
PAN-BAGNAT
BUREK
PIADINA ROMAGNOLA
BOCADILLO
BIFANA
GYROS
DÖNER KEBAB
PANELLE
TANTUNI

Everything I know about Europe

Artificial Intelligence

This talk is about **decision theory**.

Every ethical quandary I discuss applies to humans as well as machines.

Only benefit of human decision processes: **easy obfuscation**.

You can cheaply and easily run an algorithm on test data to measure an effect. You can't do the same on, e.g., a judge or loan officer.

Classical Ethical Theories

Utilitarianism

It is bad to be murdered, raped, or sent to jail.

Utilitarianism tries to minimize the bad things in the world while maximizing the good things. In math terms, find the policy which minimizes:

$$\text{Harm(policy)} = A \times (\text{\# of murders}) + B \times (\text{years people are stuck in jail}) + \ldots$$

A:B is a conversion between murders and jail time. We are indifferent between jailing someone for B years and preventing A murders.

Procedural fairness

A classical belief is that decisions should be blind to certain individual **traits** (t in this case):

```
∀t1 ∀t2 f(x, t1) == f(x, t2)
```

Intuitively: Me (a foreigner), my Brahmin wife, our non-Brahmin maid or Prime Minister Modi should get the same justice given the same facts.

San Francisco Ethical Theories

Epistemic note: I am attempting to mathematically state premises, but the proponents of those premises often prefer for them to be kept informal:

> “As engineers, we’re trained to pay attention to the details, think logically, challenge assumptions that may be incorrect (or just fuzzy), and so on. These are all excellent tools for technical discussions. But they can be terrible tools for discussion around race, discrimination, justice...because questioning the exact details can easily be perceived as questioning the overall validity of the effort, or the veracity of the historical context.”
>
> - Urs Hölzle, S.V.P. at Google

Allocative Fairness

Important concept is **protected class**. What are these?

- In US: Blacks/Hispanics. Asians are a de-jure protected class, but de-facto not. Women, sometimes homosexuals.
- In India: Scheduled Castes and OBCs. Muslims/other religious minorities only in Tamil Nadu and Kerala.

Allocative fairness is when a certain statistic is equal across protected classes.

(Some variants choose favored classes and replace “=” with “<=”. Indian college admissions have favored castes, Americans have favored races.)

Allocative fairness, base rates and group boundaries

Example:

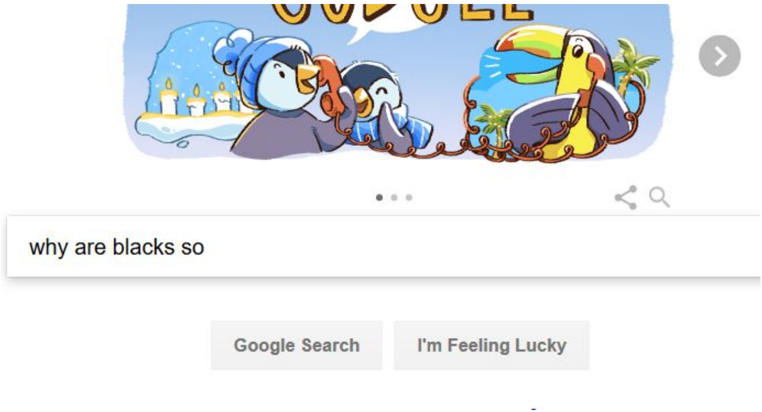
- 25% of both Scotsmen and Englishmen get 1200 on SAT.
- Cutoff for getting into college is 1200 on SAT.
- **No allocative harm.**

No True Scotsman will score < 1200 on their SAT

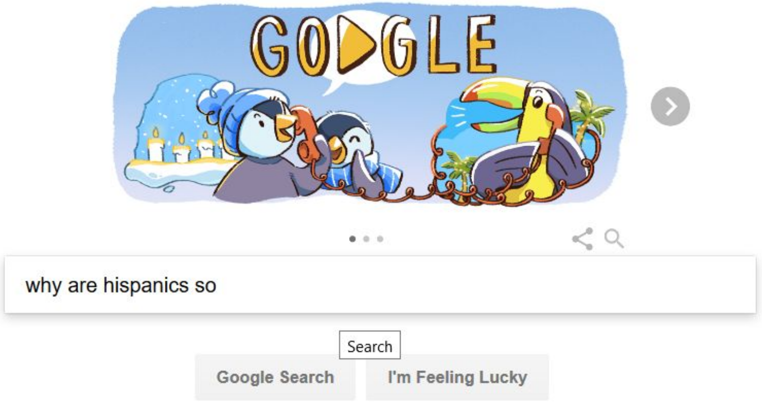
- Suddenly 100% of True Scotsmen get into college vs 25% of Englishmen and 0% of False Scotsmen.
- **Allocative harm is created!**

Representational Fairness/Honor Culture

San Francisco Google notices nothing

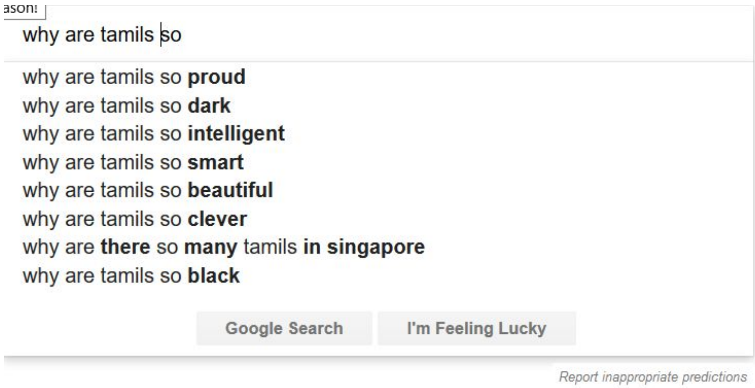


why are blacks so



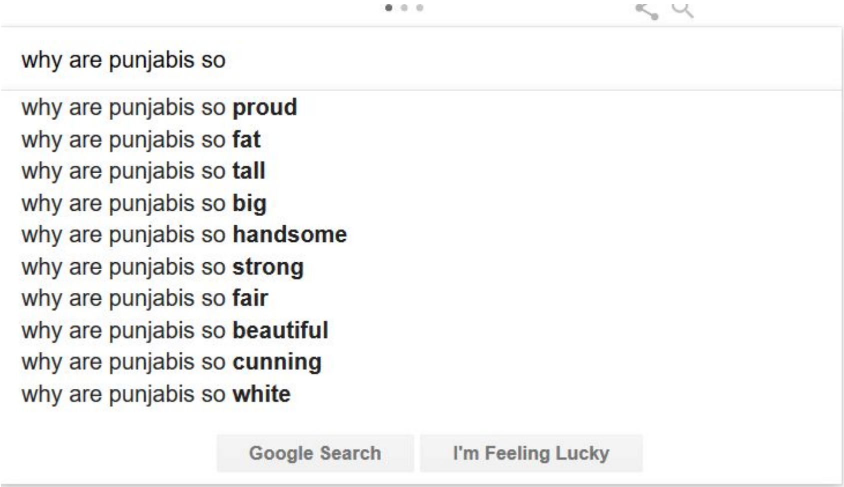
why are hispanics so

Indian Google notices everything



why are tamils so

why are tamils so **proud**
why are tamils so **dark**
why are tamils so **intelligent**
why are tamils so **smart**
why are tamils so **beautiful**
why are tamils so **clever**
why are **there** so **many** tamils **in singapore**
why are tamils so **black**



why are punjabis so

why are punjabis so **proud**
why are punjabis so **fat**
why are punjabis so **tall**
why are punjabis so **big**
why are punjabis so **handsome**
why are punjabis so **strong**
why are punjabis so **fair**
why are punjabis so **beautiful**
why are punjabis so **cunning**
why are punjabis so **white**

AI may notice things we don’t want it to

“Bias should be the expected result whenever even an unbiased algorithm is used to derive regularities from any data; bias is the regularities discovered.”

Semantics derived automatically from language corpora necessarily contain human biases

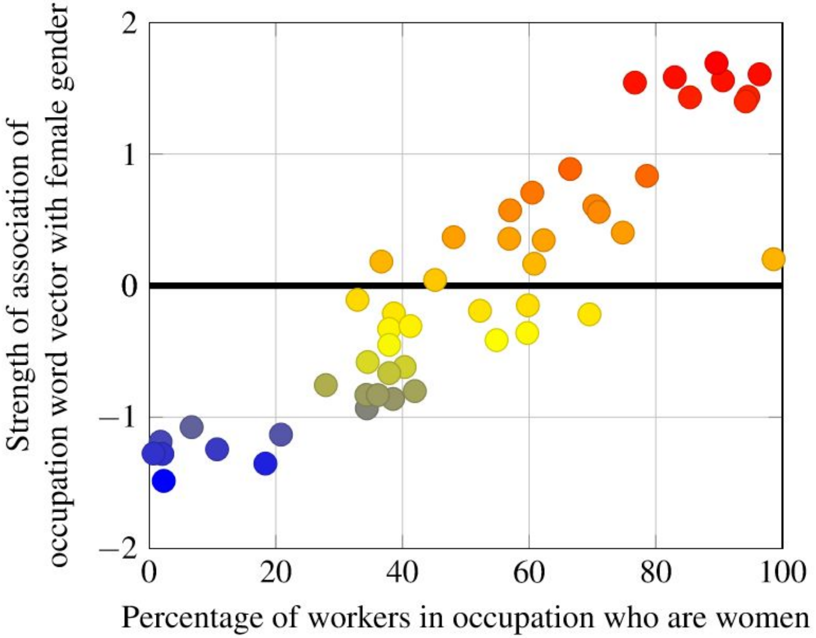
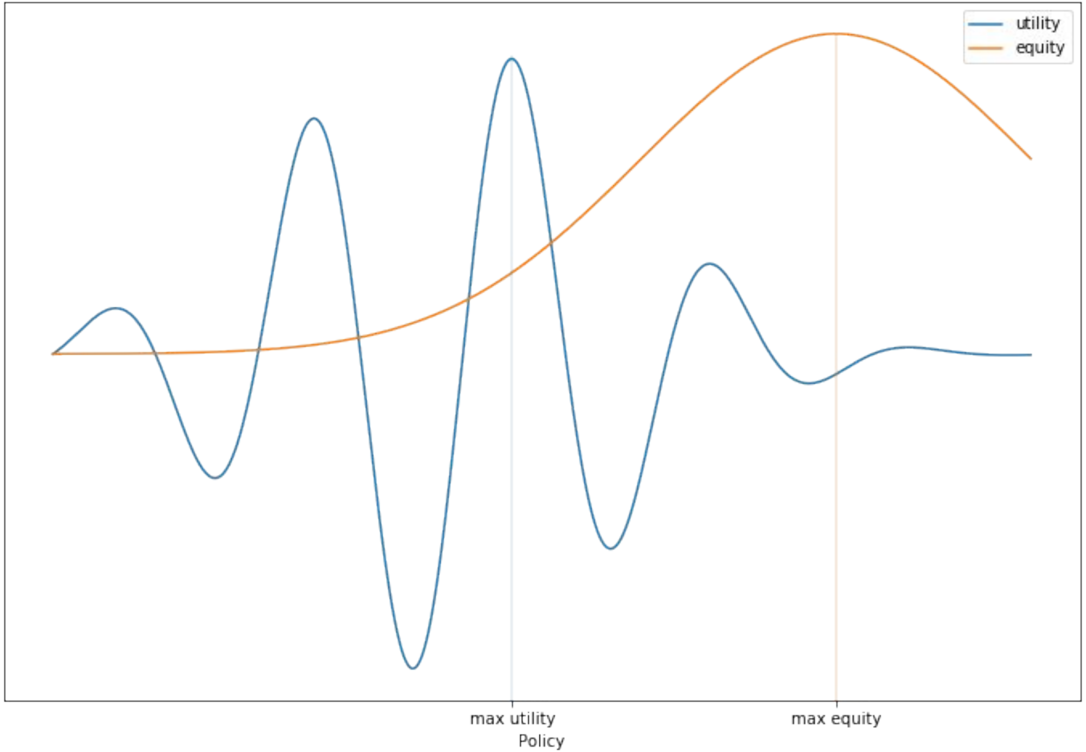


