# Fairness in Machine Learning

## Why, How, and What's Next?

Simon Caton
simon.caton@ucd.ie

School of Computer Science, University College Dublin, Ireland

# Background

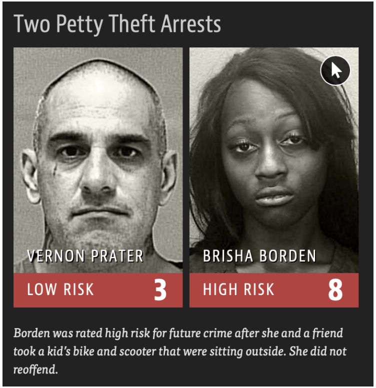
A lot of this presentation is based on:

[CH24]: Simon Caton and Christian Haas. Fairness in Machine Learning: A Survey. ACM Computing Surveys, Vol. 56 (7). 2024. Open Access Link

It was written as an entry point to the area for researchers and practitioners not familiar with Fairness in Machine Learning.

# Fairness in Machine Learning / AI — Why?

# Why? AI-driven Recidivism Risk Calculations





https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

# Why? AI-based Judgement

> **A beauty contest was judged by AI and the robots didn't like dark skin**
>
> The first international beauty contest decided by an algorithm has sparked controversy after the results revealed one glaring factor linking the winners

https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people

# Why We Should Care

Machine Learning (ML) and AI are being used A LOT in scenarios that **involve people**.

It's easy to (un)intentionally learn from the "wrong parts" of the data.

# Then what is “fair”?

# Then what is “fair”?

Decision makers may not be able to define what it means to be “fair” but they may recognize “unfairness” when they see it [GJ$^+$18].

# Then what is “fair”?

Decision makers may not be able to define what it means to be “fair” but they may recognize “unfairness” when they see it [GJ+18].

Your personal definition is likely rooted in your own world view (sociocultural norms, political view point, religion etc.) and demographics (see e.g.: [Pie17]).

This makes fairness very hard to define universally.

# Fairness in ML / AI — How?
# $\rightarrow$ Measurement and Intervention

# Measuring Fairness: Sensitive Variables (I)

Sensitive Variable: any data (or feature of the data) that refers or relates to humans [BHN19].

Specific legal frameworks also provide concrete sets.

**Some are very obvious**

age, gender, race, marital status, sexuality, religion . . .

Others are not; they are correlated with sensitive variables. These are proxy-sensitive variables (or proxies or quasi-identifiers).

# Measuring Fairness: Sensitive Variables (II)

| Sensitive Variable | Example Proxies |
|---|---|
| Gender | Education Level, Income, Occupation, Felony Data, Keywords in User Generated Content (e.g. CV, Social Media etc.), University Faculty, Working Hours |
| Marital Status | Education Level, Income |
| Race | Felony Data, Keywords in User Generated Content (e.g. CV, Social Media etc.), Zipcode |
| Disabilities | Personality Test Data (inferrable from social media posts) |
| Immigration Status | Social Media Posts |

See: [BC+17, p. 1014], and [FRD18, Ber19, SHS19, BC+17, Sel17, HC17, Yar10, SE+13, WD14, MD93, KYJ+23]

# Measuring Fairness: (Un)privileged Groups

Using some set of sensitive variables, we can define:

- Privileged Group(s): disproportionately **more** likely to be positively classified (handled)
- Unprivileged Group(s): disproportionately **less** likely to be positively classified (handled)

### Deriving a Fairness Metric

Sensitive Variable(s) + (Un)privileged Group(s) $\rightarrow$ Fairness Metric

# Measuring Fairness: Metrics (there are lots!)

| | Group-based Fairness | | | | Individual and Counterfactual Fairness |
|---|---|---|---|---|---|
| | Parity-based Metrics | Confusion Matrix-based Metrics | Calibration-based Metrics | Score-based Metrics | Distribution-based Metrics |
| Concept | Compare predicted positive rates across groups | Compare groups by taking into account potential underlying differences between groups | Compare based on predicted probability rates (scores) | Compare based on expected scores | Calculate distributions based on individual classification outcomes |
| Abstract Criterion | Independence | Separation | Sufficiency | - | - |
| Examples | Statistical Parity, Disparate Impact | Accuracy equality, Equalized Odds, Equal Opportunity | Test fairness, Well calibration | Balance for positive and negative class, Bayesian Fairness | Counterfactual Fairness, Generalized Entropy Index |

Can you spot the dilemma of being able to “measure” something?

