Fairness and discrimination

William Lowe

Hertie School

11th November 2020

Rules

People and organisations *have* rules and *make* decisions

→ Decisions are made *according to*, *mostly according to*, or *despite* the rules
→ Rules may be internally inconsistent and require *balancing* or weighting (looking at you, lawyers)
→ When there are no applicable rules, decisions are either idiosyncratic (sub-organisationally rather organisationally determined) or governed by rules that *could* in principle be made explicit, but have not been

The relationship between explicit, implicit, contradictory, or absent rules and institutional decisions is a key issue in much public administration theory (and practice)

We won't have much to say there, except…

Algorithms and rules

It is often argued that these issues are made worse by the presence of 'algorithmic' or machine learning decision-making tools

→ This is false. All explicit decision-making processes are algorithms

Thank you, Muhammad ibn Mūsa al-Khwarizmī!

→ This was anyway always true of algebra, due its concerns for maintaining equalities (al-K's big hit was 'The Compendious Book on Calculation by Completion and Balancing')
→ Used to establish 'fair division' in inheritance problems

Now algorithms are realised in machines, the field of *algorithmic fairness* in computer science / machine learning is a great place to study their fairness

Algorithmic performance

In many domains, 'algorithmic' decision making is equivalent or superior in performance to human judgment, e.g.

- → Information extraction by experts vs undergraduates vs machines (King & Lowe, 2003)
- → Clinical decisions (Grove et al., 2000)
- → Recidivism predictions (Lin et al., 2020)

and not obviously *less* transparent than humans

- → Humans can be asked for reasons, but they may not be causes
- → Machines can be asked for a lot more

Explain, please

European Union legal constraints (sensibly) do not distinguish between who – humans, algorithms, or both – does the data processing

> *"The data subject shall have the right to obtain [. . .] confirmation as to whether or not personal data concerning him or her are being processed, and [. . .] access to the personal data [. . .] and [. . .] meaningful information about the logic involved."*
>
> *(GDPR Art. 15)*

Consequently, there is a demand for 'Explainable AI' (XAI) [link]

→ So, what counts as an explanation?

Explanation and 'explanation'

Well, we want an interpretation

Explanation and 'explanation'

Well, we want an interpretation

> *Model interpretation is the ability to explain and validate the decisions of a predictive model to enable fairness, accountability, and transparency in the algorithmic decision-making*
>
> *[Skater] (Oracle)*

Not enormously illuminating…

Explanation and 'explanation'

Well, we want an interpretation

> *Model interpretation is the ability to explain and validate the decisions of a predictive model to enable fairness, accountability, and transparency in the algorithmic decision-making*
>
> *[Skater] (Oracle)*

Not enormously illuminating...

Scope of Interpretation	Algorithms	
Global Interpretation	Model agnostic Feature Importance	
Global Interpretation	Model agnostic Partial Dependence Plots	
Local Interpretation	Local Interpretable Model Explanation(LIME)	
Local Interpretation	DNNs	• Layer-wise Relevance Propagation (e-LRP): image • Integrated Gradient: image and text
Global and Local Interpretation	• Scalable Bayesian Rule Lists • Tree Surrogates	

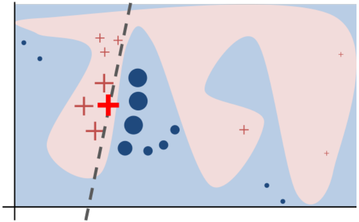
Explanation

Scope of Interpretation	Algorithms	
Global Interpretation	Model agnostic Feature Importance	
Global Interpretation	Model agnostic Partial Dependence Plots	
Local Interpretation	Local Interpretable Model Explanation(LIME)	
Local Interpretation	DNNs	• Layer-wise Relevance Propagation (e-LRP): image • Integrated Gradient: image and text
Global and Local Interpretation	• Scalable Bayesian Rule Lists • Tree Surrogates	

Translation:

