

Digital Inequalities

2025 Elisa Rubegni and Emily Winter

This lecture – learning outcomes

- To understand how digital technologies (particularly AI algorithms) may reinforce existing inequalities
- To consider possible responses to challenges inequalities generated by digital technology e.g. algorithmic auditing, fairness metrics

Understanding digital inequality

Unequal access:

e.g., digital exclusion was covered in-depth in week 15

Unequal outcomes:

e.g., those subject to algorithmic decision-making, etc. it is the main focus of today's lecture

Understanding digital inequality: Unequal outcomes

- Artificial intelligence (AI) and machine-learning tools promise of efficiency make algorithmic systems attractive
- This leads to complex social issues being increasingly automated, creating a false sense of solution and safety.
- Algorithmic decision making increasingly pervades the social sphere having a great impact on medical care, to predicting crimes, selecting social welfare beneficiaries, and identifying suitable job candidates

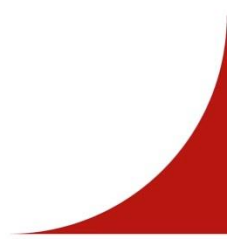
Understanding digital inequality: Unequal outcomes

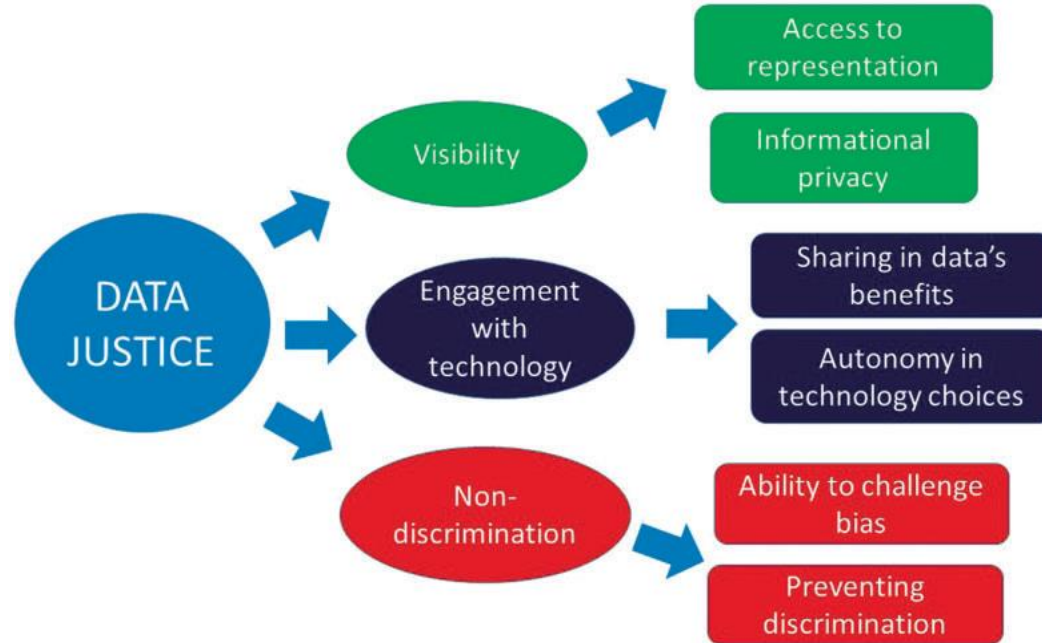
- Computer scientists are not simply dealing with purely technical aspects but are engaged in making moral and ethical decisions that impact on people's life
- The harm, bias, and injustice that emerge from algorithmic systems varies and is dependent on the training and validation data used, the underlying design assumptions, and the specific context in which the system is deployed
- As results it impacts individuals and communities that are at the margins of society

Unequal outcomes

- For challenging the power asymmetries and structural inequalities which are engrained into the society
- A shift from asking “how can we make a certain dataset representative?”
- To be focused on “what is the product or tool being used for? Who benefits? Who is harmed?”
- The idea of focusing on the people disproportionately impacted is aligned with participatory design and human-centred design which are based on the concept that the design process as a fundamentally participatory process

Data Justice

- Data justice concerns the ways in which (big) data systems can discriminate, discipline and control
 - The use of data for governance to support power asymmetries (Johnson, 2014),
 - The way data technologies can provide greater distributive justice through making the poor visible (Heeks and Renken, 2016)
 - The impact of dataveillance practice on the work of social justice organisations (Dencik et al., 2016).
- 



Taylor, Linnet. "What is data justice? The case for connecting digital rights and freedoms globally." *Big Data & Society* 4.2 (2017): 2053951717736335.