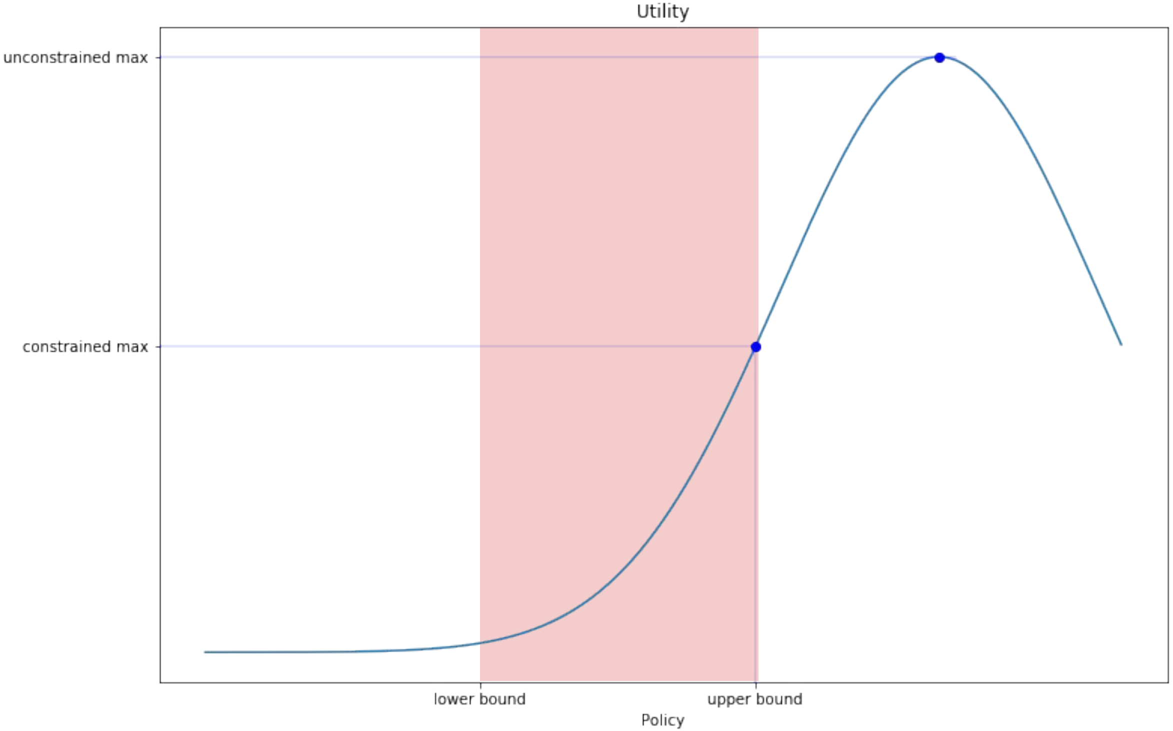
Figure 1. Occupation-gender association
Pearson’s correlation coefficient $\rho = 0.90$ with p-value $< 10^{-18}$.

Core problems in AI ethics

Can’t simultaneously maximize two objectives



Constrained max <= Global max



Outcomes and protected classes are correlated

All About Hyderabad

Things from Hyderabad:

- Great Biryani
- Dum ke Roat (best cookies in India)
- My wife
- **Pervasive fraud** on many lending platforms (including Simpl)

Simpl’s Underwriting Algo

Simpl (my employer) is an Indian microlending platform/payment processor.

Input data: Old data + specific fraud behavior.

Algorithm: A big, unstructured black box (think random forest or neural network).

Prediction target: 30 day delinquency, i.e. “has the user paid their bill within 30 days of the first bill due date”.

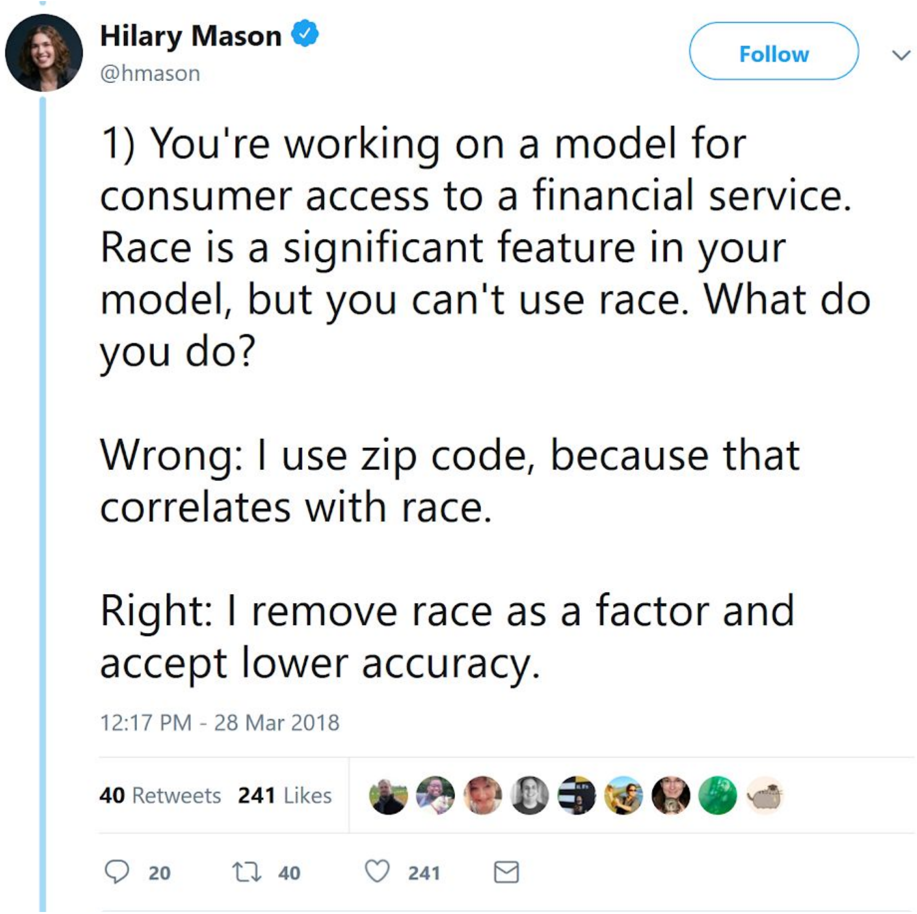
(I can’t reveal what specific fraud behavior is - think of it as something like installing the चोर App on the Evil Play Store.)

Simpl’s Underwriting Algo

If we exclude चोर app, Hyderabad is a strong indicator of delinquency.

If we include चोर app, that is the dominant feature. It’s also highly correlated with Hyderabad and results in a very high rejection rate there.

Fact: Lending is a low margin business. Lower accuracy results in fraudsters stealing all the money.



Hilary Mason
@hmason

1) You're working on a model for consumer access to a financial service. Race is a significant feature in your model, but you can't use race. What do you do?

Wrong: I use zip code, because that correlates with race.

Right: I remove race as a factor and accept lower accuracy.

12:17 PM - 28 Mar 2018

40 Retweets 241 Likes

Tradeoffs

Utilitarianism: We can offer loans to many Mumbaikars and Delhiites, and a smaller number of Hyderabadis. That’s a strict Pareto improvement over offering no loans to anyone.

Procedural fairness: Hyderabadis who install the चोर App (that’s “thief” in Hindi) are treated the same as Punekars who do the same (and vice versa).

Group unfairness: Our policy has a disparate impact on Hyderabadis - they get fewer loans issued.

Group reputation: We have learned a true but unflattering fact about Hyderabad: there is a disproportionate number of fraudsters there [1].

[1] Another possibility is a proportionate number of disproportionately active fraudsters.

100%

This is the rejection rate for Hyderabad loan applications at many other NBFCs.

In the American context this is called **redlining**.

(Context: In 1934, the USA Federal Housing Association drew a red line around black neighborhoods and told banks not to issue mortgages there.)