→ Marginal effects (Lundberg & Lee, 2017; Shrikumar et al., 2019)
→ Conditional effects, e.g. WhatIf [link]
→ A simpler model, local to a data point (LIME; Ribeiro et al., 2016)
→ Even more marginal effects
→ An equivalent decision tree (Craven & Shavlik, 1995; Wang et al., 2020)

Local explanation



Intuitively, an explanation is a local linear approximation of the model's behaviour. [LIME]

Local Interpretable Model-Agnostic Explanations (LIME; Ribeiro et al., 2016)

Explanation and 'explanation'

We can also ask for *counterfactuals*

→ Note: humans can answer counterfactual questions, but they are sometimes (often?) wrong about what they would do

Explanation and 'explanation'

We can also ask for *counterfactuals*

- → Note: humans can answer counterfactual questions, but they are sometimes (often?) wrong about what they would do

Aside: Roughly speaking,

- → when psychologists study decision making they make sure not pay subjects to minimise decision bias
- → when economists study decision making they make sure to pay subjects to minimise decision bias

Explanation and ‘explanation’

We can also ask for *counterfactuals*

- → Note: humans can answer counterfactual questions, but they are sometimes (often?) wrong about what they would do

Aside: Roughly speaking,

- → when psychologists study decision making they make sure not pay subjects to minimise decision bias
- → when economists study decision making they make sure to pay subjects to minimise decision bias

Counterfactuals are a class of conditional effect. However…

- → if they are *outside the convex hull of the training data* it’s not clear whether a surrogate global model, e.g. a decision tree, will agree with the ‘real’ model about them

Rules and fairness

Rules, decisions, or both may be unfair, which we will understand as a form of undesirable *bias / discrimination*. Reminder:

- → Many of forms of *discrimination* are considered desirable
- → We saw previously that even *bias* has virtues for (machine) learning. It reduces variance and therefore over-fitting

We will treat the determination of which forms of bias and discrimination forms are undesirable as given by an external source, and ask:

- → How to make decisions without the *undesirable* bias / discrimination?

Fairness with respect to what?

Most of this literature treats discrimination as concerning 'protected attributes', e.g. race, gender, religion (or lack thereof), etc.

- → Naturally understood as a variable, A
- → Measurable on the individual level and as defining a population quantity

We can define fairness at

- → the individual level
- → the group level
- → a mixture of both

and whether it is determined by

- → outcomes $\hat{Y}_i$ vs $\hat{Y}_j$, or $E[Y \mid A = 1]$ vs $E[Y \mid A = 0]$
- → the procedure that generates $\hat{Y}$

Lots of possibilities (Barocas et al., 2019)

Problem setup

Consider variables X, U, A, and Y

- → Y the outcome we want to predict / make decision with respect to, e.g. loan-worthiness, recidivism
- → $\hat{Y}$ our prediction of Y, e.g. probability (or amount) of eventual loan repayment, whether caught committing another crime. A function of X, A, or both. Often thresholded at τ to make a decision
- → X non-protected observed features we might use to use to create $\hat{Y}$, e.g. previous payment history, or criminal record
- → A protected features we want our predictions / decisions to be fair with respect to

Desiderata

A natural baseline expectation from $\hat{Y}$ when it is a probability is that it is *calibrated*

calibration

$$P(Y = 1 \mid \hat{Y} = v, A = 1) = P(Y \mid \hat{Y} = v, A = 0) \qquad \forall v$$

This is a *relative* calibration

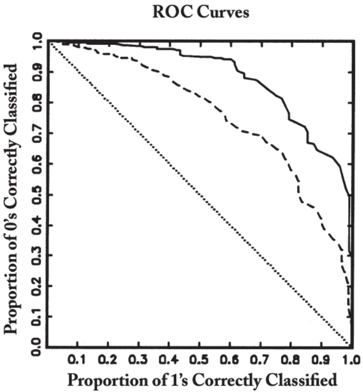
- → does not require $P(Y = 1 \mid \hat{Y} = v) = v$
- → does require the (mis)calibrations to be the same across A

