Algorithmic Discrimination

Carlos Castillo
Web Science and Social Computing
ETIC — UPF

upf. Universitat Pompeu Fabra Barcelona

HUMAINT Workshop, March 2018

Session question addressed

Q2. How do algorithms have the potential to modify the way humans make decisions based on them?

My answer: algorithms can discriminate

Next, I present a precise formulation and examples

Generic discrimination

X discriminates against someone Y in relation to Z if:

1. Y has property P and Z does not have P
2. X treats Y worse than s/he treats or would treat Z
3. It is because Y has P and Z does not have P that X treats Y worse than Z

(also applies if X believes Y has P and Z does not have P)

Kasper Lippert-Rasmussen: Born Free and Equal? A Philosophical Inquiry Into the Nature of Discrimination. Oxford University Press, 2013.

Generic discrimination

X discriminates against someone Y in relation to Z if:

1. Y has property P and Z does not have P
2. X treats Y worse than s/he treats or would treat Z
3. It is because Y has P and Z does not have P that X treats Y worse than Z

Disadvantageous differential treatment

Kasper Lippert-Rasmussen: Born Free and Equal? A Philosophical Inquiry Into the Nature of Discrimination. Oxford University Press, 2013.

Group discrimination

X group-discriminates against Y in relation to Z if:

1. X generically discriminates against Y in relation to Z
2. P is the property of belonging to a socially salient group
3. This makes people with P worse off relative to others
   or X is motivated by animosity towards people with P,
   or by the belief that people with P are inferior
   or should not intermingle with others

Kasper Lippert-Rasmussen: Born Free and Equal? A Philosophical Inquiry Into the Nature of Discrimination. Oxford University Press, 2013.

Statistical discrimination

X statistically discriminates against Y in relation to Z if:

1. X group-discriminates against Y in relation to Z
2. P is statistically relevant
   (or X believes P is statistically relevant)

Kasper Lippert-Rasmussen: Born Free and Equal? A Philosophical Inquiry Into the Nature of Discrimination. Oxford University Press, 2013.

Example (statistical / non-statistical)

a. Not hiring a highly-qualified woman because women have a higher probability of taking parental leave (statistical discrimination)
b. Not hiring a highly-qualified woman because she has said that she intends to have a child and take parental leave (non-statistical discrimination)

Kasper Lippert-Rasmussen: Born Free and Equal? A Philosophical Inquiry Into the Nature of Discrimination. Oxford University Press, 2013.

In statistical machine learning

An algorithm developed through statistical machine learning can statistically discriminate if we:

1. Disregard intentions/animosity
2. Understand statistically relevant as
   any information derived from training data

Some examples

1. Disparate impact

   The model gives people with P a bad outcome more often

2. Directly and indirectly discriminatory rules

   The model associates P (or Q, which depends on P) to a bad outcome

3. Lack of calibration

   The same output translates to different bad outcome probs. for P and not P

4. Disparate mistreatment / Lack of equal opportunity

   The false positive rate of the bad outcome is higher for P than not P

5. Unfair rankings

   The model gives people with P a lower ranking

Disparate impact

Example:

"Protected group" = "people with disabilities"

"Benefit granted" = "getting a scholarship"

	benefit		
group	denied	granted	
protected	a	b	n_1
unprotected	c	d	n_2
	m_1	m_2	n

Intuitively, if

a/n_1, the **risk** that **people with disabilities** face of not getting a scholarship is much larger than

c/n_2, the **risk** that **people without disabilities** face of not getting a scholarship,

then people with disabilities could claim they are being discriminated.

D. Pedreschi, S. Ruggieri, F. Turini: A Study of Top-K Measures for Discrimination Discovery. SAC 2012.

Directly discriminatory rules

From a database of decisions made in the past, we learn that gender = female ⇒ credit = no

P(gender=female, credit=no) / P(gender=female) > θ

This means we have found evidence of **direct discrimination**

gender	has_job	credit
male	true	yes
male	false	yes
male	true	yes
female	false	no
female	true	no
female	true	yes
...	...	...

Hajian, S., Domingo-Ferrer, J. (2013). A methodology for direct and indirect discrimination prevention in data mining. IEEE TKDE 25(7), 1445-1459.

