Trust and Trustworthiness for Dependable Machine Learning

Flavio Figueiredo (flaviovdf@dcc.ufmg.br)
Universidade Federal de Minas Gerais

Examples from NeuRIPS 2018 -- Human Side of Machine Learning

Outline

Initial Discussions

- Trust and trustworthiness
- When a system dependable? How does machine learning affect all of this?

Overview of ML systems and definitions

- Fairness
- Accountability
- Transparency, and
- Interpretability

Joint Work

Atmosphere's WP6 Team

- Leandro Balby (UFCG)
- Vasiliki Diamantopoulou (UPRC)
- Wagner Meira (UFMG)
- + others

Trust and Trustworthiness

Simplified Background

- Mathematically speaking, what is the goal of a **supervised** learning system?
  $\mathcal{D} = \{(\mathbf{x}_1, y_1), \cdots (\mathbf{x}_N, y_N)\}$

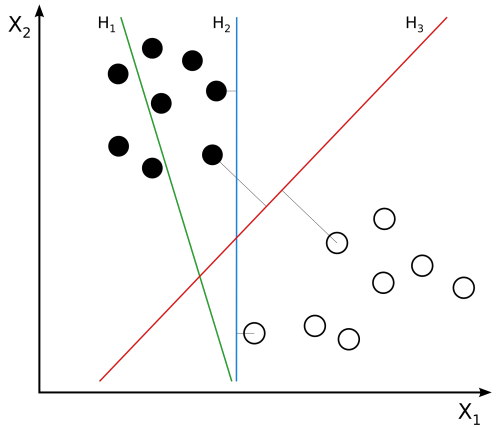
- The goal is to learn some parameters
  $\boldsymbol{\Theta}$

- Where these parameters maximize some prediction function across y
  $\hat{\boldsymbol{\Theta}} = \arg\max_{\boldsymbol{\Theta}} P(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\Theta})$

This is just one view. Optimization v. Bayesian and other topics out of the scope.

Simplified Background

- The goal of a supervised learning algorithm is to **discriminate**

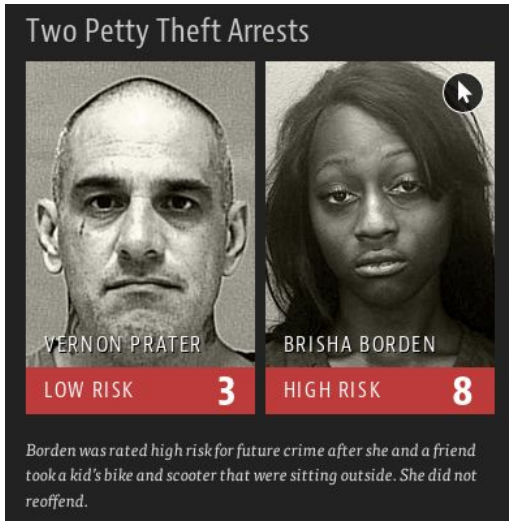


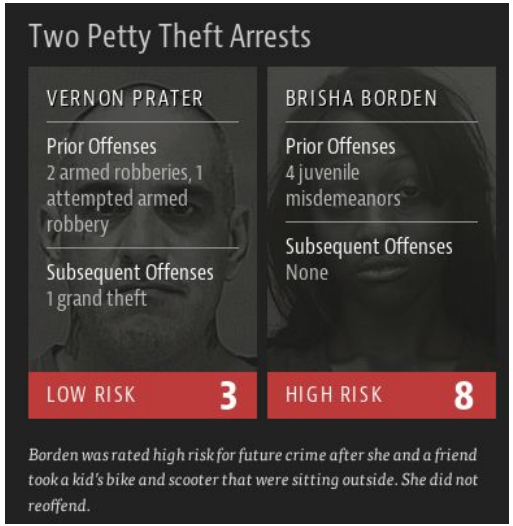
Simplified Background

- The goal of a supervised learning algorithm is to **discriminate**
- Why are we now so worried that it does? It seems we can **trust** them.

Machine Bias

- Pro Publica analysis of COMPAS (which stands for Correctional Offender Management Profiling for Alternative Sanctions)
  https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing





What is trust?