What is it about the social impacts of digital data that suggests a social justice agenda is important?

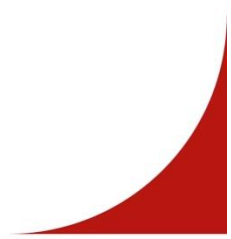
As the social impacts of big data are very different depending on:

- one's socio-economic position,
e.g. data-driven law enforcement focuses unequally on poor neighbourhoods which experience certain types of criminality (O'Neil, 2016);
- gender, ethnicity and place of origin,
e.g. transgender citizens have been dealt with by population databases in the US shows that one's ability to legally identify as a different gender depends to a great extent on one's income (Moore and Currah's 2015)

Algorithmic Social Justice

Algorithmic Social Justice addresses how AI-driven systems reinforce, mitigate, or reshape social inequalities.

For instance, digital content produced by generative AI has a propensity to reinforce prevalent stereotypes by representing people in traditional, normative gender roles and appearances which can significantly impact people's perceptions and attitudes about themselves and others (Metaxa et al., 2021)



Algorithmic bias examples



Case study 1: Algorithmic bias in medicine and health

AI's use in medicine

- AI and ML increasingly used in medicine and healthcare, often for diagnostic purposes and offers many advantages
- Quick diagnosis:
 - Often very accurate (but not for everyone, as we will see...)
 - Can speed up processes in under-pressure healthcare services

Racial bias example: skin cancer identification

AI - used to distinguish between images of malignant and benign skin lesions

Existing datasets (e.g., International Skin Imaging Collaboration) – images largely from US, Europe and Australia, featuring a majority of lighter skin tones
→ models are less effective on darker skin tones

Key issue: lack of diversity in the training data – data gap

Gender bias example: liver disease screening

- AI used to screen for liver disease from blood test results
- Study found that the AI models missed 44% of cases of liver disease in women (compared with 23% of men)
- The 2 most accurate modules (in terms of overall accuracy) had the largest gender gaps

Source: www.ucl.ac.uk/news/2022/jul/gender-bias-revealed-ai-tools-screening-liver-disease

Gender bias example: liver disease screening

Dr Isabel Straw:

“When we hear of an algorithm that is more than 90% accurate at identifying disease, we need to ask: accurate for who? High accuracy overall may hide poor performance for some groups.”

<https://www.ucl.ac.uk/news/2022/jul/gender-bias-revealed-ai-tools-screening-liver-disease>



Gender bias example: liver disease screening

Why does the gender gap in accuracy exist?

- Liver disease AI models – use biochemical markers of disease (such as lower albumin levels) which are more indicative of the disease in men than they are in women
- Women were already less likely to be successfully diagnosed with liver disease
- Not a new thing - reflects historic and current gender inequalities in clinical practice – e.g., women under-represented in clinical trials, unconscious bias, men as ‘default’
- The use of these models represents ‘the digitisation of inequalities into algorithmic systems’

Case study 2: Algorithmic bias in criminal justice



2a) Predictive policing

What is predictive policing?

- Use of historic crime data (e.g., arrest numbers) to determine how to allocate police geographically
- Model updated with new data of observed crime

What is the problem?

- Bias in what data is recorded, e.g., where police are originally patrolling (which is influenced by bias, such as consideration of race and ethnicity) certain neighbourhoods over-represented in data
- Police records not an exact measure of true crime rates
- Predictive policing does what it says: predicts future policing patterns more than it does crime

Kristian Lum, William Isaac 'To Predict and Serve?', Significance, Volume 13, Issue 5, October 2016, Pages 14–19, <https://doi.org/10.1111/j.1740-9713.2016.00960.x>

What is the problem?