Indirectly discriminatory rules

If we learn the rule

zip = 8002 ⇒ credit = no

… and we know …

zip = 8002 ⇒ origin = foreign

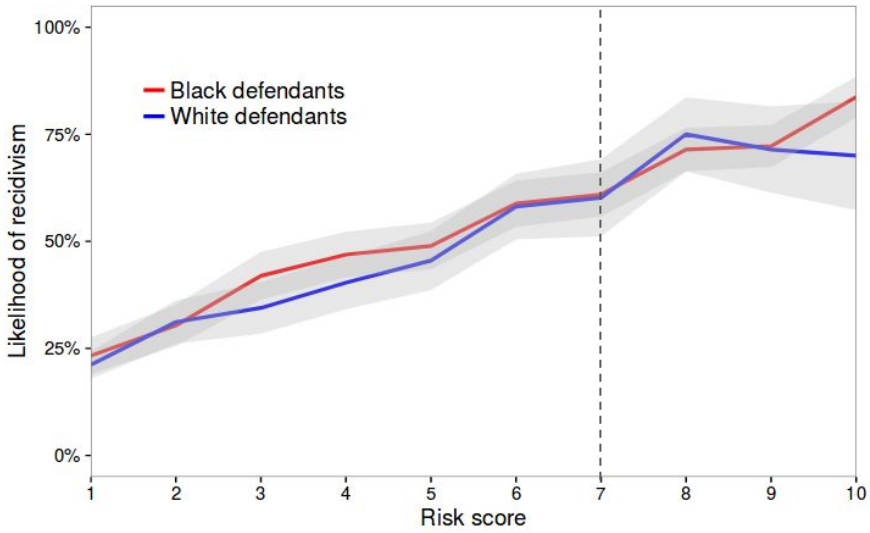
We have **indirect discrimination**

origin	zip	credit
national	8001	yes
national	8001	yes
national	8001	yes
foreign	8002	no
foreign	8002	no
foreign	8002	yes
...	...	...

Lack of calibration

A model lacks calibration if the probability of an actual bad outcome depends on the class and not only on the output

Source: Corbett-Davies et al. 2017



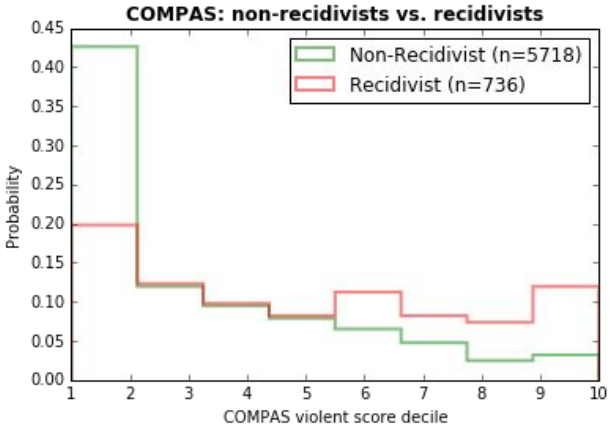
This is well calibrated: among people with a risk score of 7:

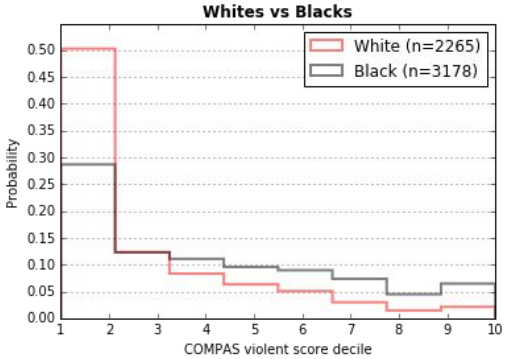
60% of blacks reoffended and

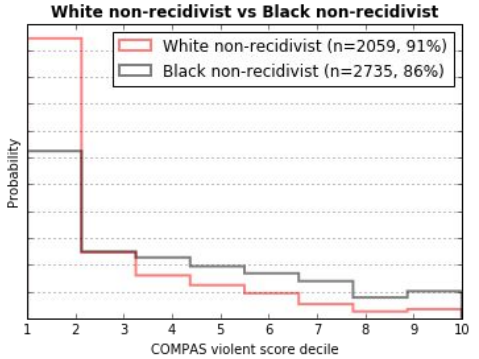
61% of whites reoffended

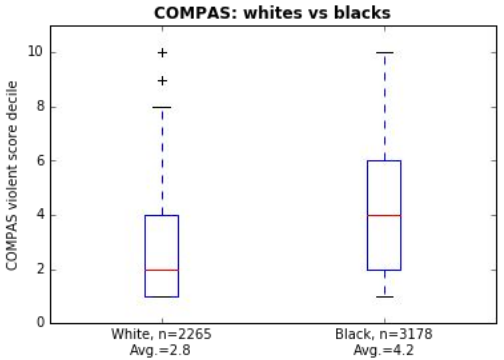
Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. In Proc. ITCS.
Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. In Proc. of KDD.