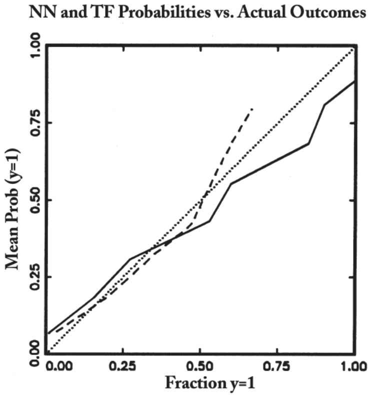
A calibrated measure is *free from predictive bias* (though may have some regular statistical bias)

- → v ‘means the same thing’ across groups

Calibration and error rates

From King and Zeng (2001)





Desiderata

A related requirement is that the proportion of cases above τ threshold is the same across groups

PREDICTIVE PARITY

$$P(Y = 1 \mid \hat{Y} > \tau, A = 1) = P(Y \mid \hat{Y} > \tau, A = 0)$$

Relatedly, the positive predictive value

$$\text{PPV} = \frac{1}{N} \sum_{i}^{N} I\left[\hat{Y}_i > \tau\right]$$

is the same across groups

Desiderata

A related requirement is that the proportion of cases above τ threshold is the same across groups

PREDICTIVE PARITY

$$P(Y = 1 \mid \hat{Y} > \tau, A = 1) = P(Y \mid \hat{Y} > \tau, A = 0)$$

Relatedly, the positive predictive value

$$\text{PPV} = \frac{1}{N} \sum_{i}^{N} I\left[\hat{Y}_i > \tau\right]$$

is the same across groups

Note: This may not follow from calibration, *if* $\hat{Y}$ is not itself a probability *and* the distribution of $\hat{Y}$ differs across A

Desiderata

equal error rates

Equal false positive rate

$$P(\hat{Y} > \tau \mid Y = 0, A = 1) = P(\hat{Y} > \tau \mid Y = 0, A = 0) \qquad \text{FPR}$$

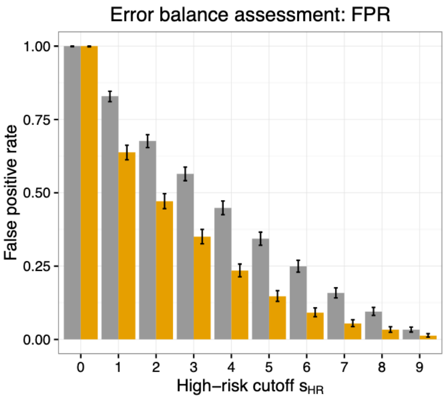
and equal false negative rate

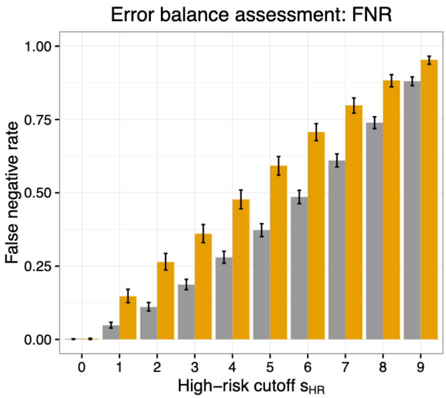
$$P(\hat{Y} \leq \tau \mid Y = 1, A = 1) = P(\hat{Y} \leq \tau \mid Y = 1, A = 0) \qquad \text{FNR}$$

It seems natural to ask that classification *mistakes* not be different across groups

