


OXFORD
MARTIN
SCHOOL




erc
European Research Council
Established by the European Commission

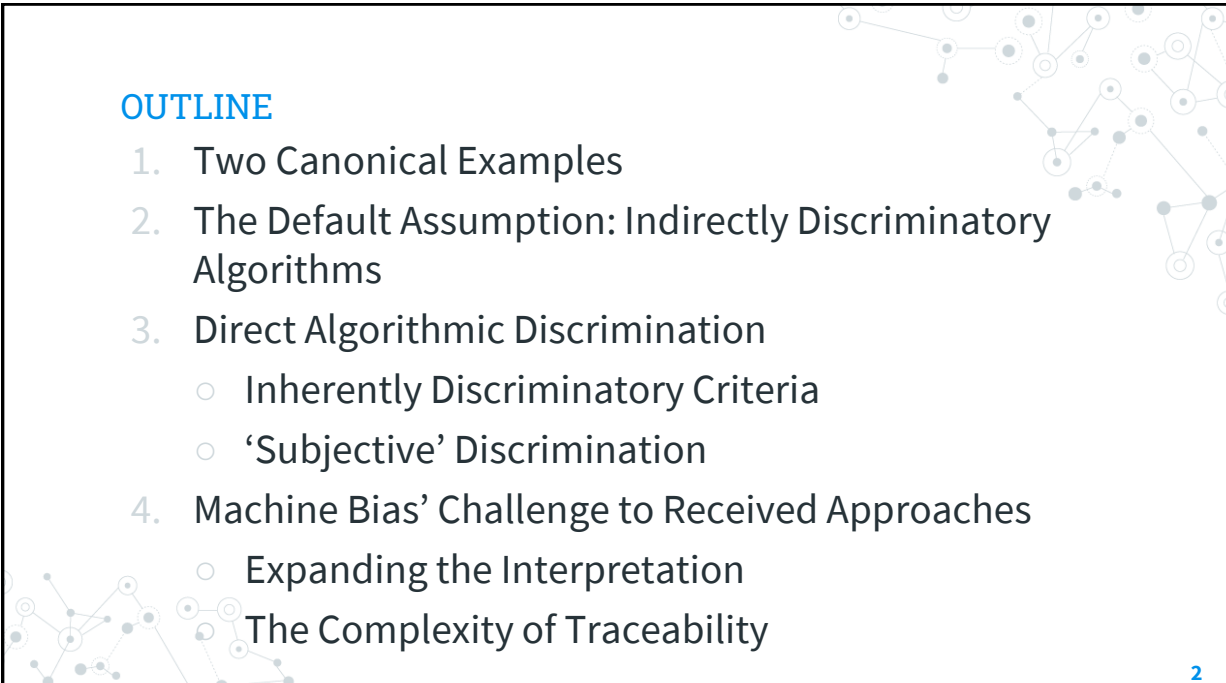
LEVERHULME
TRUST

Directly Discriminatory Algorithms

Jeremias Adams-Prassl, Reuben Binns, Aislinn Kelly-Lyth

 This training session is funded under the 'Rights, Equality and Citizenship Programme 2014-2020' of the European Commission.

1



OUTLINE

1. Two Canonical Examples
2. The Default Assumption: Indirectly Discriminatory Algorithms
3. Direct Algorithmic Discrimination
 - Inherently Discriminatory Criteria
 - 'Subjective' Discrimination
4. Machine Bias' Challenge to Received Approaches
 - Expanding the Interpretation

The Complexity of Traceability

2

2

Algorithmic discrimination: two canonical examples

Amazon's hiring algorithm

- ML algorithm looked for patterns between successful software engineering applicants
- Began to use gender indicators to predict success
- Example of **proxy discrimination**

In effect, Amazon's system taught itself that male candidates were preferable. It penalized resumes that included the word "women's," as in "women's chess club captain." And it downgraded graduates of two all-women's colleges, according to people familiar with the matter. They did not specify the names of the schools.

Amazon edited the programs to make them neutral to these particular terms. But that was no guarantee that the machines would not devise other ways of sorting candidates that could prove discriminatory, the people said.