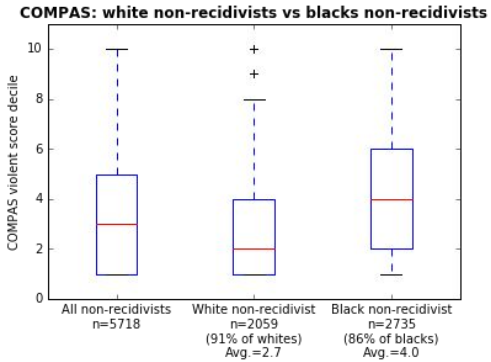
Disparate mistreatment / lack of E. O.





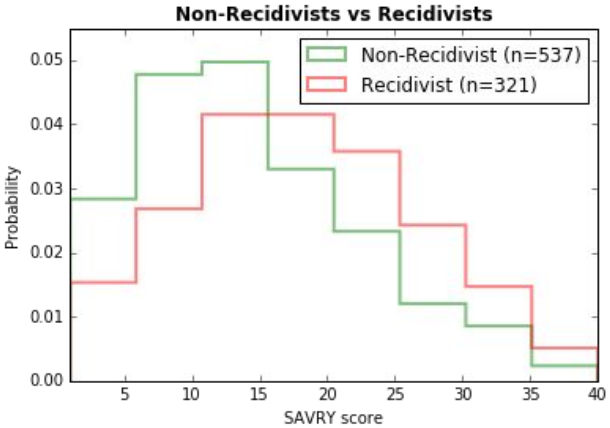


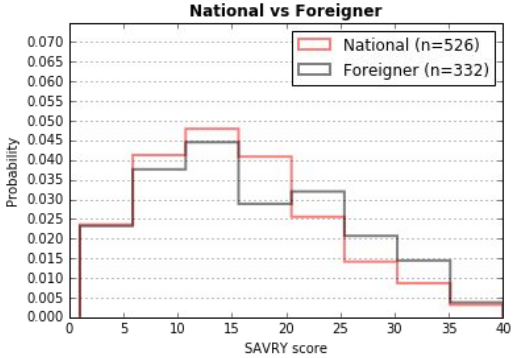


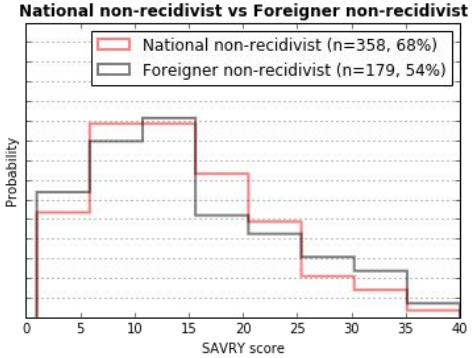


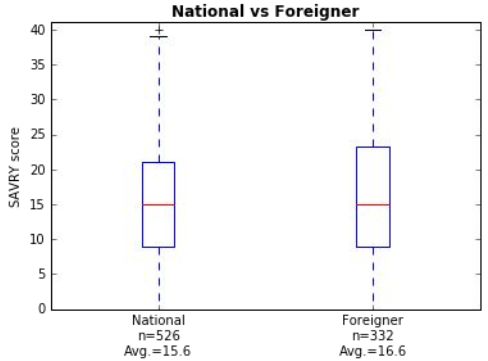
ProPublica 2016 https://github.com/propublica/compas-analysis

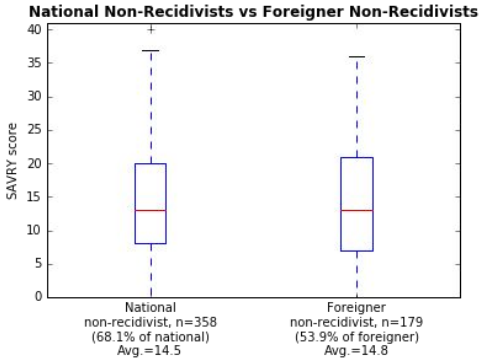
No disparate mistreatment / E.O.