# Implications of Measurement

Being able to mathematically define fairness is key to technical approaches and interventions.

BUT the idea of measurement can be precarious as it implies:

- a straightforward process [BHN19], which it isn't
- responsible use, which it might not be (see EU AI Act)

Being “fair” is not doing whatever we did last time

# Doing “something” is easy, but... (I)

- this is is going to be hard
- doing “something” might be worse than doing nothing
- expect performance (e.g. accuracy) to worsen (this might be the desired outcome)

**Suppose**

Simon went away to “look at” the data. Wait! There’s a gender variable, I’ll remove it! Situation resolved, right!?

# Doing "something" is easy, but... (II)

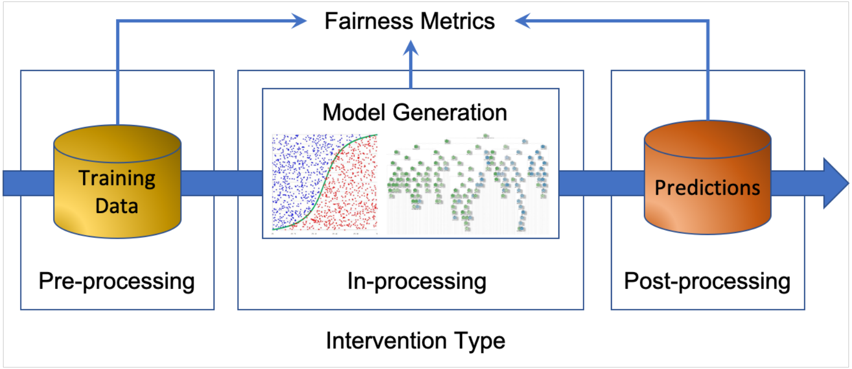
Variable omission (or blinding) has been shown to amplify biases in the data (see: [CW$^{+}$17, KC12, DH$^{+}$12]) because of proxy variables.

Many seemingly obvious approaches to improving model fairness can actually make things worse depending on your perspective(s).

**Standard Fairness Methodology**

Technical interventions at different parts of the pipeline.

# Technical Interventions in the ML Pipeline




The survey covers 11 different characterizations of intervention.

# Example: Regularization $\rightarrow$ In-Processing

We modify the optimisation function of the model: add a penalty term that regularises the model w.r.t. to (un)fair outcomes.

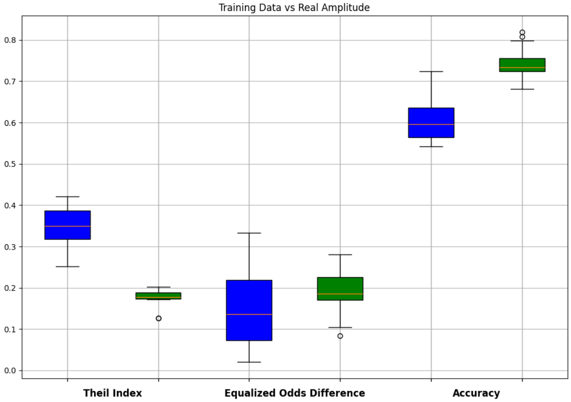
This punishes the model for “poor” scores in the fairness metric during training.

This is a fairly common approach in making models fair(er).