Statistical flaws:

- Model susceptible to runaway feedback loops, where police are repeatedly sent back to the same neighbourhood
- Crime data from this neighbourhood fed into model training
- Self-reinforcing: minimal crime data for areas that are not regularly patrolled (but this doesn't indicate a lack of crime)
- Traditional batch learning frameworks inappropriate and not a correct indication of true crime rates

Social consequences

Study by Lum and Isaac found that Oakland Police targeted two specific neighbourhoods for policing efforts related to drug crime.

These neighbourhoods were both low-income and inhabited mostly by racialized communities

- Focusing on drug-related crime, they estimate that such crimes are in fact much more evenly distributed across the city
- Disproportionate policing of particular neighbourhoods: high arrests, criminal records, etc.

2b) Predicting recidivism (criminal reoffending)



Predicting recidivism

- Most well known: Northpointe's COMPAS tool (Correctional Offender Management Profiling for Alternative Sanctions)
- COMPAS gives people who have been arrested a 'risk score' (from 1 [lowest risk] to 10 [highest risk]) that predicts a person's likelihood to reoffend within 2 years
- Score based on the defendant's answers to 137 questions
- Risk scores help determine who is incarcerated and for how long – used in pretrial (e.g., bail conditions), parole and sentencing decisions

Why are these algorithms used?

- Argument made that will be accurate and less biased than human decision-making
- To boost efficiency
- To aid resource allocation

Criticisms of COMPAS

- ProPublica investigated the 2013/14 use of the COMPAS algorithm in Broward County, Florida, comparing the risk scores with actual rates of recidivism within 2 years
- ProPublica's analysis found that the algorithm was more likely to wrongly label black defendants as high risk and more likely to wrongly label white defendants as low risk.



What did ProPublica find?

- The score correctly predicted recidivism in 61% of cases
- The algorithm correctly predicted recidivism for both black and white defendants at approx. same rate
- BUT for defendants who did not reoffend, black defendants were nearly twice as likely to have been misclassified as high risk compared to white defendants
- For defendants who did reoffend, white defendants were nearly twice as likely to have been misclassified as low risk compared to black defendants

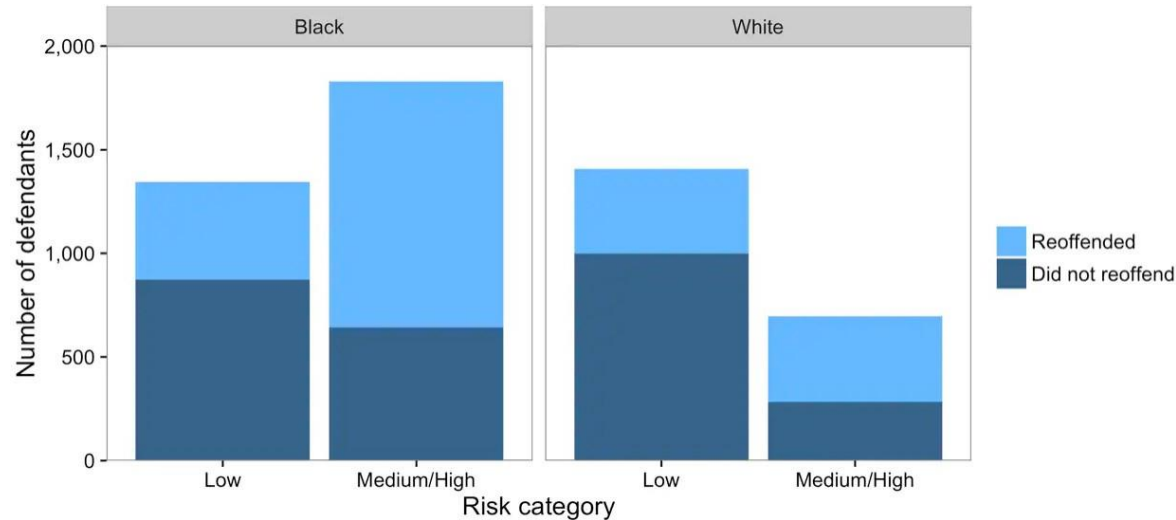
Northpointe response

Northpointe response: argued that COMPAS was fair. Why?