Controversy

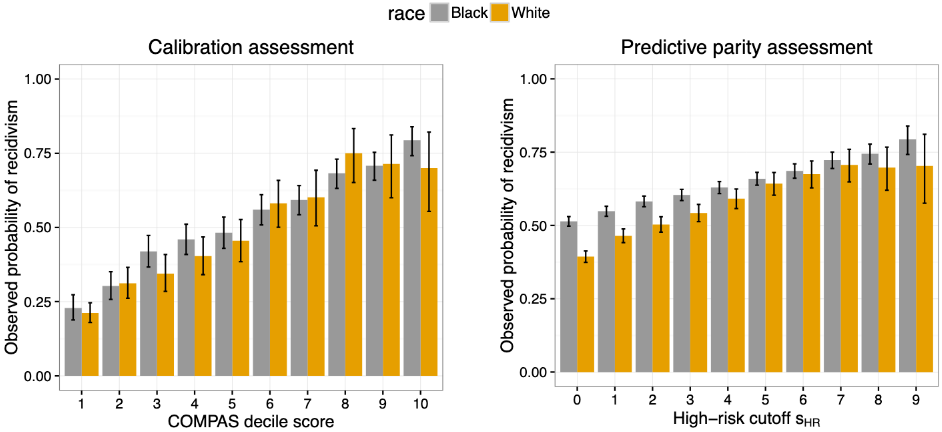
ProPublica (Jeff Larson et al., 2016) noted that a commercial recidivism prediction tool COMPAS had quite different error rates by race





Controversy

The company involved responded that, sure, but the classifier was well calibrated



A fundamental problem

When recidivism *prevalence* $p = P(Y = 1 \mid A = a)$ differs with a, we *cannot* have calibration and (both) error rates equal (Chouldechova, 2017; Kleinberg et al., 2016)

A fundamental problem

When recidivism *prevalence* $p = P(Y = 1 \mid A = a)$ differs with a, we *cannot* have calibration and (both) error rates equal (Chouldechova, 2017; Kleinberg et al., 2016)

Recall from our (second) ML lecture on classification

For any threshold τ on $\hat{Y}$, any binary classifier performance is described by the following table

	$\hat{Y} \leq \tau$	$\hat{Y} > \tau$	
$Y = 0$	TN	FP	$1 - p$
$Y = 1$	FN	TP	p
	1-PPV	PPV	

We will get one of these per value of A

A fundamental problem

	$\hat{Y} \leq \tau$	$\hat{Y} > \tau$	
$Y = 0$	TN	FP	$1 - p$
$Y = 1$	FN	TP	p
	1-PPV	PPV	

implies lots of numerical constraints. Chouldechova shows that

$$\text{FPR} = \frac{p}{1-p} \frac{1-\text{PPV}}{\text{PPV}} (1 - \text{FNR})$$

> *if an instrument satisfies predictive parity – that is, if the PPV is the same across groups – but the prevalence differs between groups, the instrument cannot achieve equal false positive and false negative rates across those groups.* *(Chouldechova, 2017)*

Diagnosis?

Notions of fairness are

→ incomplete
→ inconsistent
→ trying to work on too many levels at once

Responses:

→ Nihilism: fairness is incoherent, so choose your poison
→ Optimism: there is a coherent notion of fairness but we haven't got it yet
→ Causal inference: Let's look at this a new way (Kusner et al., 2017)

INDIVIDUAL FAIRNESS

Start at the individual again with a basic intuition:

INDIVIDUAL FAIRNESS

Define the *distance* d between two individuals i and j as a function of their measured features X and A

→ Think matching

$$\text{if } d(i,j) \text{ is small then } \hat{Y}_i \approx \hat{Y}_j$$

Similar people should get similar predictions / decisions

→ formal, outcome-oriented
→ incomplete, and hard to verify without specification of what d or $\approx$ mean
→ difficult to justify e.g. careful choice of d can make a *wide* range of prediction / decision differences 'fair'
→ Almost always unwise to take distances (or inversely 'similarity') as a theoretical primitive

COUNTERFACTUAL FAIRNESS

COUNTERFACTUAL FAIRNESS

$$P(\hat{Y}_i^{A=a} \mid A_i = a, X_i = x) = P(\hat{Y}^{A=a'} \mid A_i = a, X_i = x)$$

where $\hat{Y}^{A=a}$ is the prediction an individual gets, and $\hat{Y}^{A=a'}$ is the prediction they would have received if their protected characteristic had instead been a'.