Centre d'Estudis Jurídics i Formació Especialitzada (CEFJE) 2017: http://cejfe.gencat.cat/ca/recerca/cataleg/crono/2017/reincidencia-jj/

Unfair rankings

	Position										top 10 male	top 10 female	top 40 male	top 40 female
	1	2	3	4	5	6	7	8	9	10				
Economist	f	m	m	m	m	m	m	m	m	m	90%	10%	73%	27%
Market analyst	f	m	f	f	f	f	f	m	f	f	20%	80%	43%	57%
Copywriter	m	m	m	m	m	m	f	m	m	m	90%	10%	73%	27%

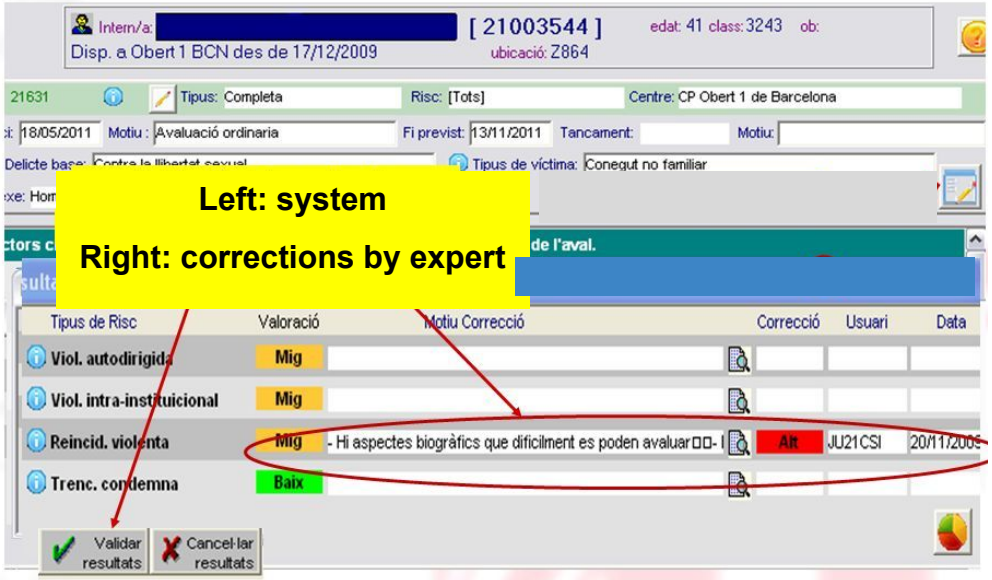
Top-10 results for job searches in XING (a recruitment site similar to LinkedIn), for selected professions.

Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., & Baeza-Yates, R. (2017, November). FA* IR: A fair top-k ranking algorithm. In CIKM.

Algorithm-human interaction

Humans need to be able to receive explanations, and to correct outcomes.

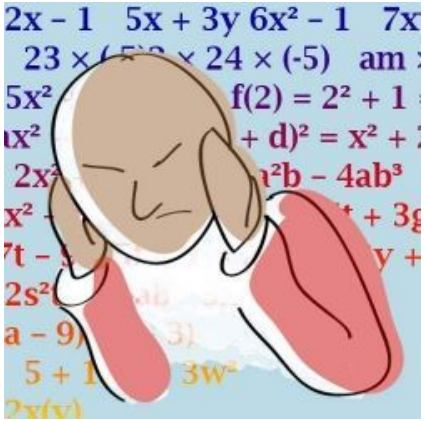
Effective transparency does not mean source code, it means the human can understand and challenge the algorithmic decisions.



RISCANVI, 2013 (Generalitat Cat.)
http://slideplayer.es/slide/7242758/

A personal opinion on transparency ...

Many "customers" of algorithmic decision making systems do not value transparency. They want certainty, not doubts.



This is a perverse incentive for developers/providers

Numbers, plots, charts, suggest objectivity
Lack of math literacy becomes problematic

We need good evaluation frameworks

How can this be fixed? Improving math literacy and using evaluation frameworks that integrate multiple dimensions

In addition to accuracy: "dollars saved, lives preserved, time conserved, effort reduced, quality of living increased" and respect to privacy, fairness, accountability, transparency

Wagstaff, K. (2012). Machine learning that matters. arXiv:1206.4656.

Conclusion

To the extent that algorithms can
engage in disadvantageous differential treatment
that leaves people of a socially salient group worse-off,
based on statistical information
… algorithms can discriminate

Current research looks at trade-offs of utility and fairness
and at mechanisms for mitigating unfairness

carlos.castillo@upf.edu