Predictive parity: likelihood of recidivism among offenders deemed high risk is the same regardless of race

Accuracy equity: can discriminate between recidivists and non-recidivists equally well regardless of race

Calibration: the likelihood of recidivism for any score is the same regardless of race



Taken from: <https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/>

Other analyses of COMPAS

- Dressel and Farid 2018 – found that COMPAS was no more accurate or fair than predictions made by humans with no criminal justice expertise
- Rudin et al. 2020 – transparency is a better goal than fairness (due to competing – and incompatible – definitions of fairness).

What's at stake here?

- Different understandings of fairness
- Can't satisfy all definitions of fairness simultaneously
- Fairness vs. predictive accuracy
- Bigger question of whether decisions about individuals should be guided by demographic patterns
- Need to understand this in context of broader criminal justice landscape (e.g., use of predictive policing, etc.)

Responses to algorithmic bias: Algorithmic Auditing

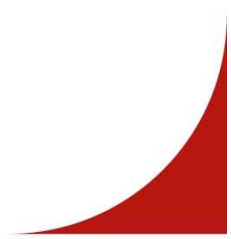


Algorithmic auditing

What is Algorithmic Auditing?

- A method to analyse AI models by repeatedly querying them.
- Helps detect biases or unintended behaviours in AI systems.

Why is it important?

- AI models, especially generative AI, function as 'black boxes'.
 - Identifies potential societal biases embedded in training data.
- 

Methodologies in Algorithmic Auditing

Expert-led audits

- Conducted by researchers or dedicated auditing teams.
- Methods include:
 - Data scraping
 - Creation of synthetic user profiles (sock puppets)

Example:

- Detecting bias in AI-generated images favouring certain demographics

Expansion to Non-Expert Auditing

Involving Everyday Users

- Expands auditing beyond experts
- Engages individuals directly impacted by AI models

Youth Participation

- Frequent users of AI technologies
- Offer valuable insights into biases affecting younger audiences

Effectiveness:

- Users identify biases that resonate with their lived experiences
- 

Auditing and Participatory Approaches

Why Include Non-Experts?

- Experts may overlook biases not present in their lived experiences.
- Everyday users can highlight real-world impacts of AI bias.

Participatory research techniques

- Improve AI fairness and accountability.

The Importance of Auditing

- Algorithmic auditing is essential to detect and mitigate AI biases
- Including non-experts, particularly youth, offers unique and impactful insights
- Algorithmic auditing aims at expanding AI fairness

Responses to algorithmic bias: technical/statistical responses – fairness metrics

With many thanks to Dr James Grant (Department of Mathematics and Statistics, Lancaster University) for providing the fairness metric definitions and graphs that have been adapted for use in this lecture ©

Fairness metrics

- Narayanan (2018) – identifies over 20 mathematical definitions of fairness
- IBM AIF360 toolkit – an open-source toolkit that includes over 70 different bias detection metrics
- Challenges:
 - different definitions of fairness;
 - how to choose between different fairness metrics;
 - interaction between fairness and accuracy

See:

- <https://github.com/Trusted-AI/AIF360> and resources)
- <https://www.youtube.com/watch?v=DBqRmvXwwUA> (IBM
- You can experiment with different fairness metrics on sample data sets here: <https://aif360.res.ibm.com/data>

Fairness metric example: statistical parity difference

This measures fairness by focussing on the rate of positive outcomes in two groups.

If an algorithm assigns positive outcomes to members of the two groups at equal rates, then it is judged to be fair.

$$\text{Statistical Parity Difference} = \text{Probability of an individual in Group A receiving a positive outcome} - \text{Probability of an individual in Group B receiving a positive outcome}$$

The closer to 0 the statistical parity difference (whether positive or negative), the fairer.

Fairness metric example: equal opportunity difference

- This measures fairness by focusing on the rate of true positive outcomes in two groups—assessing whether those who should receive a positive outcome do.
- If an algorithm assigns positive outcomes to members of the two groups who are suited to positive outcomes at equal rates, then it is judged to be fair.

$$\text{Equal Opportunity Difference} = \text{Probability of an individual in Group A who should get a positive outcome receiving one} - \text{Probability of an individual in Group B who should get a positive outcome receiving one}$$

The closer to 0 the equal opportunity difference (whether positive or negative), the fairer.

Does it matter?

Yes, because different fairness metrics will have different outcomes...

Let's explore this with a very simple mocked-up example...