- Inspired by Onora O'Neill
  https://www.youtube.com/watch?v=XWwTYy9k5nc

- Consider a question? Do we trust politicians?
  - This question has had the same answer for a long time
  - People usually don’t trust politicians

What is trustworthiness?

- Inspired by Onora O'Neill
  https://www.youtube.com/watch?v=XWwTYy9k5nc

- Consider a question? Do we trust politicians?
  - This question has had the same answer for a long time
  - People usually don't trust politicians
- **Trustworthiness**
  - Evidence of **why** can I trust you
  - Evidence is observable (though hard to quantify): competence, reliability
  - We trust a science not because it came from a scientist, it is **testable**
- We need to direct our **trust** to **trustworthy** properties.
  - Why and when can I trust

Services need to earn trust. They need to be trustworthy

Trustworthiness changes over time

Dependability

Dependability

From Wikipedia

Dependability

From Wikipedia, the free encyclopedia

In systems engineering, **dependability** is a measure of a system's **availability**, **reliability**, and its **maintainability**, and **maintenance support performance**, and, in some cases, other characteristics such as **durability**, **safety** and **security**.[1] In software engineering, **dependability** is the ability to provide services that can defensibly be trusted within a time-period.[2] This may also encompass mechanisms designed to increase and maintain the dependability of a system or software.[3]

The International Electrotechnical Commission (IEC), via its Technical Committee TC 56 develops and maintains international standards that provide systematic methods and tools for dependability assessment and management of equipment, services, and systems throughout their life cycles.

Dependability can be broken down into three elements:

- **Attributes** - A way to assess the dependability of a system
- **Threats** - An understanding of the things that can affect the dependability of a system
- **Means** - Ways to increase a system's dependability

However, over time these properties get more complex.

ML systems currently impact society.

It is interesting that it mentions: **trust over time**

It’s a hard problem

- **Uptime.** Do we want more or less of it?

It’s a hard problem

- **Uptime.** Do we want more or less of it?
- Probably more, there are clear problems that require more uptime.
- Maybe cost and energy are issues, but they are quantifiable issues

It’s a hard problem

- **Fairness.** Do we want more or less of it?

It’s a hard problem

- **Fairness.** Do we want more or less of it?
- Name one problem solved by machine learning fairness?

It’s a hard problem

- **Fairness.** Do we want more or less of it?
- Name one problem solved by machine learning fairness?
- What is fairness?!

It’s a hard problem

- **Fairness.** Do we want more or less of it?
- Name one problem solved by machine learning fairness?
- What is fairness?!

Let’s try and help

- On a credit scoring system that helps one decide loans to give out. Do we want more fairness?

It’s a hard problem

- **Fairness.** Do we want more or less of it?
- Name one problem solved by machine learning fairness?
- What is fairness?!

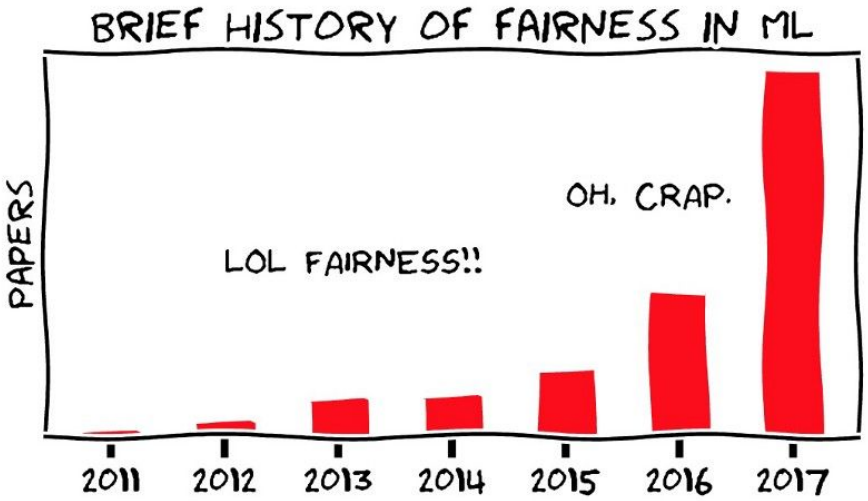
Let’s try and help

- On a credit scoring system that helps one decide loans to give out. Do we want more fairness?
- What if the system simply denies all loans?