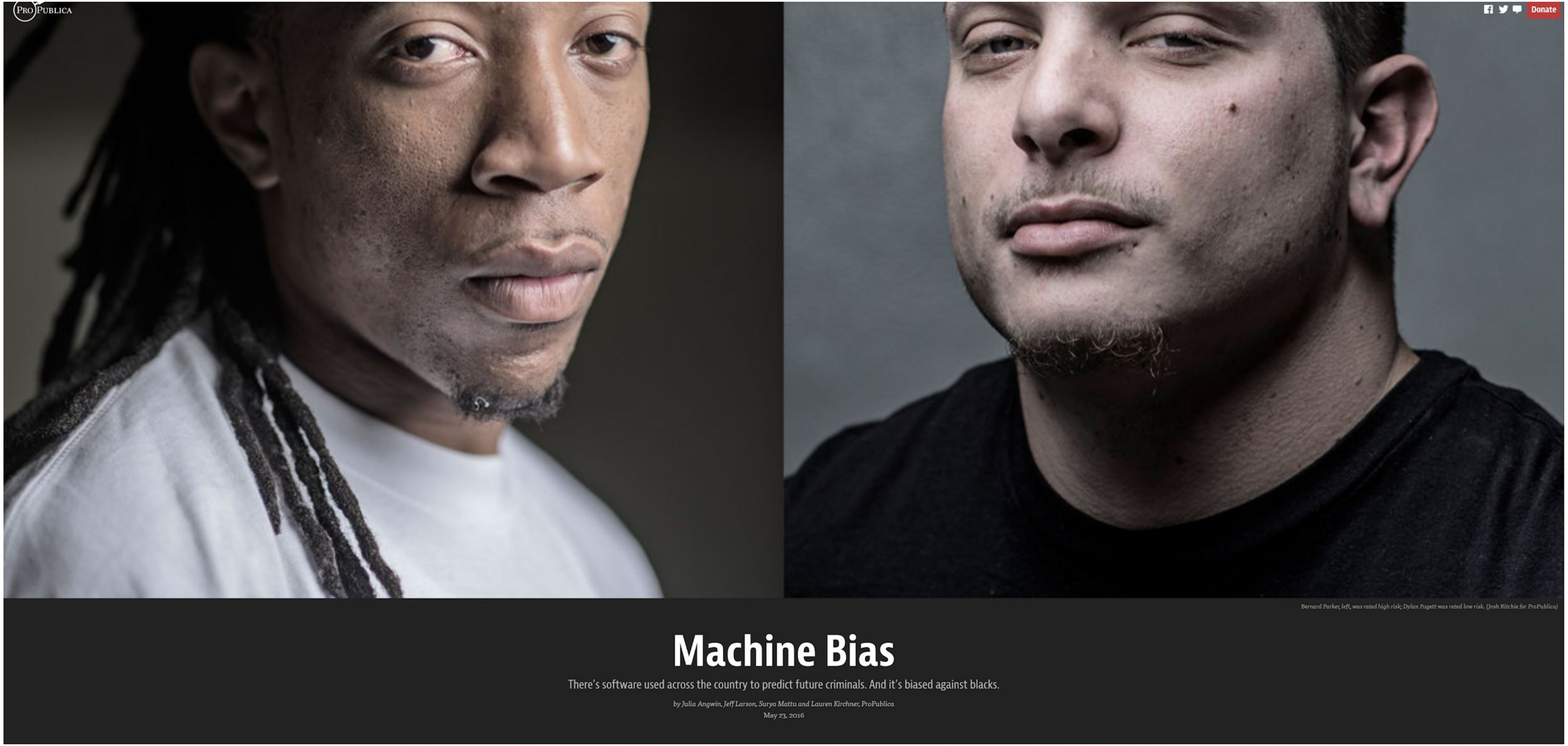
Simpl lives in a competitive market

If we choose to service Hyderabad with no disparities, we’ll run out of money and stop serving Hyderabad. The other NBFCs won’t.

Net result: Hyderabad is redlined by competitors and still gets no service.

Our choice: Keep the fraudsters out, utilitarianism over group rights.

A couple of weeks ago my mother in law - who lives in Hyderabad - informed me that Simpl approved her credit line.



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

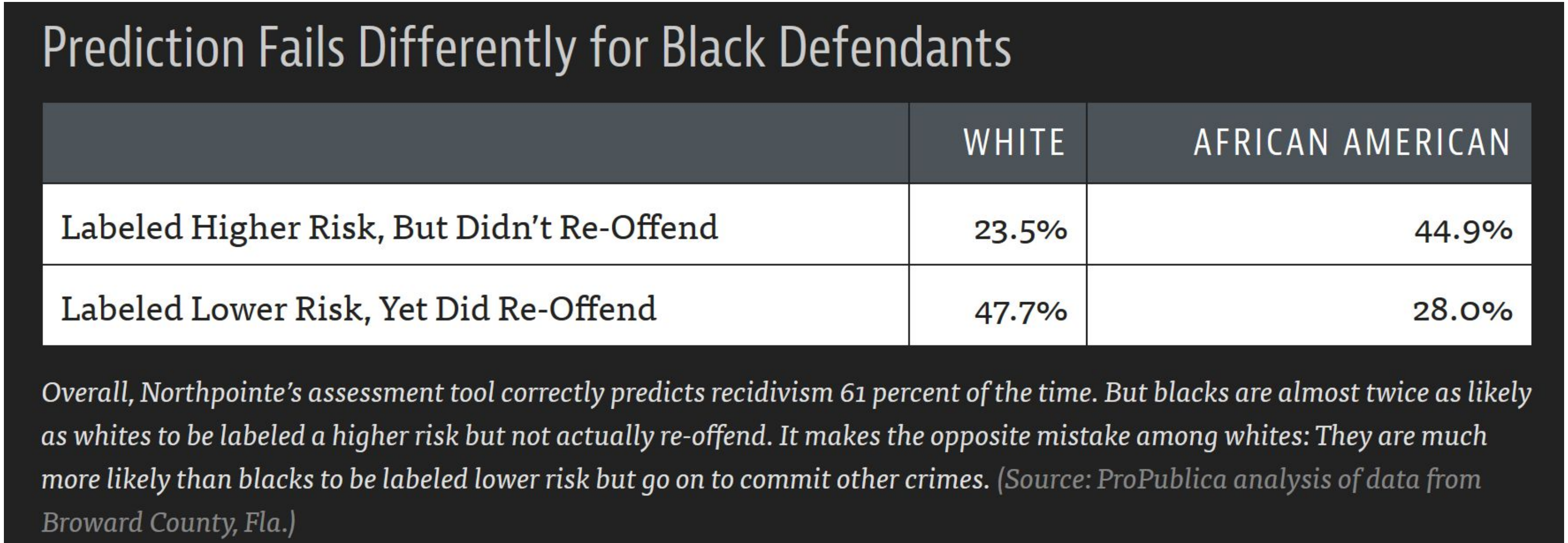
(Screenshot of ProPublica's Article)

Computational Criminology

COMPAS Algorithm

137 factors go into a black box model - age, gender, criminal history, single mother, father went to jail, number of friends who use drugs, etc.

ProPublica claims it’s “biased against blacks”.



Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn’t Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe’s assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

How does COMPAS work?

Dressel and Farid replicated COMPAS predictions using Logistic Regression on only 7 features: age, sex, #juvenile misdemeanors, #juvenile felonies, #adult crimes, crime degree (most recent), and crime charge (most recent).

Goal: Explainable model with same predictions as COMPAS. It’s predictions:

- 25 year old male who kidnapped and raped 6 women: high risk
- 43 year old female who shoplifted a toy for her kid one Christmas: low risk

(Adding race does not significantly improve accuracy.)

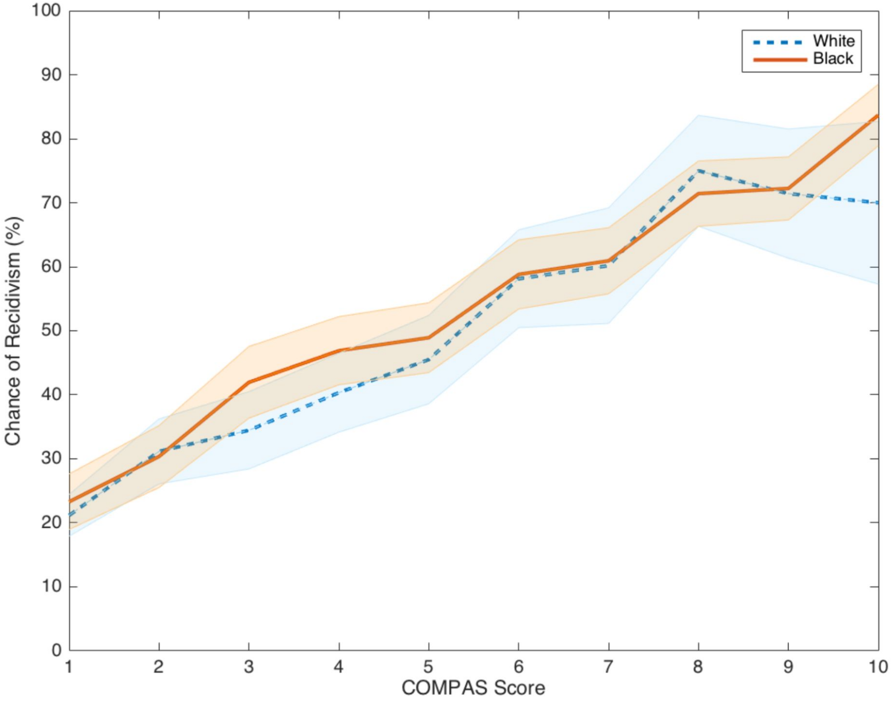
Checking Calibration

ProPublica checked the calibration of the algorithm, and found a disparity that was “almost statistically significant” at p=0.057.

(Flashback to CrunchConf 2015: Multiple Comparisons - Make your boss happy with false positives. Correcting ProPublica’s multiple comparisons, p=0.114.)

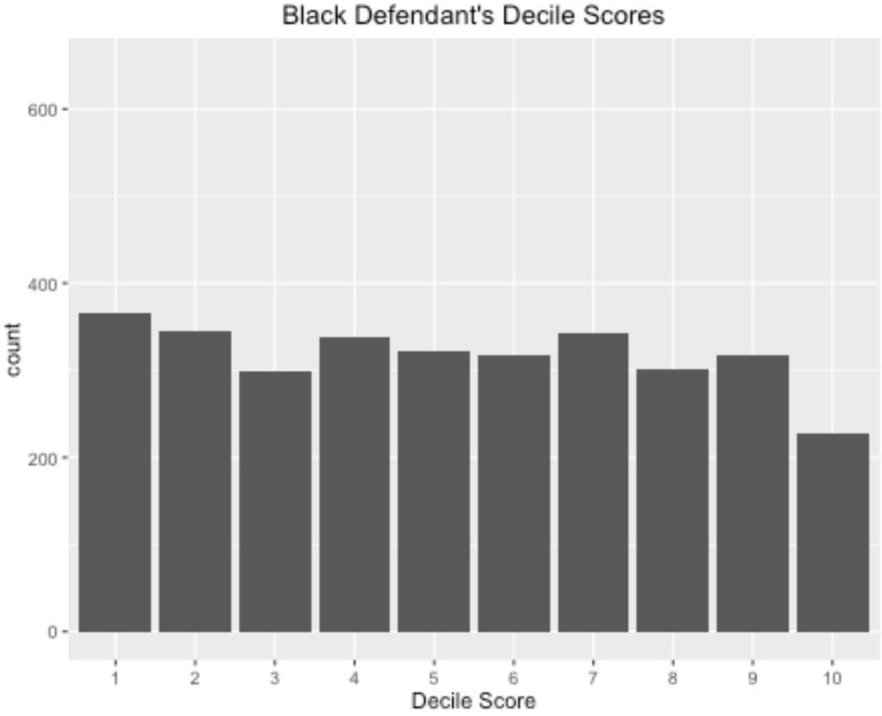
Conclusion: A black or white person with a risk score of 5 have equal probability of recidivism.