Scenario

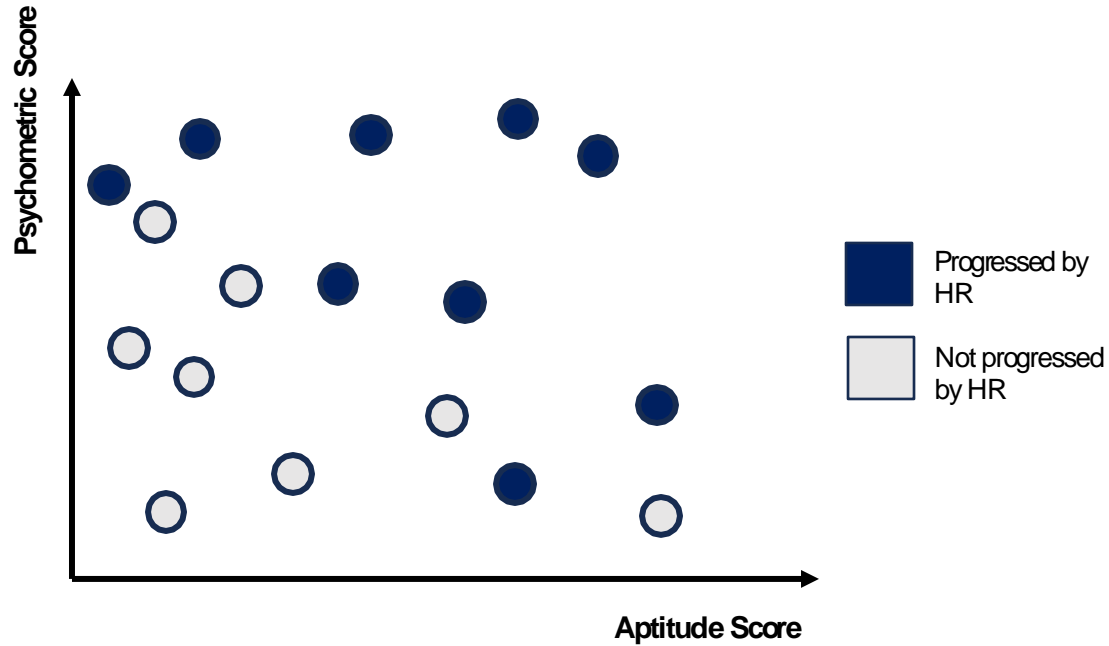
A large company is looking to hire a new communications director.

The HR team are excellent at identifying suitable candidates but would like to explore using an algorithm to speed up the process of selecting candidates to progress to the next stage.

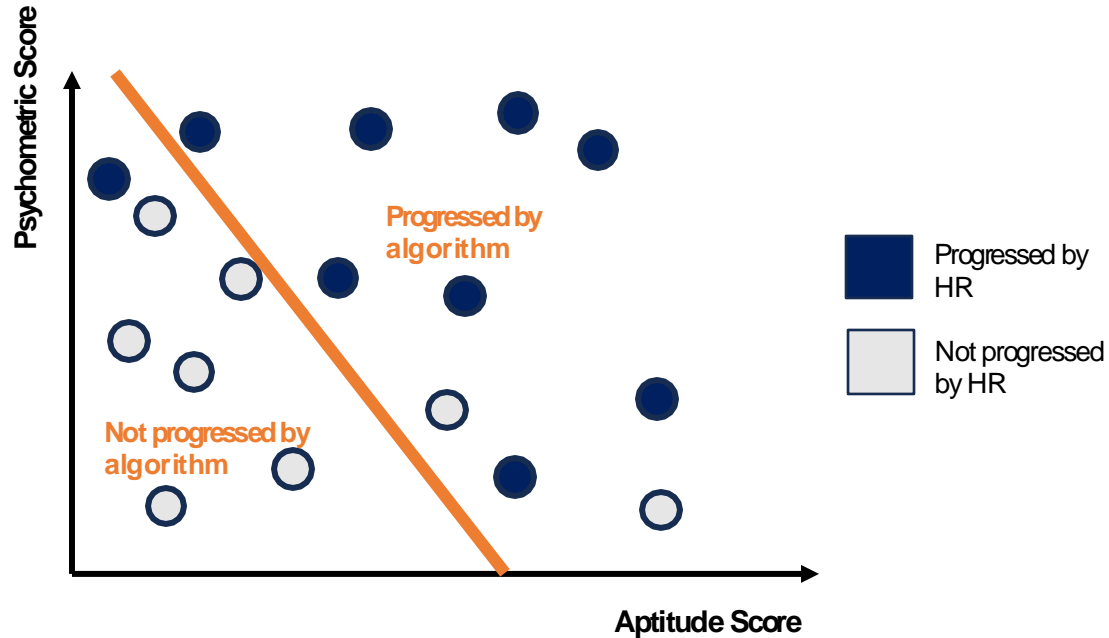
The algorithm has been trained on previous applicant data including whether they were deemed suitable to progress or not (decision made by HR team).