Machine Learning Systems

- Need to specify their trustworthiness
  - Fairness
  - Transparency
  - Accountability
  - Interpretability

- I can’t even trust the definition of fairness



https://fairmlclass.github.io

It is a human problem

Nitin Koli, Joshua Kroll (NeuRIPS, 2018)

- Issues of fairness, transparency, accountability, transparency and interpretability are **social-technological**

  “Technologies don’t live in a vacuum and if we pretend that they do we kind of have put our blinders on and decided to ignore any human problems.”

Dependability

From the Working Group

About IFIP Working Group 10.4

Increasingly, individuals and organizations are developing or procuring sophisticated computing systems on whose services they need to place great reliance. In differing circumstances, the focus will be on differing properties of such services -- e.g., continuity, performance, real-time response, ability to avoid catastrophic failures, prevention of deliberate privacy intrusions.

The notion of **dependability**, defined as *the trustworthiness of a computing system which allows reliance to be justifiably placed on the service it delivers,* enables these various concerns to be subsumed within a single conceptual framework. Dependability thus includes as special cases such attributes as **reliability, availability, safety, security**.

The Working Group is aimed at identifying and integrating approaches, methods and techniques for specifying, designing, building, assessing, validating, operating and maintaining computer systems which should exhibit some or all of these attributes.

Specifically, the Working Group is concerned with progress in:

1. Understanding of faults (accidental faults, be they physical, design-induced, originating from human interaction; intentional faults) and their effects.
2. Specification and design methods for dependability.
3. Methods for error detection and processing, and for fault treatment.
4. Validation (testing, verification, evaluation) and design for testability and verifiability.
5. Assessing dependability through modeling and measurement.

Dependability

- Assumes that the human negatively impacts system dependability
- We are now learning that with machine learning systems it’s the other way around
- Systems now negatively impact social structures

Dependability

- Assumes that the human negatively impacts system dependability
- We are now learning that with machine learning systems it’s the other way around
- Systems now negatively impact social structures

How social media filter bubbles and algorithms influence the election

With Facebook becoming a key electoral battleground, researchers are studying how automated accounts are used to alter political debate online

Revealed: Facebook’s internal rules on sex, terrorism and violence

▲ A Facebook Live broadcast hosted by ITV News had Theresa May answering questions sent in by users of the site. Photograph: Facebook/PA

Facebook’s WhatsApp limits text forwards to 5 recipients to curb rumors

PUBLISHED MON, JAN 21 2019 • 4:46 AM EST | UPDATED TUE, JAN 22 2019 • 9:58 AM EST

REUTERS

SHARE

KEY POINTS

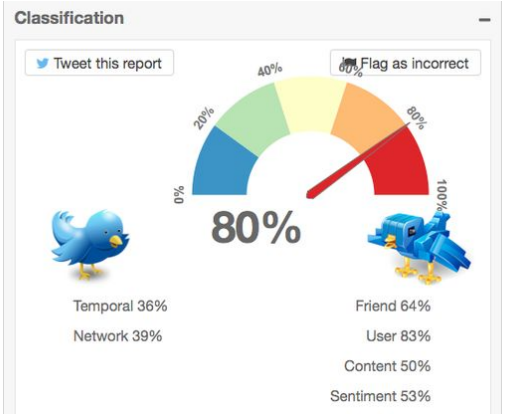
- WhatsApp expanded a cap on message forwarding it first introduced in India after the spread of rumors led to killings and lynching attempts.
- The messaging app will roll out an update to activate the new forward limit, starting Monday, WhatsApp’s head of communications said.

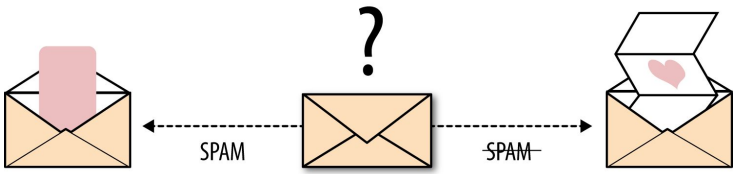
TRENDI

Chris Ratcliffe | Bloomberg | Getty Images

The old arms race

- W can tackle these problems as an arms race





Social Technological

- Computer Science
- Data Science
- Law
- Health
- Politics

  etc.

https://vimeo.com/149389876

FAT*

What is Fairness?