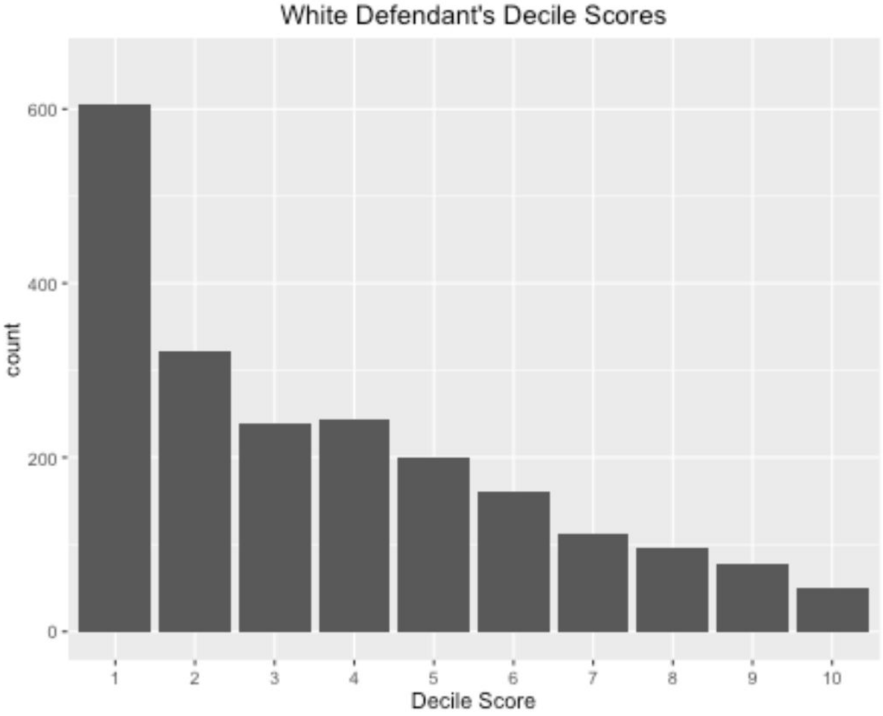
(Key point is cells 28-29 in their R Script.)



Accuracy and Racial Biases of Recidivism Prediction Instruments, Julia J. Dressel

Distribution of scores





The “bias” comes from base rates

Simplified numbers: High risk == 60% chance of recidivism, low risk = 20%.

Black people: **60% labelled high risk** * 40% chance of no recidivism/ = 24% chance of “labelled high risk, didn’t recidivate”.

White people: **30% labelled high risk** * 40% chance of no recidivism = 12% chance of “labelled high risk, didn’t recidivate”.

(Not going to do a calculation with 10 deciles in a 40 minute talk.)

The “bias” comes from base rates

Calibration means: P(~recidivate|black, high risk) = P(~recidivate|white, high risk)

Group fairness means: P(high risk | black, ~recidivate) = P(high risk | ~recidivate)

Bayes theorem provides a relationship:

P(high risk | ~recidivate) = P(~recidivate | high risk) P(high risk) / P(recidivate)

The “bias” comes from base rates

P(high risk | ~recidivate) = P(~recidivate | high risk) P(high risk) / P(recidivate)

Disparity is caused by base rates being different. P(high risk) is significantly higher for blacks than whites, and as a result, P(high risk | ~recidivate) must also be higher.

If calibration is equal across groups, then by necessity false positive rates must differ if base rates do.

Impossibility Theorem

Theorem (Kleinberg, Mullainathan, Raghavan): A predictive algorithm can only be well calibrated and have equal false positive/negative rates if it achieves either perfect accuracy or base rates are equal.

Advice for journalists: If you run an analysis the way ProPublica did, you are mathematically guaranteed to get a conclusion of “bias.” There’s no risk of getting a bad story.

How to fix the disparity

Racially specific thresholds: Raise the “high risk” cutoff for one group, not the other.

Raising risk thresholds makes these terms decrease.

$$P(\text{high risk} \mid {\sim}\text{recidivate}) = P({\sim}\text{recidivate} \mid \text{high risk})\ P(\text{high risk}) \,/\, P(\text{recidivate})$$

Which in turn makes this decrease.

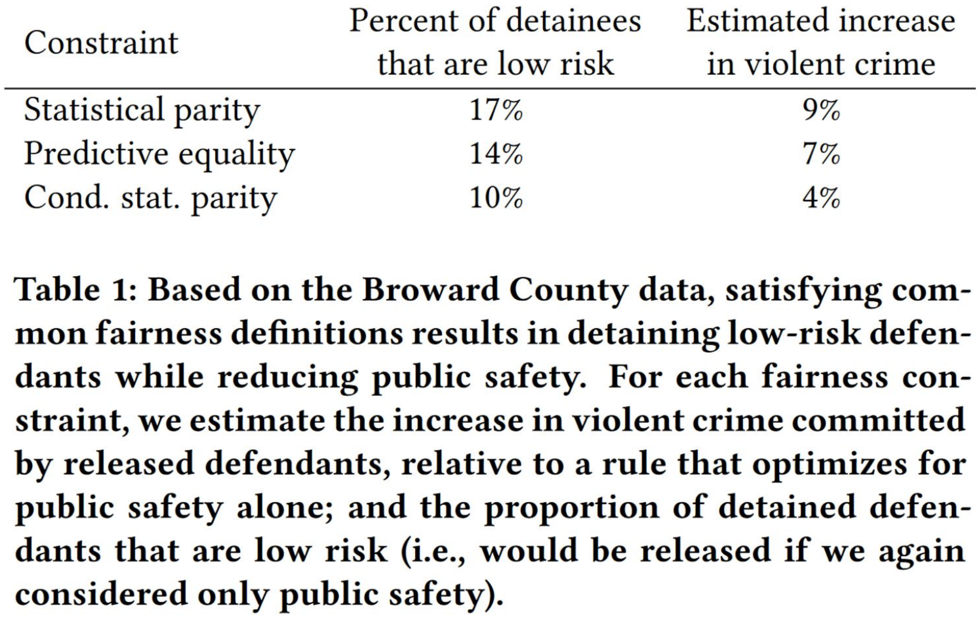
Procedurally: If the high risk threshold is 3 robberies, a black criminal at this threshold is paroled while a white criminal is jailed.

Utilitarian cost

Theorem: For a utility function of the form U = A*Crime - B*#people jailed, the maxima is achieved when all risk thresholds are equal.

Proof: Ordinary calculus. (Technical assumption: pdf of risk scores is continuous, non-vanishing.)

Intuitive meaning: To reduce crime as much as possible, put people in jail in order of highest risk first.



Constraint	Percent of detainees that are low risk	Estimated increase in violent crime
Statistical parity	17%	9%
Predictive equality	14%	7%
Cond. stat. parity	10%	4%

Table 1: Based on the Broward County data, satisfying common fairness definitions results in detaining low-risk defendants while reducing public safety. For each fairness constraint, we estimate the increase in violent crime committed by released defendants, relative to a rule that optimizes for public safety alone; and the proportion of detained defendants that are low risk (i.e., would be released if we again considered only public safety).

[Algorithmic decision making and the cost of fairness]

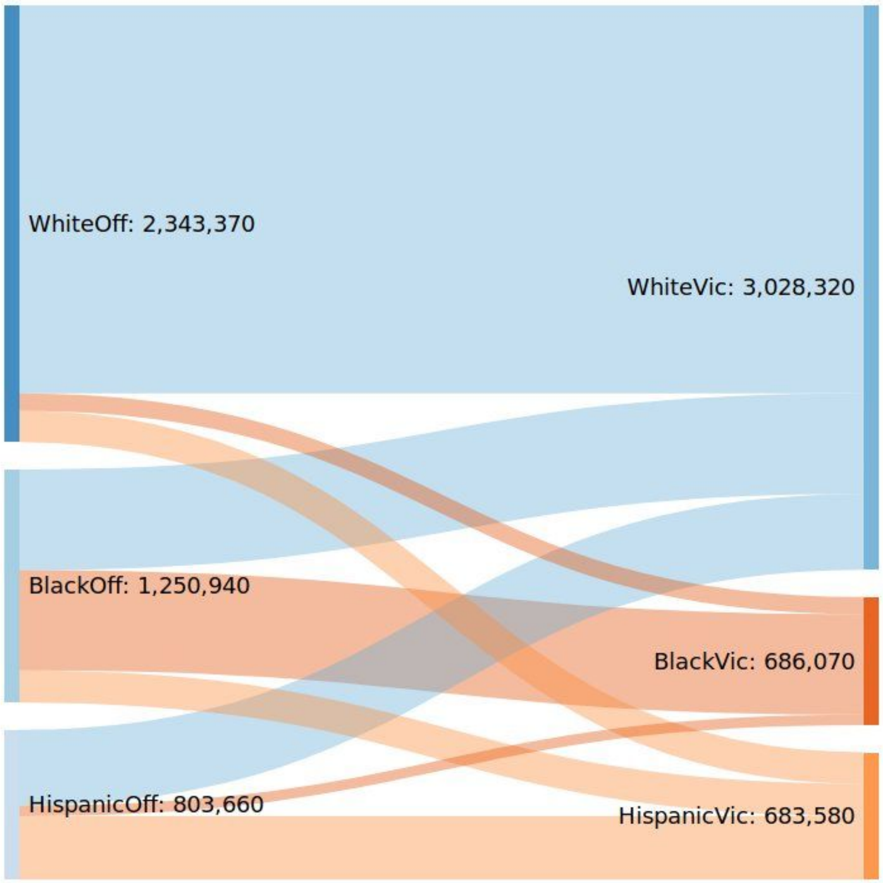
Victimization disparity

Crime victimization is disproportionately intraracial.

If crime goes up by 9% due to releasing black criminals from jail, this will have the following (ballpark) effect.

- Crime victimization will go up by about 7.5% among non-hispanic whites.
- Crime victimization will go up by about 37.5% among blacks.

[1] These stats are approximate, gained by multiplying Broward County crime increase percentages by national demographic and crime victimization statistics.