To make the decision about candidate suitability, the algorithm uses a psychometric score, and an aptitude score derived from an initial screening test.

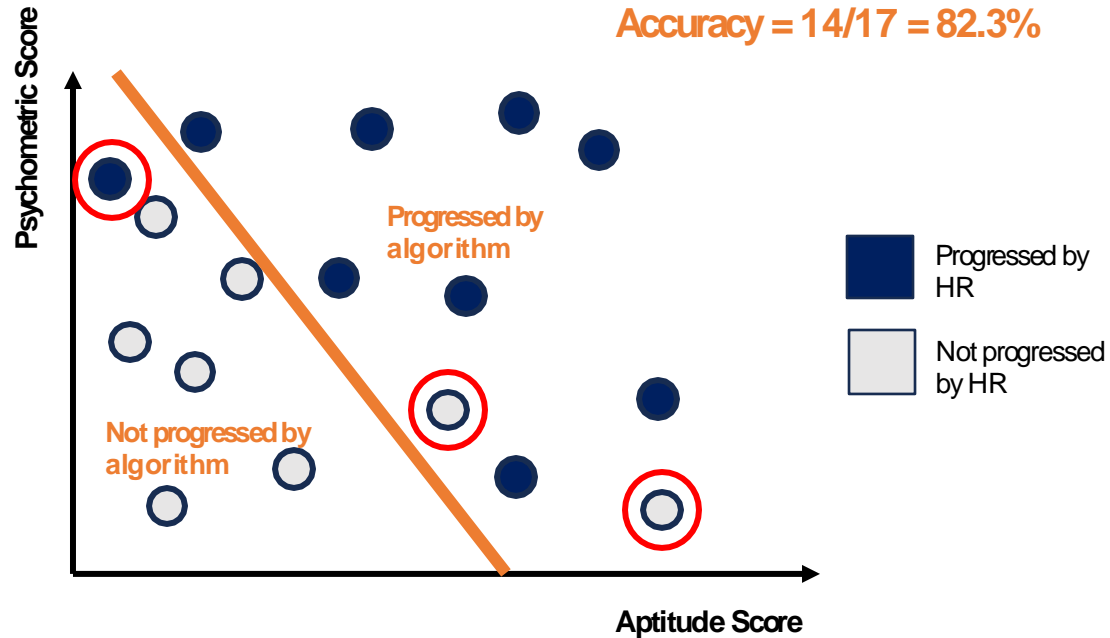
Historical candidate data



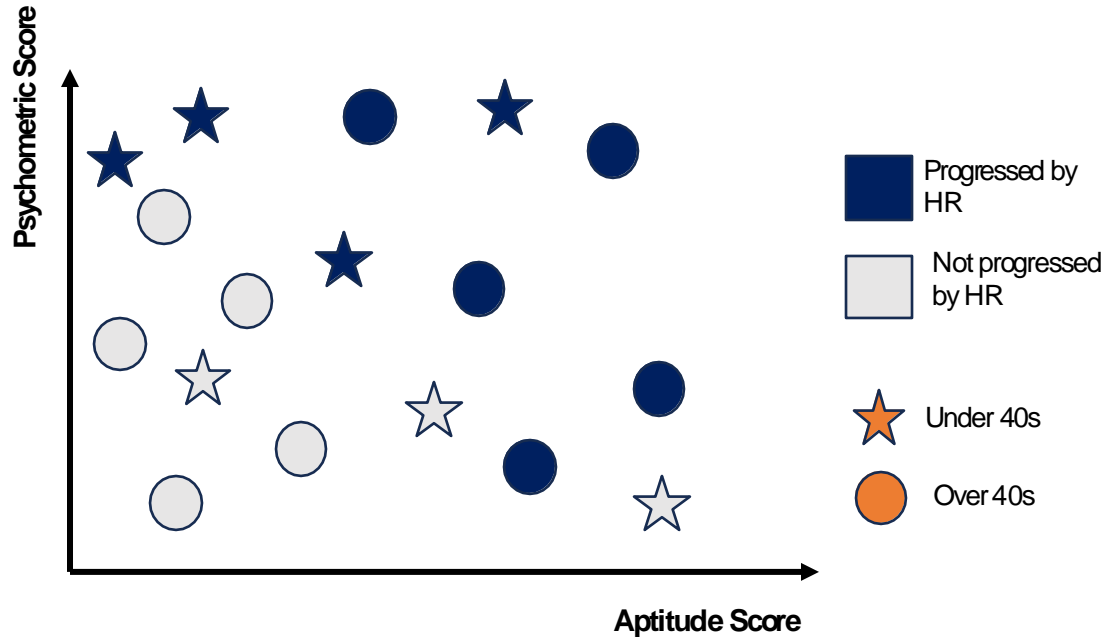
Algorithm Illustration



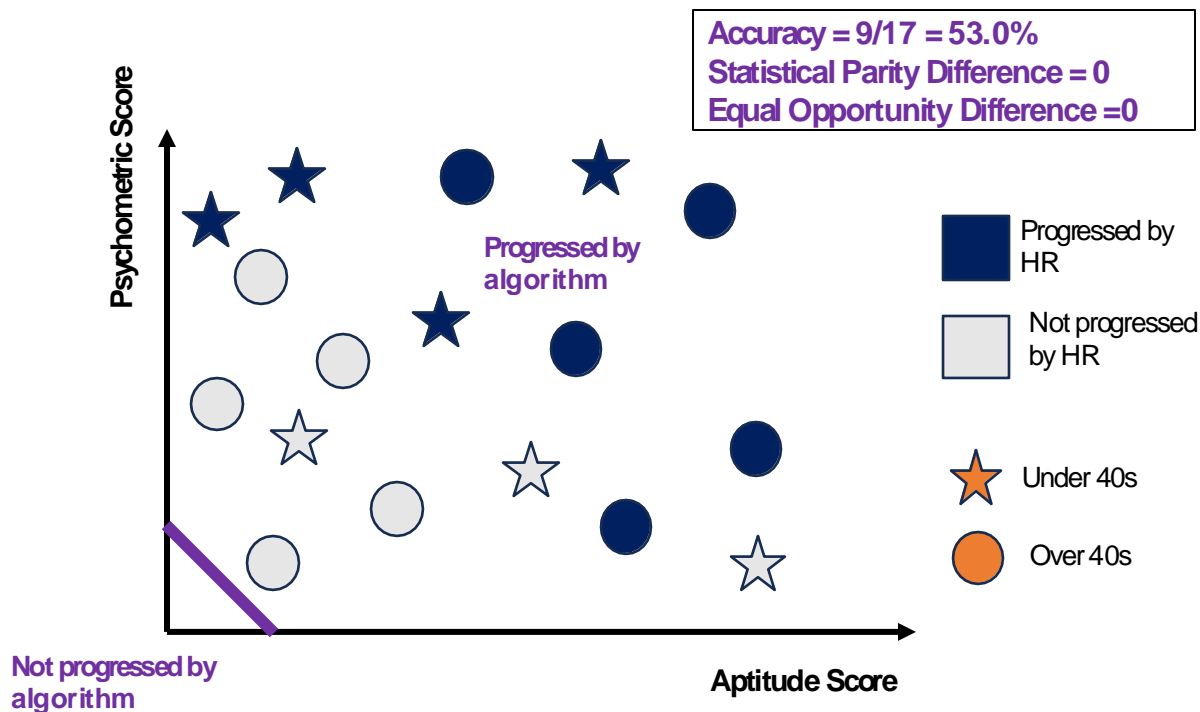
Algorithm Illustration



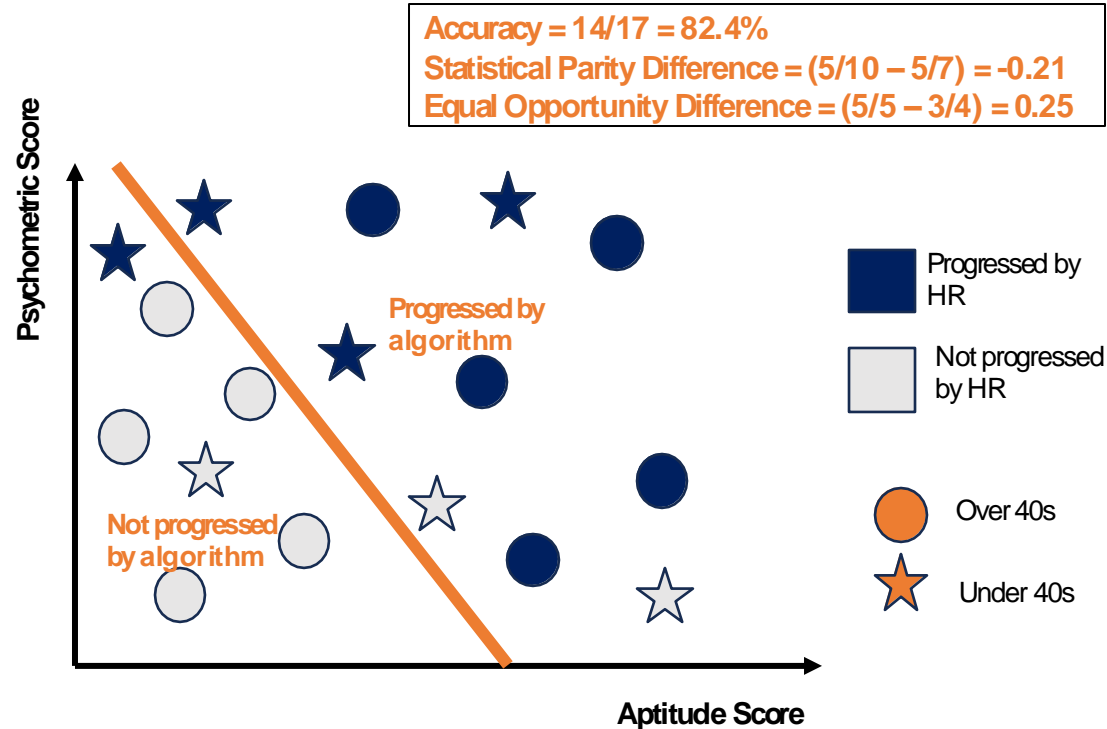
Bias Illustration



Optimising for fairness?



Bias Illustration



How to choose a fairness metric?

- Lots to consider!
- How critical is the outcome?
- Some situations in which it is very important to eliminate false negatives are likely to be very important