Let’s begin with fairness as it closely relates to all other metrics.

21 fairness definitions and their politics

Arvind Narayanan - FAT Conference 2018 Tutorial

- Computer Scientist on a wild goose chase for a single definition
- There is value to various definitions
- Each can lead to **trustworthiness**

What is Fairness?

Sahil Verma and Julia Rubin (2018) -- Fairness Definitions Explained

- A lot of these metrics worry about some form of equality

  $\hat{\Theta} = \arg\max_{\Theta} P(\mathbf{y} \mid \mathbf{X}, \Theta)$

- Let S be some subset of sensitive attributes.

  S = ***{ col(j, X) | column j is sensitive }***

  N = ***{ col(i, X) | column i is not-sensitive }***

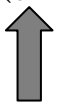
	Definition	Paper	Citation #	Result
3.1.1	Group fairness or statistical parity	[12]	208	×
3.1.2	Conditional statistical parity	[11]	29	✓
3.2.1	Predictive parity	[10]	57	✓
3.2.2	False positive error rate balance	[10]	57	×
3.2.3	False negative error rate balance	[10]	57	✓
3.2.4	Equalised odds	[14]	106	×
3.2.5	Conditional use accuracy equality	[8]	18	×
3.2.6	Overall accuracy equality	[8]	18	✓
3.2.7	Treatment equality	[8]	18	×
3.3.1	Test-fairness or calibration	[10]	57	✓
3.3.2	Well calibration	[16]	81	✓
3.3.3	Balance for positive class	[16]	81	✓
3.3.4	Balance for negative class	[16]	81	×
4.1	Causal discrimination	[13]	1	×
4.2	Fairness through unawareness	[17]	14	✓
4.3	Fairness through awareness	[12]	208	×
5.1	Counterfactual fairness	[17]	14	–
5.2	No unresolved discrimination	[15]	14	–
5.3	No proxy discrimination	[15]	14	–
5.4	Fair inference	[19]	6	–

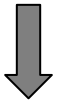
Table 1: Considered Definitions of Fairness

Balanced Representation

Is this fairness?

$$\hat{\Theta} = \arg\max_{\Theta} P(\mathbf{y} \mid \mathbf{X}, \Theta)$$





$\mathbf{X} =$

1	1	1	0	1
1	0	1	1	0
0	0	1	0	1
1	0	1	1	0
1	1	0	1	0

N **S**

Parity in Predictions

	Actual – Positive	Actual – Negative
Predicted – Positive	**True Positive (TP)** PPV = $\frac{TP}{TP+FP}$ TPR = $\frac{TP}{TP+FN}$	**False Positive (FP)** FDR = $\frac{FP}{TP+FP}$ FPR = $\frac{FP}{FP+TN}$
Predicted – Negative	**False Negative (FN)** FOR = $\frac{FN}{TN+FN}$ FNR = $\frac{FN}{TP+FN}$	**True Negative (TN)** NPV = $\frac{TN}{TN+FN}$ TNR = $\frac{TN}{TN+FP}$

Some Examples

- Demographic Parity

  $$P(\hat{y} = 1|s = 0) = P(\hat{y} = 1|s = 1)$$

- Equality in Opportunity (FNR)

  $$P(\hat{y} = 0|s = 0, y = 0) = P(\hat{y} = 0|s = 1, y = 0)$$

- Calibration

  $$E[y|s = 0, \hat{y} = p] = E[y|s = 1, \hat{y} = p] \;\; \forall p \in [0, 1]$$

No free Lunch

Arvind Narayanan - FAT Conference 2018 Tutorial

		True condition				
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$	F_1 score = $\frac{1}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}{2}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$		

Feedback Loops and Utility

How safe do we want a city to be?