→ $\hat{Y}_i$ is fair if it would not have been different had A_i taken a different value

This has a number of interesting properties:

→ individual
→ exactly *half* outcome-oriented (the observed $\hat{Y}^{A=a}$)
→ A special case of IF that specifies the distance function

COUNTERFACTUAL FAIRNESS

COUNTERFACTUAL FAIRNESS

Sufficient (but not necessary) condition

→ Lemma: Conditioning on *non-children* of A will always be fair

This does not hold for some other definitions…

COUNTERFACTUAL FAIRNESS

COUNTERFACTUAL FAIRNESS

Sufficient (but not necessary) condition

→ Lemma: Conditioning on *non-children* of A will always be fair

This does not hold for some other definitions…

We are now at the state of the art:

→ Group-based fairness
→ Individual-based fairness
→ Counterfactual individual-based fairness

COUNTERFACTUAL FAIRNESS

Looking backwards

- → Counterfactually defined fairness is not *entirely* new
- → We met it with mediation analysis
- → discrimination is proved when there is a *direct* effect of gender on hiring decisions; *indirect* effects via choice of job to apply for are not (legally) discrimination

COUNTERFACTUAL FAIRNESS

Looking backwards

- → Counterfactually defined fairness is not *entirely* new
- → We met it with mediation analysis
- → discrimination is proved when there is a *direct* effect of gender on hiring decisions; *indirect* effects via choice of job to apply for are not (legally) discrimination

Maybe this approach is general. I hope so, but I'm biased…

References

Barocas, S., Hardt, M. & Narayanan, A. (2019). 'Fairness and machine learning'. fairmlbook.org.

Chouldechova, A. (2017, February 28). *Fair prediction with disparate impact: A study of bias in recidivism prediction instruments*. Retrieved November 10, 2020, from http://arxiv.org/abs/1703.00056

Craven, M. & Shavlik, J. W. (1995). 'Extracting tree-structured representations of trained networks'. *Proceedings of the 8th International Conference on Neural Information Processing Systems*, 7.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E. & Nelson, C. (2000). 'Clinical versus mechanical prediction: A meta-analysis.' *Psychological Assessment*, *12*(1), 19–30.

Jeff Larson, Surya Mattu, Lauren Kirchner & Julia Angwin. (2016, May 23). *How we analyzed the compas recidivism algorithm*. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

References

King, G. & Lowe, W. (2003). ‘An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design’. *International Organization*, *57*(3), 617–642.

King, G. & Zeng, L. (2001). ‘Improving forecasts of state failure’. *World Politics*, *53*(4), 623–658.

Kleinberg, J., Mullainathan, S. & Raghavan, M. (2016, November 17). *Inherent trade-offs in the fair determination of risk scores*. Retrieved November 9, 2020, from http://arxiv.org/abs/1609.05807

Kusner, M. J., Loftus, J., Russell, C. & Silva, R. (2017). Counterfactual fairness. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan & R. Garnett (Eds.), *Advances in neural information processing systems 30* (pp. 4066–4076). Curran Associates, Inc.

Lin, Z. ‘J., Jung, J., Goel, S. & Skeem, J. (2020). ‘The limits of human predictions of recidivism’. *Science Advances*, *6*(7).

References

Lundberg, S. M. & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan & R. Garnett (Eds.), *Advances in neural information processing systems* (pp. 4765–4774). Curran Associates, Inc.

Ribeiro, M. T., Singh, S. & Guestrin, C. (2016, August 9). *”why should i trust you?”: Explaining the predictions of any classifier*. Retrieved November 10, 2020, from http://arxiv.org/abs/1602.04938

Shrikumar, A., Greenside, P. & Kundaje, A. (2019, October 12). *Learning important features through propagating activation differences*. Retrieved November 10, 2020, from http://arxiv.org/abs/1704.02685

Wang, C., Han, B., Patel, B., Mohideen, F. & Rudin, C. (2020, May 8). *In pursuit of interpretable, fair and accurate machine learning for criminal recidivism prediction*. Retrieved November 10, 2020, from http://arxiv.org/abs/2005.04176