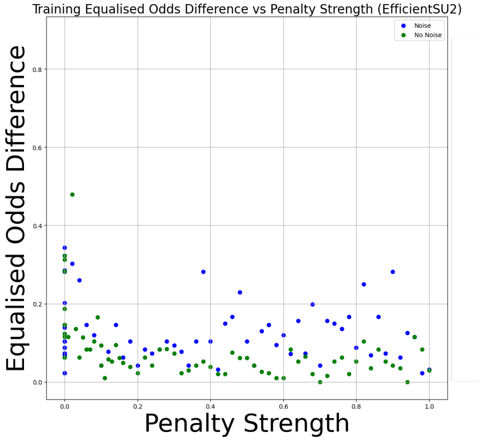
# Example: Fairness Regularization in Quantum ML

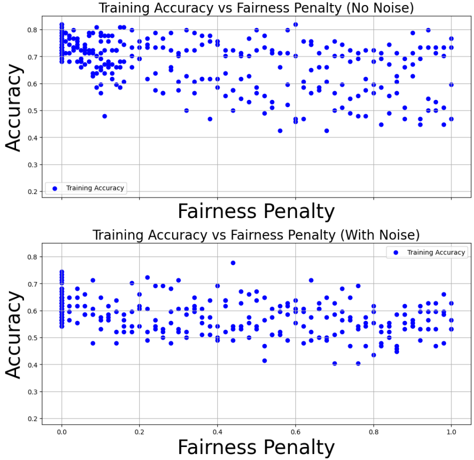
Quantum ML Models can also be unfair [BHC25].

Blue: with quantum noise
Green: without



# Example: Fairness vs. Accuracy Trade-off



# Fairness in ML / AI — What's to come?
# $\rightarrow$ (Mean) Questions & Diversification

# Standard Challenges

- Improving fairness often detriments accuracy [BH+18, DH+12, CDP+17, HP+16, Zli15, CW+17, Haa19].
- One fairness measure often detriments another [KL+18, Cho17].
- The choice of fairness measure(s) itself may even harbor, disguise, or create new underlying ethical concerns.
- Some interventions are at odds with legislation due to transformations on the data and/or model output(s).

# (Mean) Starter Questions for “Experts”

- What experience do they have with fairness in machine learning?
- Can they define fairness (if so they might be bluffing!)
- Are they worried about how to do it? (they should be!)
- Do they have a social scientist or policy person on the team?
- Do they understand the data flows and the underlying (business) objectives?

# Supporting Practitioners

ML is very accessible: so unfair practices are easier and easier to (un)intentionally arise.

A nice python library is IBM's AIF360 [BD$^+$18]

Your pre-processing is really important, e.g. [CMH22]

Interventions for diverse tasks: regression ([KT$^+$18]), reinforcement learning [GS$^+$22], quantum classifiers [BHC25], LLMs ...

More details in the survey paper!

# Final Remarks

Think very carefully what fairness means for you!

Go beyond race, gender, and age; especially into proxy-variables

Be diverse in the data you use

Go beyond making the model "fair" – look at other parts of the data analytical process too

Regardless of fairness, is your application ethically appropriate?

Sometimes it's just better to collect more data.

# References (I)


Matthew T Bodie, Miriam A Cherry, et al.
The law and policy of people analytics.
*U. Colo. L. Rev.*, 88:961, 2017.

Rachel K. E. Bellamy, Kuntal Dey, et al.
AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias.
*arXiv preprint arXiv:1810.01943*, oct 2018.

Richard Berk.
Accuracy and fairness for juvenile justice risk assessments.
*Journal of Empirical Legal Studies*, 16(1):175–194, 2019.

# References (II)


Richard Berk, Hoda Heidari, et al.
Fairness in criminal justice risk assessments: The state of the art.
*Sociological Methods & Research*, page 0049124118782533, 2018.

Sofia Badalova, Chrstian Haas, and Simon Caton.
Quantumly fair, or fairly quantum? an exploration of fairness in quantum machine learning.
In *ECAI-25 – under review*, 2025.

Solon Barocas, Moritz Hardt, and Arvind Narayanan.
*Fairness and Machine Learning*.
fairmlbook.org, 2019.

# References (III)


Sam Corbett-Davies, Emma Pierson, et al.