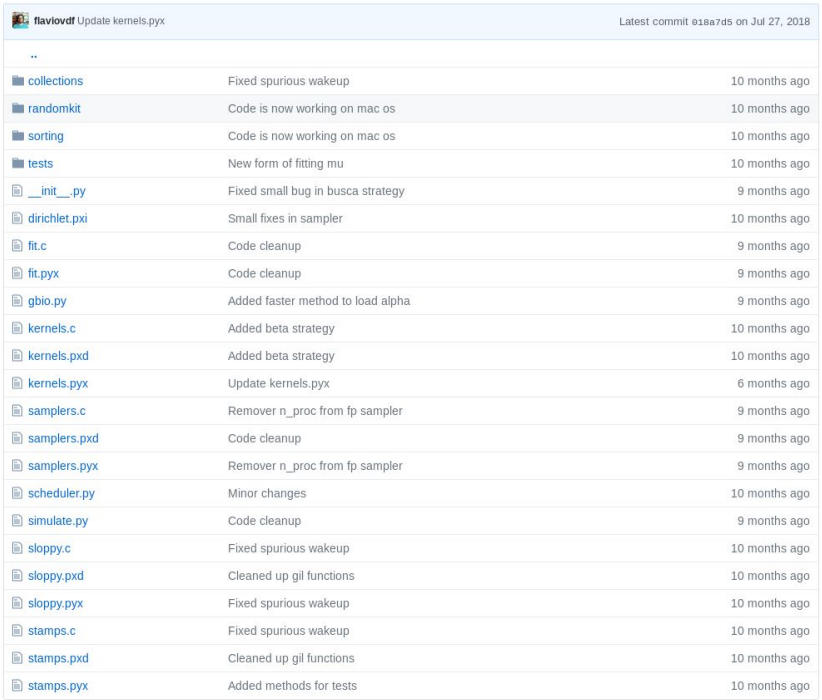
It can be shown thresholding this score, leads to unfairness.

Drawbacks

- We are mostly focused on correlations
- Maybe that nice matrix is impossible
- We are reducing a **processes** to **measures**

Free Software Approach to Transparency

- The source code is public and auditable



flaviovdf Update kernels.pyx — Latest commit 018a7d5 on Jul 27, 2018

..		
collections	Fixed spurious wakeup	10 months ago
randomkit	Code is now working on mac os	10 months ago
sorting	Code is now working on mac os	10 months ago
tests	New form of fitting mu	10 months ago
__init__.py	Fixed small bug in busca strategy	9 months ago
dirichlet.pxi	Small fixes in sampler	10 months ago
fit.c	Code cleanup	9 months ago
fit.pyx	Code cleanup	9 months ago
gbio.py	Added faster method to load alpha	9 months ago
kernels.c	Added beta strategy	10 months ago
kernels.pxd	Added beta strategy	10 months ago
kernels.pyx	Update kernels.pyx	6 months ago
samplers.c	Remover n_proc from fp sampler	9 months ago
samplers.pxd	Code cleanup	9 months ago
samplers.pyx	Remover n_proc from fp sampler	9 months ago
scheduler.py	Minor changes	10 months ago
simulate.py	Code cleanup	9 months ago
sloppy.c	Fixed spurious wakeup	10 months ago
sloppy.pxd	Cleaned up gil functions	10 months ago
sloppy.pyx	Fixed spurious wakeup	10 months ago
stamps.c	Fixed spurious wakeup	10 months ago
stamps.pxd	Cleaned up gil functions	10 months ago
stamps.pyx	Added methods for tests	10 months ago

A Lazy ML Approach to Transparency

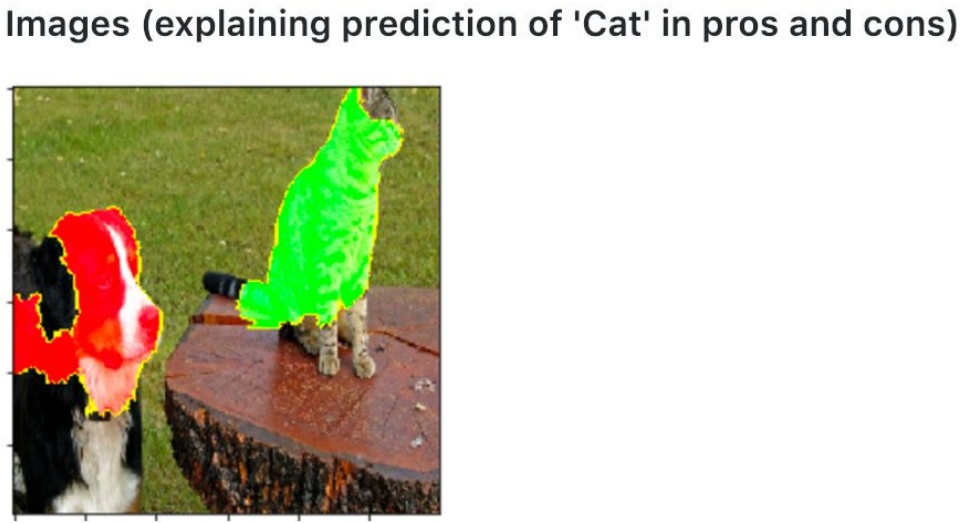
- I employed a simple model, thus it is easy to understand