3

3

Algorithmic discrimination: two canonical examples

Gender Shades

- Commercial gender classification systems were much worse at recognising darker-skinned females than lighter-skinned males (max error rates of 34.7% and 0.8% respectively)
- Example of **sampling bias**

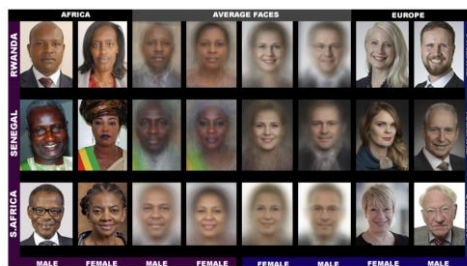


Figure 1: Example images and average faces from the new Pilot Parliaments Benchmark (PPB). As the examples show, the images are constrained with relatively little variation in pose. The subjects are composed of male and female parliamentarians from 6 countries. On average, Senegalese subjects are the darkest skinned while those from Finland and Iceland are the lightest skinned.

Buolamwini & Gebru, 2018

4

4

1.

The Default Assumption

Indirectly Discriminatory Algorithms

5

5

Algorithmic Indirect Discrimination

- ⊙ Algorithmic decision-making systems (ADMS) discriminate: extensive empirical evidence
- ⊙ How can discrimination law tackle this?
 - ⊙ US – Barocas & Selbst (2016)
 - ⊙ EU – Hacker (2018)
- ⊙ ADMS as a PCP
- ⊙ Disparate impact on protected group

6

6

“

*... direct discrimination does not cover **indirect proxy discrimination**... does not capture **sampling bias** and **incorrect labelling** unless these practices directly relate to class membership... in most cases in which bias is an accidental feature of the data processing, unfavourable treatment arguably does not occur “on grounds of” group membership and therefore **does not amount to direct discrimination**.*

Hacker (2018)

7

7

An Incomplete Assumption

1. Distinct approach to direct discrimination in EU / UK law
2. Scope of justification
3. Normative classification

8

8

Direct Discrimination	Disparate Treatment	Indirect Discrimination	Disparate Impact
<p>“One person is treated less favourably than another is, has been or would be treated in a comparable situation, on grounds of a protected characteristic.”</p> <ul style="list-style-type: none"> • No need for intention. • Covers implicit bias. 	<p>A person is treated differently because of their protected characteristic.</p> <ul style="list-style-type: none"> • Requires intention. • Contested whether this can include implicit bias. 	<p>Application of an “apparently neutral provision, criterion, or practice which would put persons with a protected characteristic at a particular disadvantage compared with other persons”.</p>	<p>Application of a facially neutral policy which causes a disparate impact on the basis of a protected characteristic.</p>
<p>Very narrow scope for justification; only possible if the protected characteristic is a genuine occupational requirement.</p>	<p>Very narrow scope for justification, e.g. if the protected class is a bona fide occupational qualification.</p>	<p>May be objectively justified as a proportionate means of achieving a legitimate aim.</p>	<p>May be justified by “business necessity”.</p>

9

9

Direct Discrimination	Disparate Treatment	Indirect Discrimination	Disparate Impact
<p>“One person is treated less favourably than another is, has been or would be treated in a comparable situation, on grounds of a protected characteristic.”</p> <ul style="list-style-type: none"> • No need for intention. • Covers implicit bias. 	<p>A person is treated differently because of their protected characteristic.</p> <ul style="list-style-type: none"> • Requires intention. • Contested whether this can include implicit bias. 	<p>Application of an “apparently neutral provision, criterion, or practice which would put persons with a protected characteristic at a particular disadvantage compared with other persons”.</p>	<p>Application of a facially neutral policy which causes a disparate impact on the basis of a protected characteristic.</p>
<p>Very narrow scope for justification; only possible if the protected characteristic is a genuine occupational requirement.</p>	<p>Very narrow scope for justification, e.g. if the protected class is a bona fide occupational qualification.</p>	<p>May be objectively justified as a proportionate means of achieving a legitimate aim.</p>	<p>May be justified by “business necessity”.</p>

10

10

Self-justifying feedback loops

- ⊙ €€€ considerations not in themselves enough, but...
 - ⊙ Example: predictive policing (Ensign et al, 2018)
 - ⊙ ADMS as a facially neutral PCP → use objectively justified?
 - ⊙ Prediction of more crime borne out by higher numbers of arrests in data
 - ⊙ If justification is permitted, feedback loop created
- Judicial interpretation could remedy

11

11

Distinguishing the concepts in the case law?