@AnechoicMedia14, avg yearly # violent crimes with injury

Tradeoffs

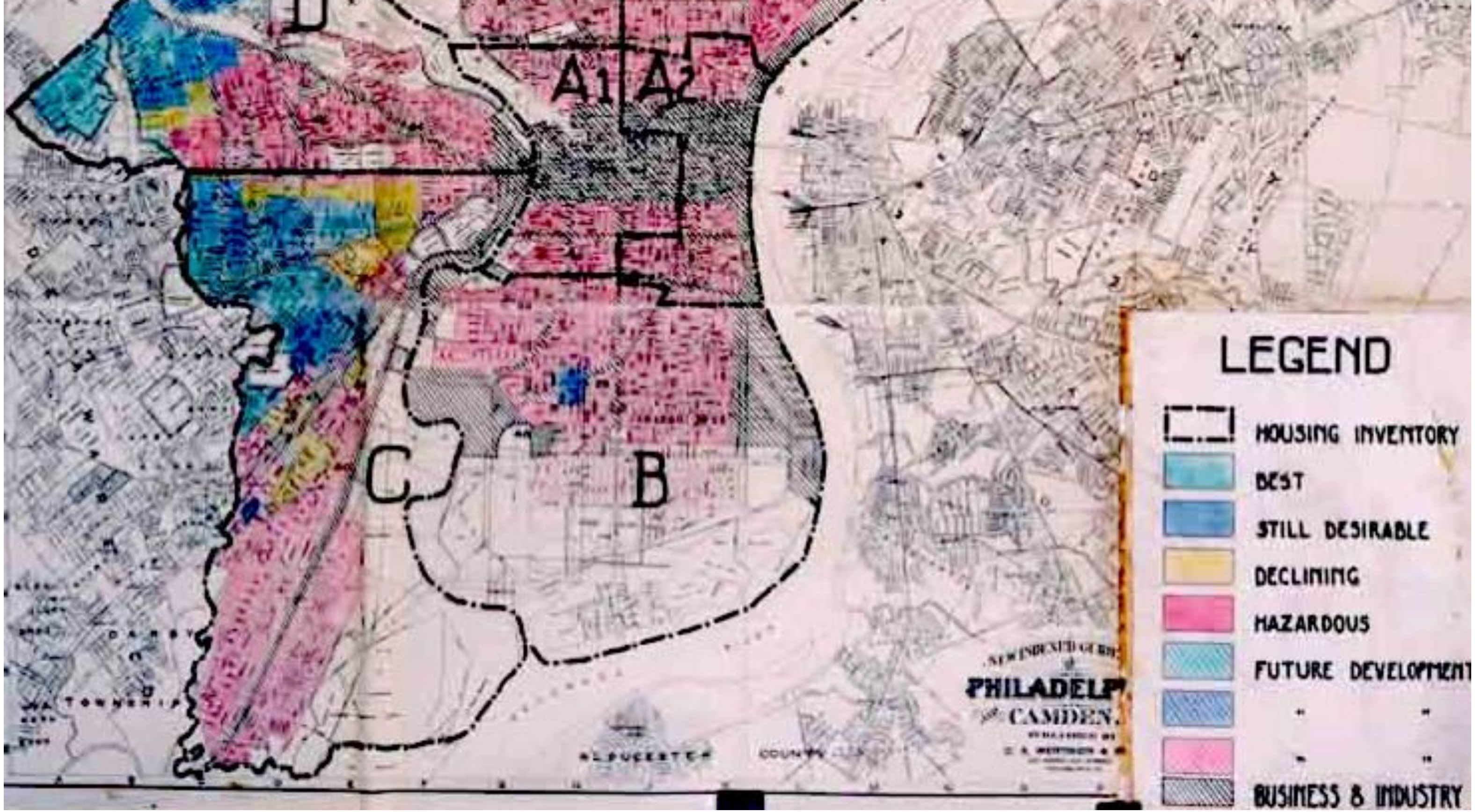
Utilitarianism: COMPAS reduces crime. Using a fairer algorithm would cause people to be raped and murdered.

Procedural fairness: COMPAS treats a black bike thief identically to a white bike thief, and a black serial killer identically to a white serial killer.

Group fairness (for victims): COMPAS reduces - but does not eliminate - racial disparities in crime victimization.

Group unfairness (for criminals): COMPAS causes a disparity in false positive rates.

Representational unfairness: COMPAS reveals that blacks are significantly more likely than whites to recidivate.



Fairness in lending

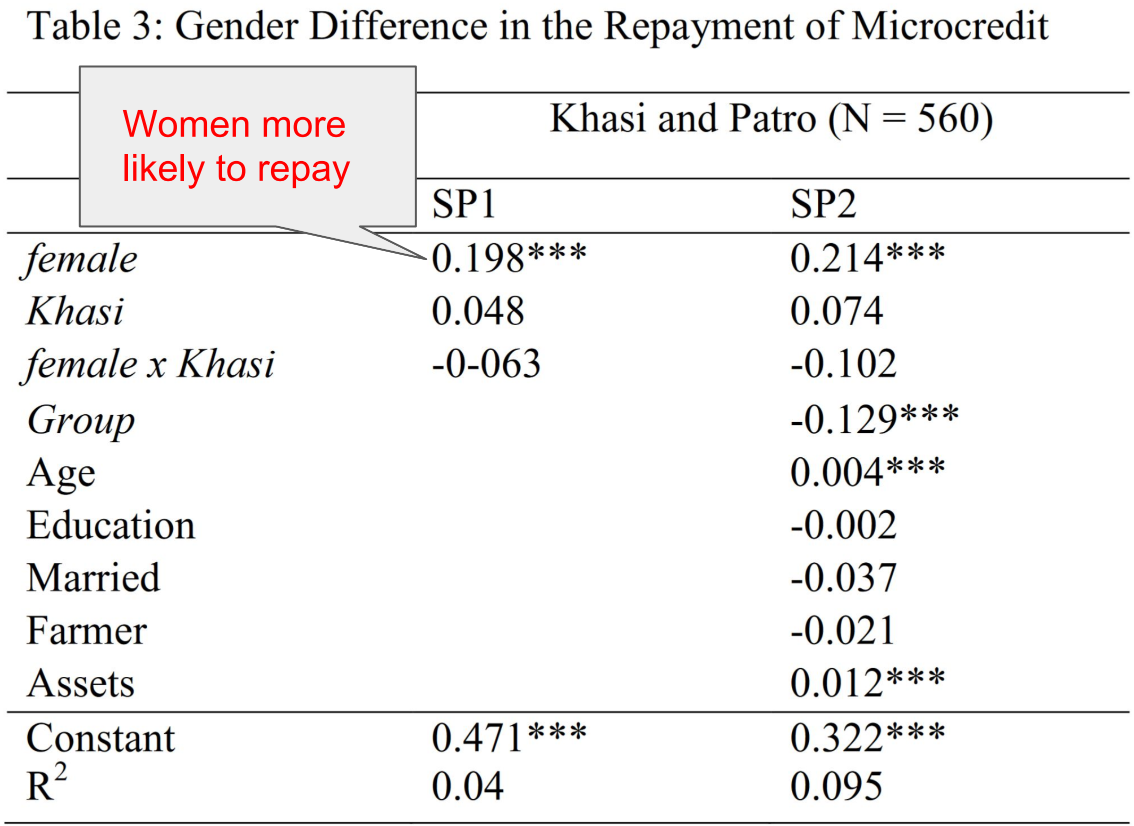


Table 3: Gender Difference in the Repayment of Microcredit

	Khasi and Patro (N = 560)	
	SP1	SP2
female	0.198***	0.214***
Khasi	0.048	0.074
female x Khasi	-0-063	-0.102
Group		-0.129***
Age		0.004***
Education		-0.002
Married		-0.037
Farmer		-0.021
Assets		0.012***
Constant	0.471***	0.322***
R^2	0.04	0.095

Women more likely to repay

Are Women “Naturally” Better Credit Risks in Microcredit?

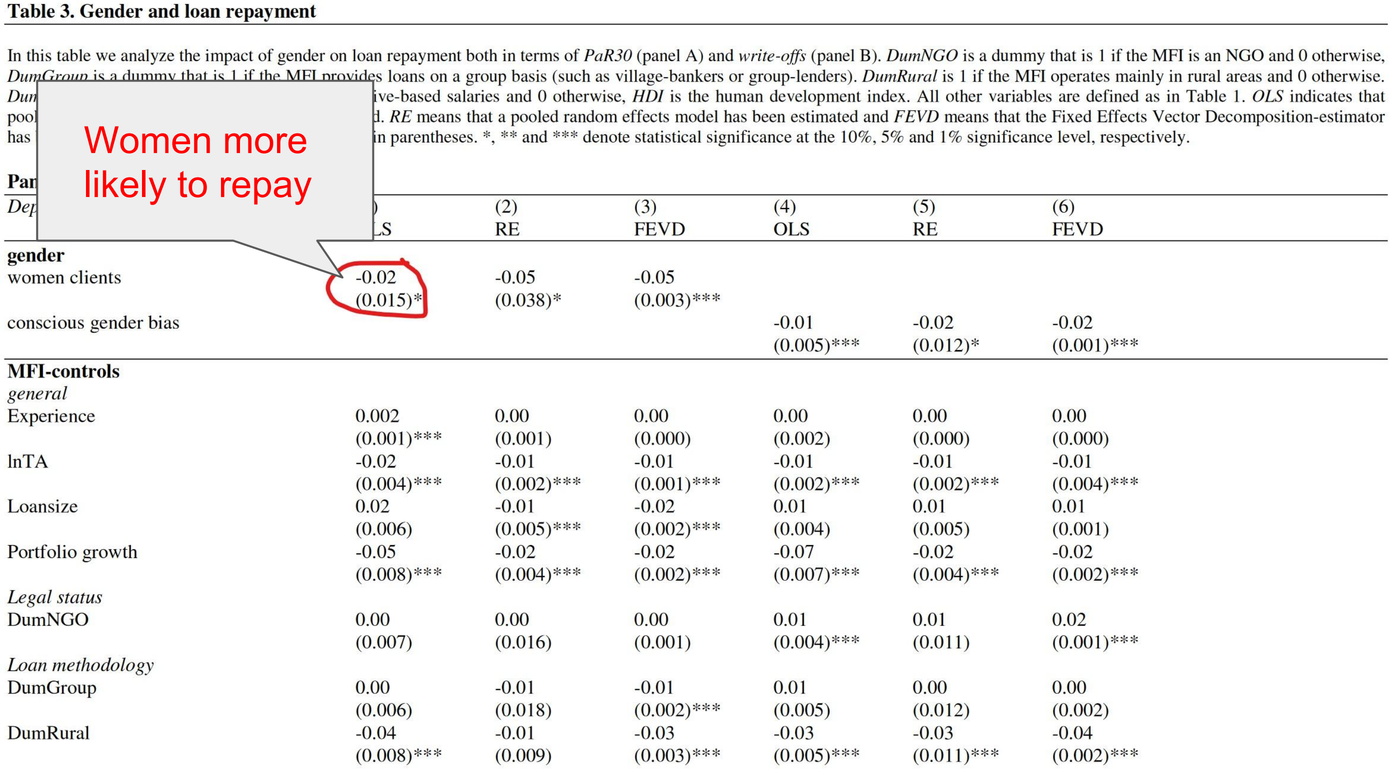


Table 3. Gender and loan repayment

In this table we analyze the impact of gender on loan repayment both in terms of *PaR30* (panel A) and *write-offs* (panel B). *DumNGO* is a dummy that is 1 if the MFI is an NGO and 0 otherwise, *DumGroup* is a dummy that is 1 if the MFI provides loans on a group basis (such as village-bankers or group-lenders). *DumRural* is 1 if the MFI operates mainly in rural areas and 0 otherwise. *Dum* ive-based salaries and 0 otherwise, *HDI* is the human development index. All other variables are defined as in Table 1. *OLS* indicates that pool d. *RE* means that a pooled random effects model has been estimated and *FEVD* means that the Fixed Effects Vector Decomposition-estimator has in parentheses. *, ** and *** denote statistical significance at the 10%, 5% and 1% significance level, respectively.