Table 4: Average value (μ) and the lower ($\downarrow$95%) and upper HPD ($\uparrow$95%) values for significant explanatory variables in each topic cluster.

	μ	$\downarrow$95%	$\uparrow$95%
(S1) Out of Scope			
Intercept	-2.11	-3.04	-1.30
Male Gender & Aud. D	-2.06	-3.98	-0.03
(S2) About Characters			
Intercept	1.66	0.80	2.46
(S3) Greetings			
Intercept	-3.32	-4.56	-2.01
(S4) Reaction to Failure			
Intercept	-4.48	-6.26	-2.78
Audience B	1.74	0.18	3.60
Direct Address	1.88	0.38	3.61
Direct Address & Aud. B	-1.74	-3.35	-0.16

Interpretability

Lime and Shap. Also limited, ML to explain ML.



Accountability

From Wikipedia:

“In ethics and governance, accountability is answerability, blameworthiness, liability, and the expectation of account-giving.”

General Data Protection Regulation

Also the California Consumer Privacy Act

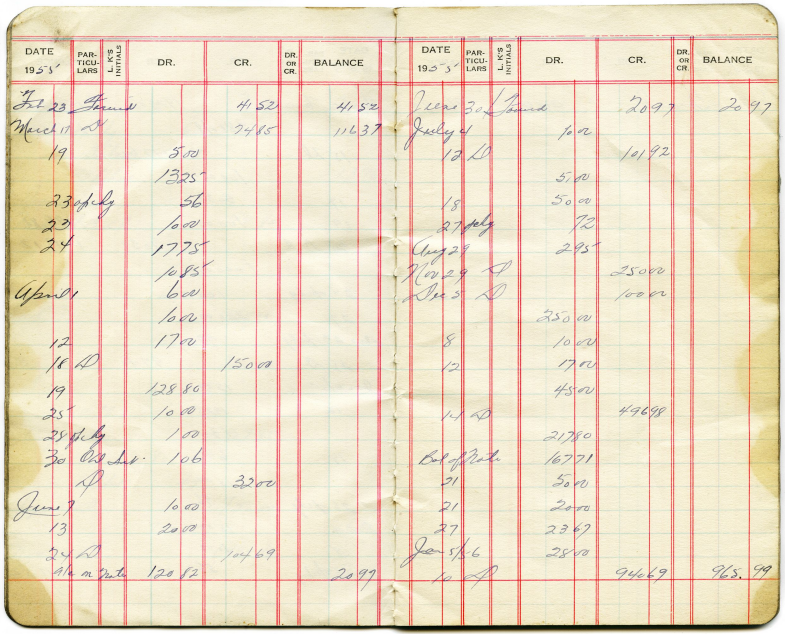
- Big companies need to be **accountable** when using your data
- However, impacts exist even when I agree to share my data
  - Let’s say most of the population agrees to share data with Facebook
- Who is **accountable?**

Accountability

- How do we measure accountability?

Accountability

- How do we measure accountability?
- Machines can keep track of records (data provenance)



Accountability

- What do we do with it?
- Conflicts with Privacy
- Hot topic in ML nowadays. Is it our job to make others accountable?

Drawbacks

- We are mostly focused on correlations
- We are reducing a **processes** to **measures**

- Kroll et al. (2018). When is an election fair (or transparent, accountable)?
  An election is a process. The whole process should be accountable, transparent and subject to recounts.