- Understanding the specifics of the context and the factors that affect discrimination and injustice for different protected characteristics

For more information, see: <https://doi.org/10.1145/3585006>



Sources and further reading

- Kasirzadeh, Atoosa. "Algorithmic fairness and structural injustice: Insights from feminist political philosophy." *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 2022.
- ProPublica case study sources: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- On methodology/analysis: <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>
- More in-depth info on data/analysis: <https://github.com/propublica/compas-analysis/blob/master/Compas%20Analysis.ipynb>
- Julia Dressel and Hany Farid, The accuracy, fairness, and limits of predicting recidivism. *Sci Adv*.4, eaao5580(2018). DOI:10.1126/sciadv.aao5580
- Rudin, C., Wang, C., & Coker, B. (2020). The Age of Secrecy and Unfairness in Recidivism Prediction. <https://doi.org/10.1162/99608f92.6ed64b30>
- Algorithmic Fairness and Structural Injustice: Insights from Feminist Political Philosophy. <https://dl.acm.org/doi/epdf/10.1145/3514094.3534188>

Sources and further reading

Video about IBM's AI Fairness tools:

https://www.youtube.com/watch?v=1RptHwfkx_k

Joy Buolamwini video exploring how AI misidentifies famous black women as men (performance poetry!):

<https://www.youtube.com/watch?v=QxuyfWoVV98>

Thank you for attending, any questions?