Women more likely to repay

Par

Dep	)	(2)	(3)	(4)	(5)	(6)
	LS	RE	FEVD	OLS	RE	FEVD
gender						
women clients	-0.02	-0.05	-0.05			
	(0.015)*	(0.038)*	(0.003)***			
conscious gender bias				-0.01	-0.02	-0.02
				(0.005)***	(0.012)*	(0.001)***
MFI-controls						
general						
Experience	0.002	0.00	0.00	0.00	0.00	0.00
	(0.001)***	(0.001)	(0.000)	(0.002)	(0.000)	(0.000)
lnTA	-0.02	-0.01	-0.01	-0.01	-0.01	-0.01
	(0.004)***	(0.002)***	(0.001)***	(0.002)***	(0.002)***	(0.004)***
Loansize	0.02	-0.01	-0.02	0.01	0.01	0.01
	(0.006)	(0.005)***	(0.002)***	(0.004)	(0.005)	(0.001)
Portfolio growth	-0.05	-0.02	-0.02	-0.07	-0.02	-0.02
	(0.008)***	(0.004)***	(0.002)***	(0.007)***	(0.004)***	(0.002)***
Legal status						
DumNGO	0.00	0.00	0.00	0.01	0.01	0.02
	(0.007)	(0.016)	(0.001)	(0.004)***	(0.011)	(0.001)***
Loan methodology						
DumGroup	0.00	-0.01	-0.01	0.01	0.00	0.00
	(0.006)	(0.018)	(0.002)***	(0.005)	(0.012)	(0.002)
DumRural	-0.04	-0.01	-0.03	-0.03	-0.03	-0.04
	(0.008)***	(0.009)	(0.003)***	(0.005)***	(0.011)***	(0.002)***

Women and Repayment in Microfinance

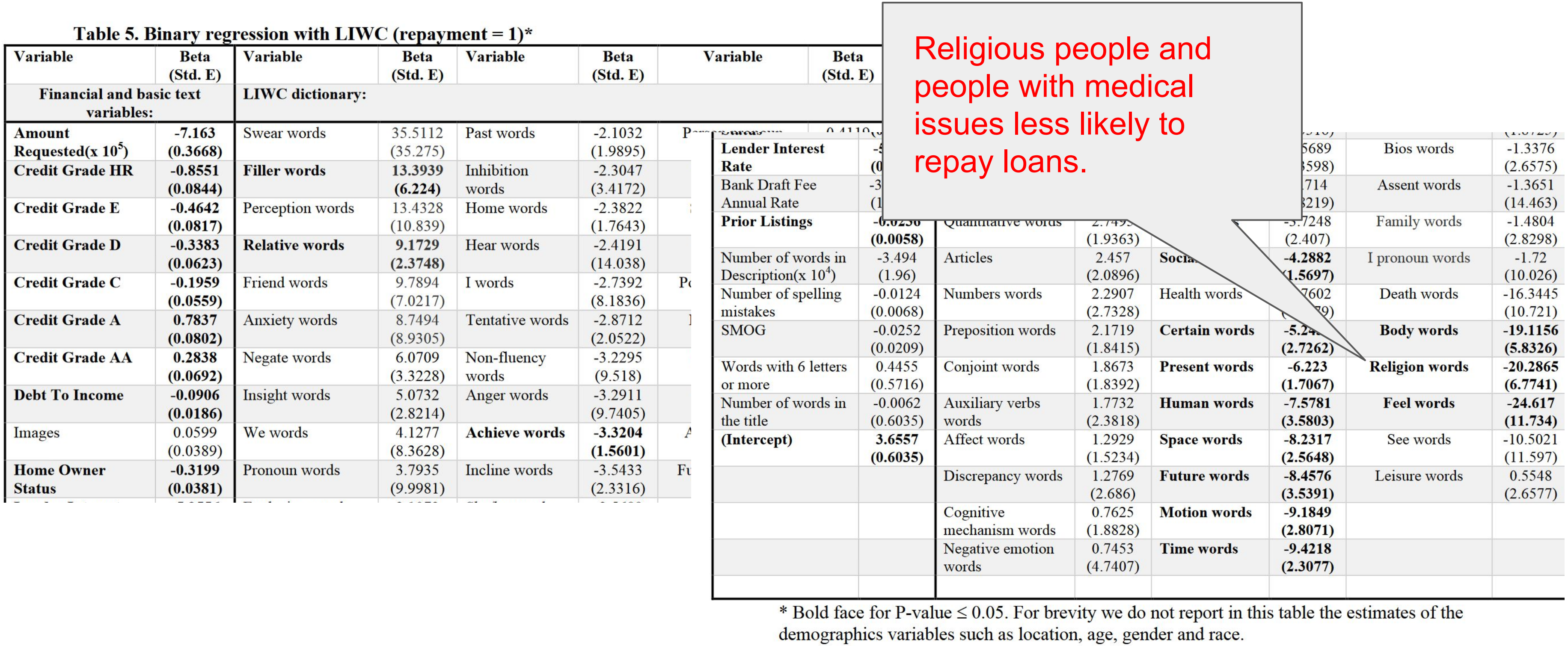


Table 5. Binary regression with LIWC (repayment = 1)*

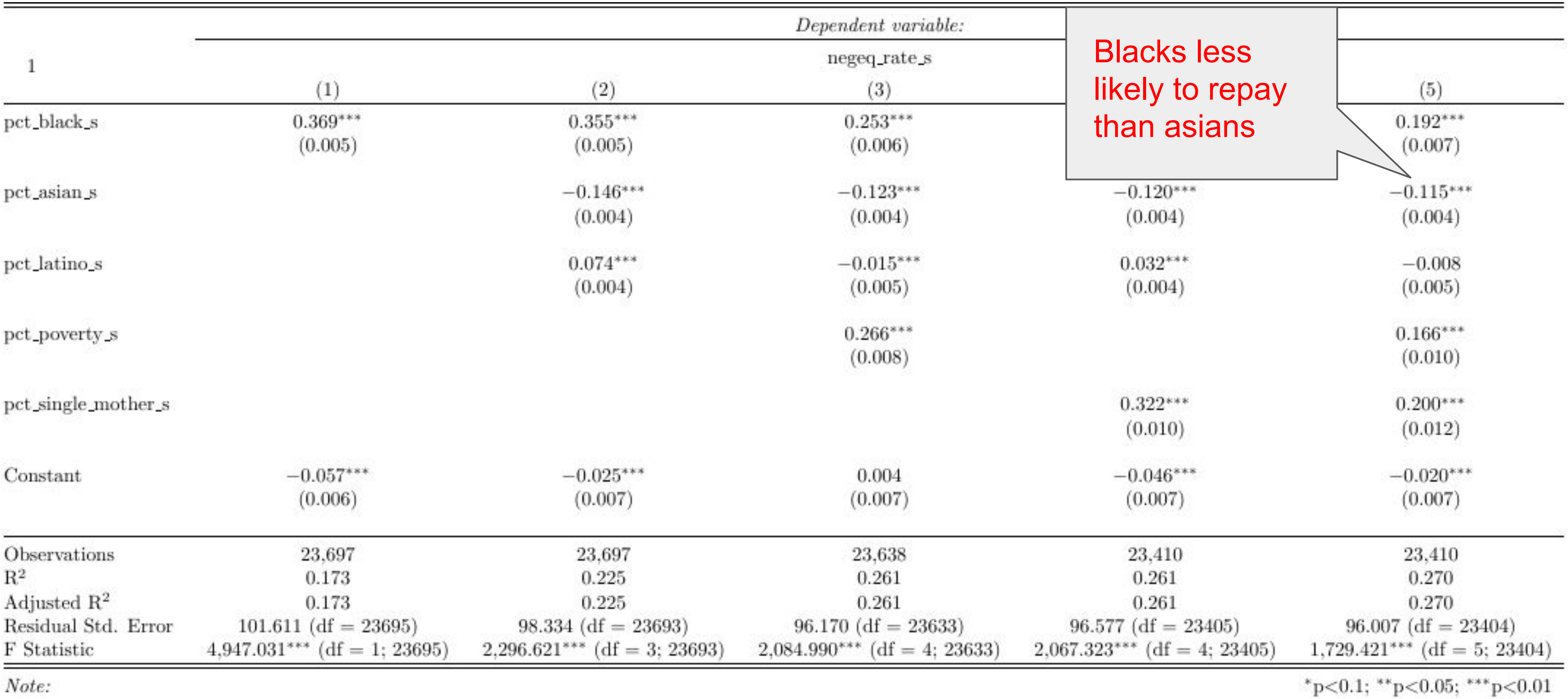
Variable	Beta (Std. E)	Variable	Beta (Std. E)	Variable	Beta (Std. E)	Variable	Beta (Std. E)
Financial and basic text variables:		LIWC dictionary:					
Amount Requested(x 10^5)	**-7.163 (0.3668)**	Swear words	35.5112 (35.275)	Past words	-2.1032 (1.9895)	P	
Credit Grade HR	**-0.8551 (0.0844)**	**Filler words**	**13.3939 (6.224)**	Inhibition words	-2.3047 (3.4172)		
Credit Grade E	**-0.4642 (0.0817)**	Perception words	13.4328 (10.839)	Home words	-2.3822 (1.7643)		
Credit Grade D	**-0.3383 (0.0623)**	**Relative words**	**9.1729 (2.3748)**	Hear words	-2.4191 (14.038)		
Credit Grade C	**-0.1959 (0.0559)**	Friend words	9.7894 (7.0217)	I words	-2.7392 (8.1836)	Po	
Credit Grade A	**0.7837 (0.0802)**	Anxiety words	8.7494 (8.9305)	Tentative words	-2.8712 (2.0522)	I	
Credit Grade AA	**0.2838 (0.0692)**	Negate words	6.0709 (3.3228)	Non-fluency words	-3.2295 (9.518)		
Debt To Income	**-0.0906 (0.0186)**	Insight words	5.0732 (2.8214)	Anger words	-3.2911 (9.7405)		
Images	0.0599 (0.0389)	We words	4.1277 (8.3628)	**Achieve words**	**-3.3204 (1.5601)**	A	
Home Owner Status	**-0.3199 (0.0381)**	Pronoun words	3.7935 (9.9981)	Incline words	-3.5433 (2.3316)	Fu	

Variable	Beta (Std. E)	Variable	Beta (Std. E)	Variable	Beta (Std. E)	Variable	Beta (Std. E)
Lender Interest Rate	**-5 (0**				5689 598)	Bios words	-1.3376 (2.6575)
Bank Draft Fee Annual Rate	-3 (1				.714 219)	Assent words	-1.3651 (14.463)
Prior Listings	**-0.0236 (0.0058)**	Quantitative words	2.749 (1.9363)		-3.7248 (2.407)	Family words	-1.4804 (2.8298)
Number of words in Description(x 10^4)	-3.494 (1.96)	Articles	2.457 (2.0896)	**Socia**	**-4.2882 (1.5697)**	I pronoun words	-1.72 (10.026)
Number of spelling mistakes	-0.0124 (0.0068)	Numbers words	2.2907 (2.7328)	Health words	7602 9)	Death words	-16.3445 (10.721)
SMOG	-0.0252 (0.0209)	Preposition words	2.1719 (1.8415)	**Certain words**	**-5.24 (2.7262)**	**Body words**	**-19.1156 (5.8326)**
Words with 6 letters or more	0.4455 (0.5716)	Conjoint words	1.8673 (1.8392)	**Present words**	**-6.223 (1.7067)**	**Religion words**	**-20.2865 (6.7741)**
Number of words in the title	-0.0062 (0.6035)	Auxiliary verbs words	1.7732 (2.3818)	**Human words**	**-7.5781 (3.5803)**	**Feel words**	**-24.617 (11.734)**
(Intercept)	**3.6557 (0.6035)**	Affect words	1.2929 (1.5234)	**Space words**	**-8.2317 (2.5648)**	See words	-10.5021 (11.597)
		Discrepancy words	1.2769 (2.686)	**Future words**	**-8.4576 (3.5391)**	Leisure words	0.5548 (2.6577)
		Cognitive mechanism words	0.7625 (1.8828)	**Motion words**	**-9.1849 (2.8071)**		
		Negative emotion words	0.7453 (4.7407)	**Time words**	**-9.4218 (2.3077)**		

* Bold face for P-value ≤ 0.05. For brevity we do not report in this table the estimates of the demographics variables such as location, age, gender and race.