Counterfactuals

- Evaluates the impact of features with counterfactual approach [Zhang and Bareinboim (2018)]
- **“Would the prediction change if the subject were black?”**

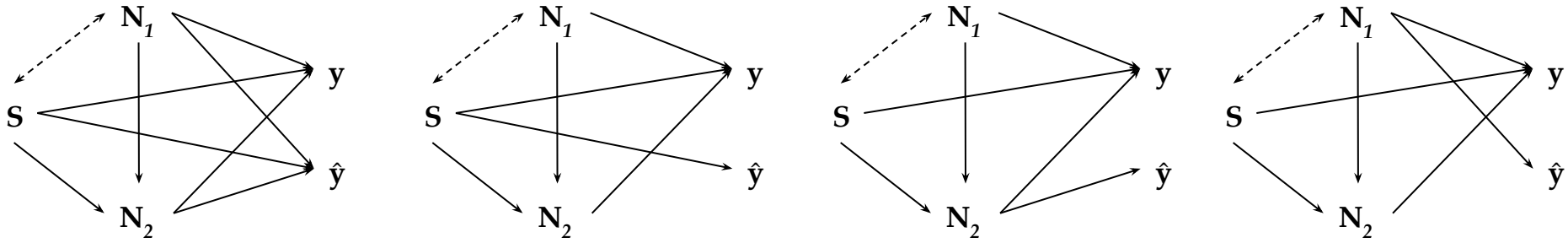
Causal Analysis

https://pair-code.github.io/what-if-tool/

1. For a given subject
2. Find the closest point with a different prediction and different sensitive attribute
3. Swap features keeping the sensitive attribute

Causal Fairness

The COMPASS model. $\mathbf{S}$ captures race. $\mathbf{N}_1$ demography. $\mathbf{N}_2$ prior convictions.
Zhang and Bareinboim (2018) -- Equality of Opportunity in Classification



Each of these classifiers have the same equalized odds.

✔

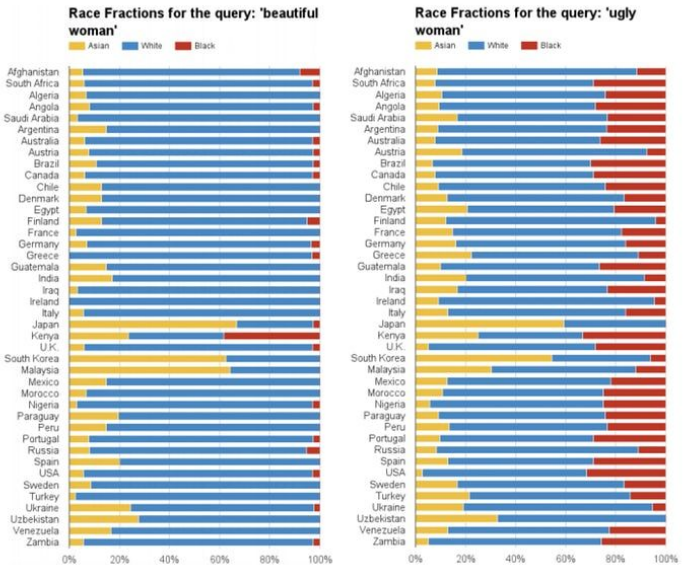
- Step in the right direction
- Human in the loop (provides the DAG asks the question)
- One step closer to a process.

However

- We are not lawmakers or sociologists
- We still need to educate and get educated

Limitations

Different machine learning tasks



Limitations

Different machine learning tasks

Man is to Computer Programmer as Woman is to Homemaker?
Debiasing Word Embeddings

Tolga Bolukbasi[1], Kai-Wei Chang[2], James Zou[2], Venkatesh Saligrama[1,2], Adam Kalai[2]
[1]Boston University, 8 Saint Mary's Street, Boston, MA
[2]Microsoft Research New England, 1 Memorial Drive, Cambridge, MA
tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

Dependable Systems

- We should not view humans as the cause of problems
- Shift in direction for dependability
- Processes not only measures
  - Systems should enable humans to take actions
  - Open data/model provenance
  - Explanations

  - Systems should enable humans to say no
  - I do not want to see certain content. Do not use my data

  - Systems should be **trustworthy** in ways the average user can understand

Hard Problem

Society (as a consequence datasets) is unfair
Accountability is difficult (who do we blame?)
Datasets and models are hard to understand

Thank You!