Direct Discrimination

- ⊙ Formal equality
- ⊙ Reason-focussed

Indirect Discrimination

- ⊙ Aimed at advancing substantive equality
- ⊙ Effects-focussed

NB: “[the] distinction between direct and indirect discrimination is hard to draw on a conceptual level... It is therefore not possible to provide a consistent and precise division between the legal categories” (Collins and Khaitan, 2018)

12

12

Returning to our: examples: putative DD?

Amazon's hiring algorithm Gender Shades

- ⊙ Formal Inequality: likes treated differently
- ⊙ Protected characteristics play a role in outcome – a 'reason' for decision

13

13

2. Direct Algorithmic Discrimination

14

14

Direct Discrimination

JFS (2009)

*“Direct discrimination can arise in one of two ways: because a decision or action was taken on a ground which was, however worthy or benign the motive, **inherently racial** within the meaning of s.1(1)(a), or because it was taken or undertaken for a reason which was **subjectively racial**”.*

Lord Phillips [78]

15

15

Inherently Discriminatory Criteria

- ⊙ *James* (UK, 1990)
- ⊙ *WABE* (EU, 2021)



16

16

Inherently Discriminatory Algorithms

- ⊙ A mortgage assessment tool programmed to consider marital status
- ⊙ Amazon's recruitment algorithm, which learned to penalise graduates of two all-women's colleges
- ⊙ Latent variable proxy: *"a machine learning tool could 'learn a perfect proxy for race... such that including race over and above this combination would have no effect on risk classifications"* (Davies and Douglas, 2020)

17

17

Subjective Discrimination

Lady Hale: *"the discriminator may... unconsciously be making his selections on the basis of race or sex. He may not realise that he is doing so, but that is what he is in fact doing."*

JFS [64]

18

18

'Subjective' Algorithmic Discrimination

- ⊙ ML algorithm as an automated version of human implicit bias
- ⊙ Example: recruitment algorithm which learns to use gender indicators (sports, language usage)
- ⊙ Legal analysis shouldn't be different because treatment is meted out by an algorithm

19

19

The Risk of Black Box Bias

Prima facie finding of direct discrimination
→ burden to disprove

20

20

3.

Machine Bias' Challenge to Received Approaches

When is an outcome 'because of' a protected characteristic?

21

21

Moving Beyond the Two Categories?

- ⊙ Amazon example fits into existing caselaw
- ⊙ What about Gender Shades?
 - ⊙ Sampling data unrepresentative – not necessarily because of any subjective discrimination
 - ⊙ No inherently discriminatory criterion
- ⊙ A need to evolve the judicial interpretation of 'because of'?

22

22

A Stricter Standard for ADMS?

- ⊙ Factors more traceable, but standards should be the same
- ⊙ Ask whether protected characteristic had a ‘significant influence’ on the outcome (*Nagarajan 1990*)
- ⊙ Computer science approaches:
 - ⊙ SHAP methods
 - ⊙ LIME-based explanations

23

23

Applying Existing Law

- ⊙ Gender Shades in the human context
 - ⊙ No scope for rationalisation of implicit biases
 - ⊙ Human decision-makers can engage in *ex post facto* rationalisation
- ⊙ In Europe, similarly qualified applicants from immigrant backgrounds have to send c. 30% more job applications than ‘majority’ applicants to get a similar success rate (GEMM Project 2018)
- ⊙ Not reflected in the caselaw: it is “*unusual to find direct evidence of racial discrimination*” because “*few [decision-makers] will be prepared to admit such discrimination even to themselves*” (King 1991)

24

24

The Complexity of Traceability

- ⦿ Quantifying and tracing the role of protected characteristic: a dangerous blessing
- ⦿ Protected characteristics will often have played some upstream role
- ⦿ Example: tainted information (*Reynolds* 2015)
 - ⦿ Poor reference from a biased ex-employer
 - ⦿ Knowledge as a tacit consideration?

25

25

“

*Much computational research on fairness is built on frameworks borrowed from discrimination law ... perhaps most crucially, the belief that fairness can be achieved by simply altering how we assess people at discrete moments of decision-making... **exposing the limits of algorithmic notions of fairness has exposed the limits of the underlying legal and philosophical notions of discrimination***

26

Abebe et al (2020)

26

Thanks!

Jeremias Adams-Prassl
@Jeremias Prassl

Aislinn Kelly-Lyth
@LawAislinn

Reuben Binns
@RDBinns



Presentation template by [SlidesCarnival](#)

27