Religious people and people with medical issues less likely to repay loans.

When Words Sweat: Identifying Signals for Loan Default in the Text of Loan Applications



	Dependent variable:				
1	negeq_rate_s				
	(1)	(2)	(3)		(5)
pct_black_s	0.369***	0.355***	0.253***		0.192***
	(0.005)	(0.005)	(0.006)		(0.007)
pct_asian_s		−0.146***	−0.123***	−0.120***	−0.115***
		(0.004)	(0.004)	(0.004)	(0.004)
pct_latino_s		0.074***	−0.015***	0.032***	−0.008
		(0.004)	(0.005)	(0.004)	(0.005)
pct_poverty_s			0.266***		0.166***
			(0.008)		(0.010)
pct_single_mother_s				0.322***	0.200***
				(0.010)	(0.012)
Constant	−0.057***	−0.025***	0.004	−0.046***	−0.020***
	(0.006)	(0.007)	(0.007)	(0.007)	(0.007)
Observations	23,697	23,697	23,638	23,410	23,410
R^2	0.173	0.225	0.261	0.261	0.270
Adjusted R^2	0.173	0.225	0.261	0.261	0.270
Residual Std. Error	101.611 (df = 23695)	98.334 (df = 23693)	96.170 (df = 23633)	96.577 (df = 23405)	96.007 (df = 23404)
F Statistic	4,947.031*** (df = 1; 23695)	2,296.621*** (df = 3; 23693)	2,084.990*** (df = 4; 23633)	2,067.323*** (df = 4; 23405)	1,729.421*** (df = 5; 23404)

Note: *p<0.1; **p<0.05; ***p<0.01

Blacks less likely to repay than asians

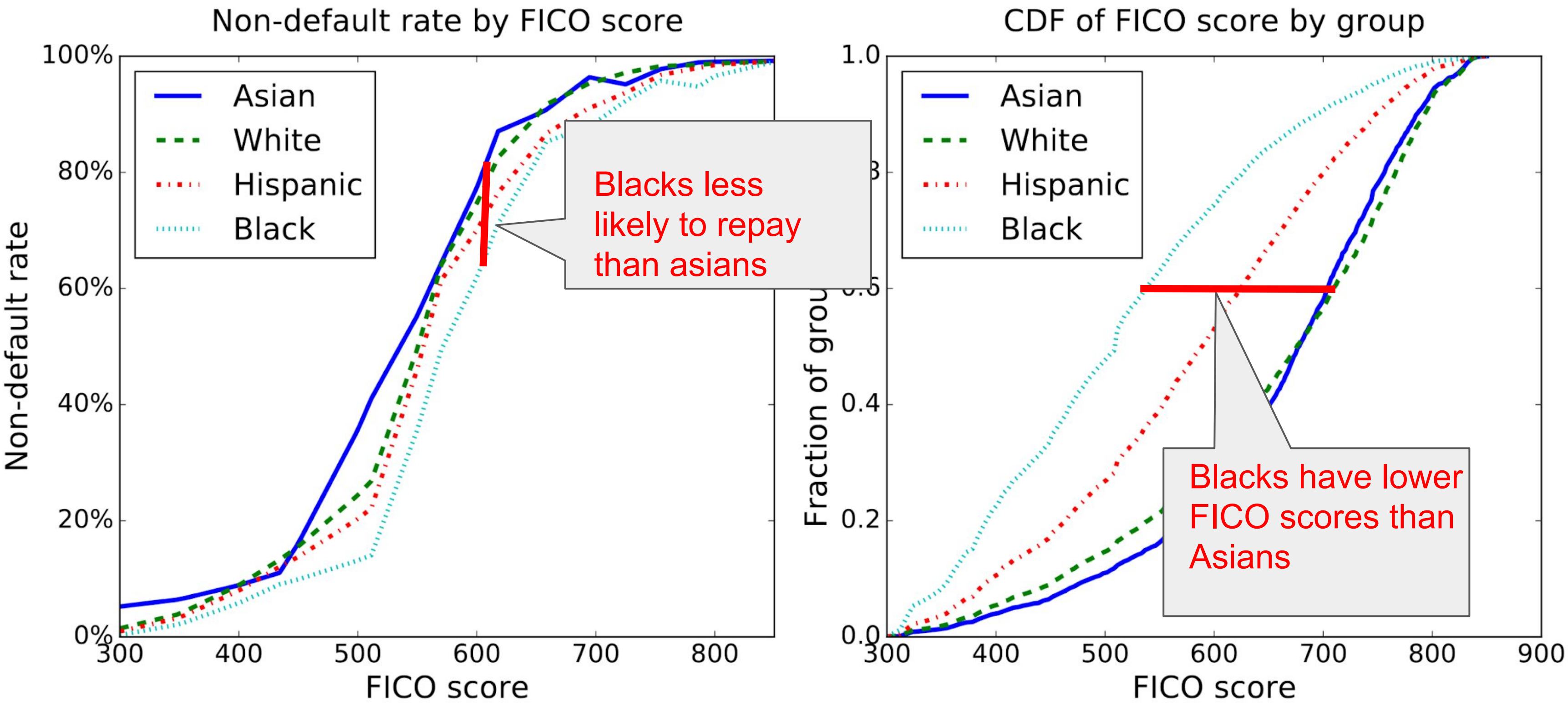
On the relationship between negative home owner equity and racial demographics

FICO Score

- 35% payment history (or lack of payments)
- 30% debt burden (how much you currently owe, relative to income, assets, assessed limits)
- 15% length of history
- 10% types of credit (credit card + mortgage + consumer unsecured > only credit card)
- 10% hard pulls (consumer explicitly applying for a loan)

Key point: 100% of FICO is based on a person’s actions (ignoring identity theft).

Procedurally fair, by definition.



Blacks less likely to repay than asians

Blacks have lower FICO scores than Asians

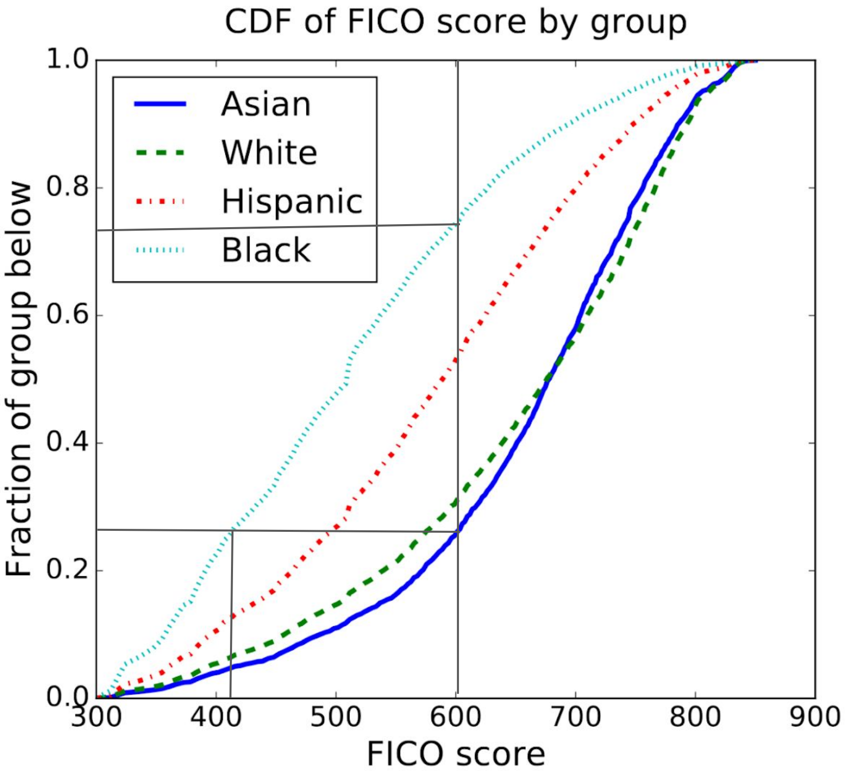
Equality of Opportunity in Supervised Learning

Exposing tradeoffs

Theorem: Using a single FICO threshold, one can only achieve the same approval rate across all groups at points x where CDF(x, A) = CDF(x, B).

Procedurally fair choice: Choose a fixed FICO cutoff, say 600. Then we reject 75% of blacks, 25% of Asians, violating the principle of group fairness.

Group fair choice: Choose a fixed approval level - say 75%, implying a risk cutoff of 600 for Asians and 410 for blacks. This violates principle of procedural fairness.

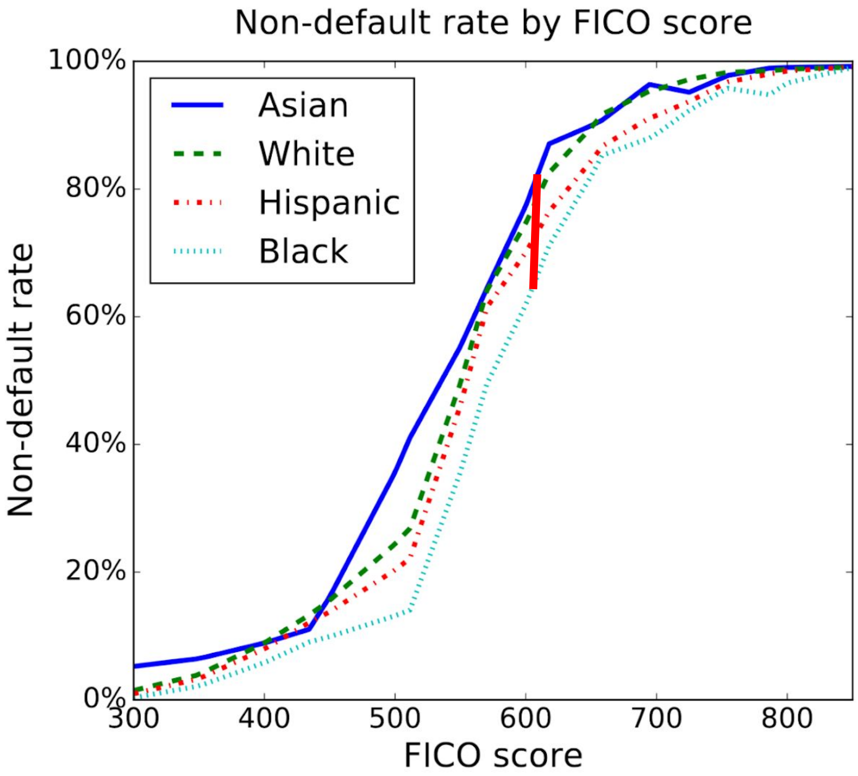


Exposing tradeoffs

Theorem: Any procedurally fair loan cutoff is not utility maximizing, except at FICO=850 (the max).