Algorithmic decision making and the cost of fairness.

In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806, New York, New York, USA, 2017. ACM, ACM Press.

Simon Caton and Christian Haas.

Fairness in machine learning: A survey.

*ACM Computing Surveys*, 2024.

# References (IV)


Alexandra Chouldechova.
Fair prediction with disparate impact: A study of bias in recidivism prediction instruments.
*Big data*, 5(2):153–163, 2017.

Simon Caton, Saiteja Malisetty, and Christian Haas.
Impact of imputation strategies on fairness in machine learning.
*Journal of Artificial Intelligence Research*, 74:1011–1035, 2022.

Flavio Calmon, Dennis Wei, et al.
Optimized Pre-Processing for Discrimination Prevention.
In *Advances in Neural Information Processing Systems*, pages 3992–4001, 2017.

# References (V)


Cynthia Dwork, Moritz Hardt, et al.
Fairness through awareness.
In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM, 2012.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici.
All models are wrong but many are useful: Variable importance for black-box, proprietary, or misspecified prediction models, using model class reliance.
*arXiv preprint arXiv:1801.01489*, 2018.

# References (VI)


- Stephen Gillen, Christopher Jung, et al.
  Online learning with an unknown fairness metric.
  In *Advances in Neural Information Processing Systems*, pages 2600–2609, 2018.
- Pratik Gajane, Akrati Saxena, et al.
  Survey on fair reinforcement learning: Theory and practice.
  *arXiv preprint arXiv:2205.10032*, 2022.
- Christian Haas.
  The Price of Fairness - A Framework to Explore Trade-Offs in Algorithmic Fairness.
  In *International Conference on Information Systems (ICIS) 2019*, 2019.

# References (VII)


Margeret Hall and Simon Caton.
Am I who I say I am? Unobtrusive self-representation and personality recognition on Facebook.
*PloS one*, 12(9):e0184417, 2017.

Moritz Hardt, Eric Price, et al.
Equality of opportunity in supervised learning.
In *Advances in neural information processing systems*, pages 3315–3323, 2016.

Faisal Kamiran and Toon Calders.
Data preprocessing techniques for classification without discrimination.
*Knowledge and Information Systems*, 33(1):1–33, oct 2012.

# References (VIII)


Jon Kleinberg, Jens Ludwig, et al.
Algorithmic fairness.
In *Aea papers and proceedings*, volume 108, pages 22–27, 2018.

Junpei Komiyama, Akiko Takeda, et al.
Nonconvex optimization for regression with fairness constraints.
In *International conference on machine learning*, pages 2737–2746. PMLR, 2018.

Arefeh Kazemi, Arjumand Younus, Mingyeong Jeon, M Atif Qureshi, and Simon Caton.
Inéire: An interpretable nlp pipeline summarising inclusive policy making concerning migrants in ireland.
*IEEE Access*, 2023.

# References (IX)


Douglas S Massey and Nancy A Denton.
*American apartheid: Segregation and the making of the underclass*.
Harvard University Press, 1993.

Emma Pierson.
Demographics and discussion influence views on algorithmic fairness.
*arXiv:1712.09124*, 2017.

Hansen Andrew Schwartz, Johannes C Eichstaedt, et al.
Toward personality insights from language exploration in social media.
In *2013 AAAI Spring Symposium Series*, 2013.

# References (X)


Andrew D Selbst.
Disparate impact in big data policing.
*Ga. L. Rev.*, 52:109, 2017.

Babak Salimi, Bill Howe, and Dan Suciu.
Data management for causal algorithmic fairness.
*Data Engineering*, page 24, 2019.

Lauren Weber and Elizabeth Dwoskin.
Are workplace personality tests fair.
*Wall Street Journal*, 29, 2014.

# References (XI)


Tal Yarkoni.

Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers.

*Journal of research in personality*, 44(3):363–373, 2010.

Indre Zliobaite.

On the relation between accuracy and fairness in binary classification.

may 2015.

UCD
DUBLIN