At FICO=600, approx 80% of Asian borrowers will repay loans and about 60% of Black borrowers will.

Utilitarian choice is to issue loans to Asian applicants with a 590 FICO (repayment probability ~78%) over black applicants with a 600 FICO.



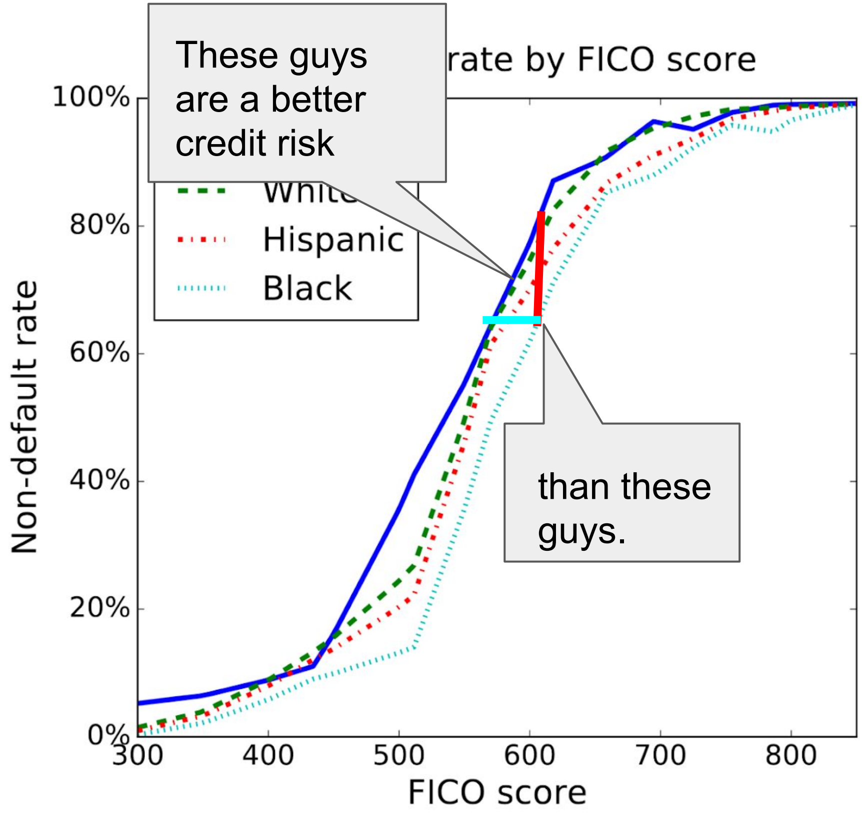
Exposing tradeoffs

Theorem: Any procedurally fair loan cutoff is not utility maximizing, except at FICO=850 (the max).

Proof: Draw a horizontal line to intersect the point where the FICO cutoff (a vertical line) meets the lower performing graph.

Find the point where horizontal line intersects the higher performing graph. This cutoff applied to the higher group, and the original cutoff applied to the lower group, is utility maximizing.

(Technical assumption: the density of the higher performing group is non-zero in this region.)



These guys are a better credit risk

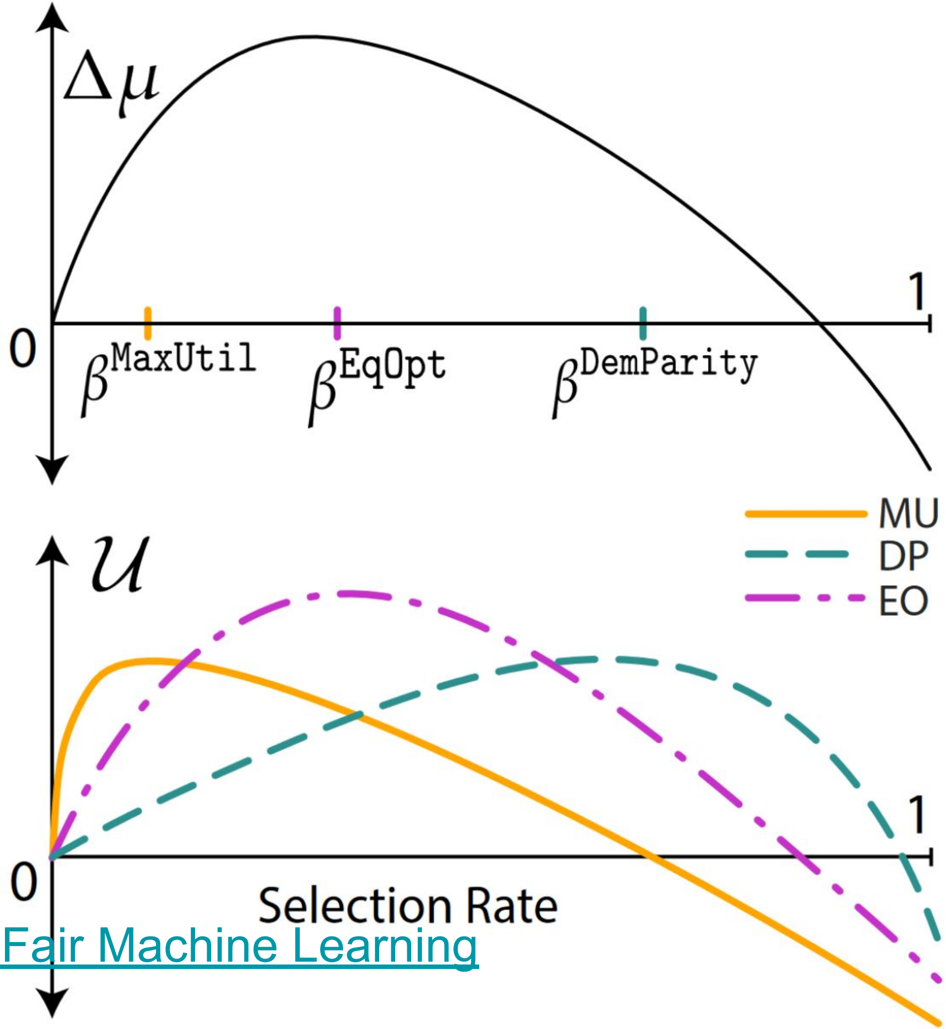
than these guys.

Predatory Lending

In 2008, the media used the term “predatory lending” to refer to making loans a bank knew would not be repaid.

Bad loans can cause many years of financial hardship, and also lower one’s credit score (exacerbating disparities).

These second order effects of attempts at fairness can drown out the first order effects, and harm those they are meant to help. (Whether this happens is a complex quantitative question.)



Delayed Impact of Fair Machine Learning

Tradeoffs

Utilitarianism: Racially discriminatory FICO cutoffs maximize utility.

Group fairness (today): Differently calibrated FICO cutoffs can result in either equal false positive rates, or equal loan acceptance rates.

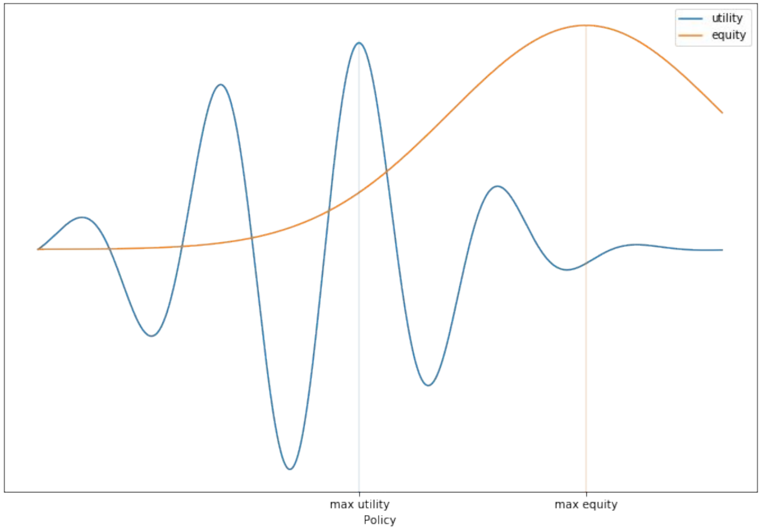
Group fairness (tomorrow): Other differently calibrated cutoffs can reduce credit score disparities (but not simultaneously with group fairness today).

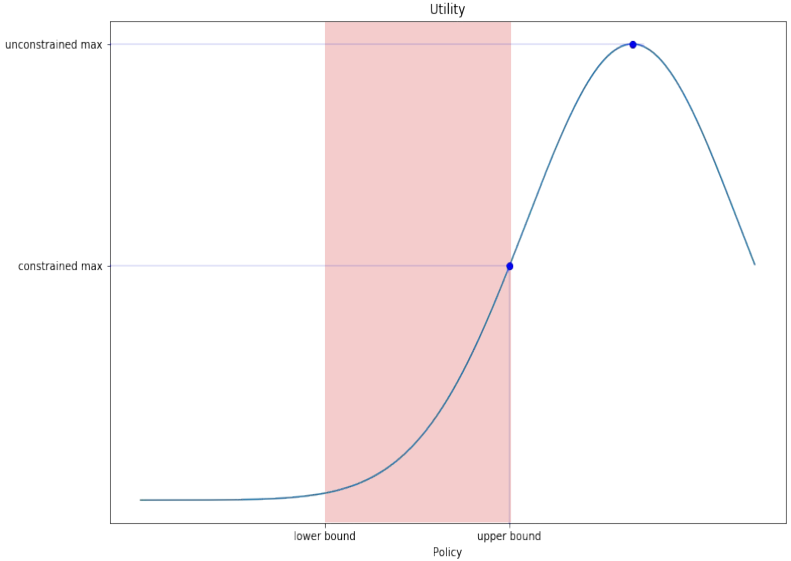
Group unfairness: Using FICO results in far fewer blacks being issued loans, and can reduce aggregate credit scores.

Representational unfairness: FICO reveals that blacks more likely to default than whites. Calibration graphs reveal that some of this default is not explained by financial history (i.e. FICO is biased in favor of blacks).

Procedural fairness: FICO score is independent of protected traits.

There is no policy choice which satisfies all ethical principles.





The laws of mathematical optimization still apply

Conclusion

Early on I said I wouldn’t be giving any ethical prescriptions.

I will, however, give one meta-ethical prescription: **formalize your ethical principles as terms in your utility function or as constraints.**

It is nearly certain that tradeoffs between these principles exist, and if we don’t acknowledge this, we run the risk of unknowingly engaging in bad actions.