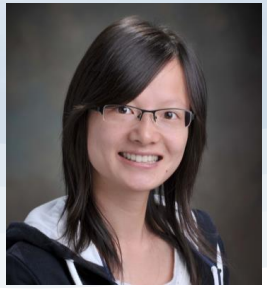
Fairness in Healthcare AI

#FairMLHealth



Muhammad Aurangzeb Ahmad [1,2], Carly Eckert MD MPH [3], Christine Allen [1], JuHua Hu [4], Vikas Kumar [1], Ankur Teredesai [4]

1. KenSci Inc.
2. Department of Computer Science, University of Washington Bothell
3. Department of Epidemiology, University of Washington
4. Department of Computer Science and Systems, University of Washington Tacoma

PAKDD 2021

#FairMLHealth Tutorial Overview

- Foundations: Fairness in Healthcare ML
- Measurement & Mismeasurement of Fairness
- Operationalizing Fairness in Healthcare ML
- Best Practices
- Library Demo
- Conclusion

Foundations: Fairness in Healthcare ML

Elements of Ethical ML in Healthcare

Explainable &
Transparent

Fair &
Unbiased

Robust

Privacy &
Security

Bias & Discrimination in Healthcare: History

The seminal figures of modern medicine (Anton van Leeuwenhoek(1632-1723), the Father of Microscopy, Marcello Malphighi(1628-1694), the Father of Histology, Carl Linnaeus(1707-1778) the Father of Biological Classification) held racial and biased beliefs that greatly influenced modern medicine and healthcare (Byrd et al 2001).

Education: Many Western physicians assumed poor health as normal for Black populations ("Negro Diseases"). This was part of medical schools' syllabi until the 1960s in the US (Savitt 2002).

Medical Profession: With few exceptions Blacks were not represented in the medical profession in the US until the late 19th century and the percentage in the profession remained at 2% from 1900 to 1980.

Sterilization: A third of Puerto Rican women of childbearing age were sterilized under coercion from 1930s to 1970s. Many Mexican and Native American women were also sterilized (de Malave 1999). International examples of sterilization pf indigenous people are abundant e.g., India in mid70s (Wilson 2017).

Fatality: Higher prevalence of death during childbirth and lower birth weight babies among pregnant Black women (Randall 1995).

Bias & Discrimination in Healthcare: History

Tuskegee Experiments: From 1932 to 1972, the US government tracked and deceived 600 hundred low-income Black men in Tuskegee, AL for a study where sham treatments were given for Syphilis. Many men needlessly passed the disease to their family, suffered and died (Thomas and Quinn 1991).

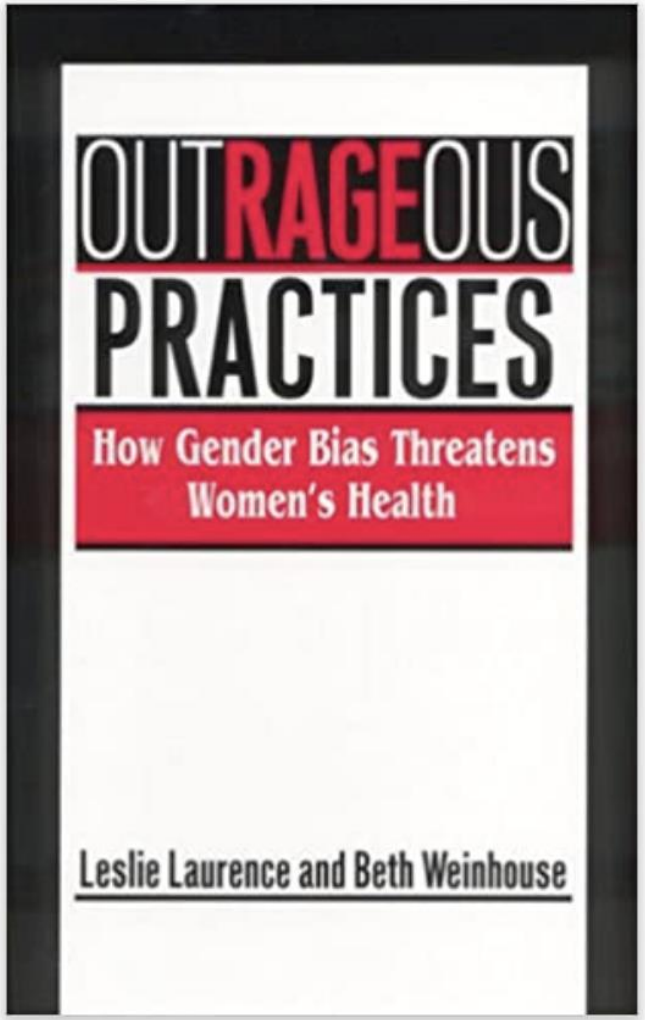
Sickle Cell Disease: which mostly affects Black populations, received less attention in research than other prominent diseases, mainly because its disproportionately affects people of color (Wailoo 2017).

Cardiac Bypass Surgery: Physicians refer significantly less Black men for cardiac bypass surgery than white men. This is due to the incorrect perception that Black patients were less well-educated and less likely to engage in physical activity after the surgery. Thus, the physicians concluded that they were poorer candidates for the surgery (Malat and Griffin 2006).

Bias in Non-Human Healthcare Research: A 2011 literature survey of 10 different fields within Biology revealed that single-sex studies of male animals outnumbered those of females 5.5 to 1. Since the 1960s male bias in non-human studies had increased (Beery and Zucker 2011).

Bias & Discrimination in Healthcare: History

- Rockefeller University's NIH-supported study the role of obesity in breast and uterine cancer did not enroll women (Simkin 1995)
- Yet the study concluded that older women were less likely to be given lifesaving interventions as compared to men (Bierman 2007)
- Other studies observe that women are less likely to be given analgesia than men (Chen 2008)
- The 1982 Multiple Risk Factor Intervention Trial explored the impact of dietary and exercise in preventing heart disease included no women out of trial size of 13,000



Bias & Discrimination in Healthcare: History

- For most of the 15 leading causes of death in the US including heart disease, cancer, stroke, diabetes, kidney disease, hypertension, liver cirrhosis and homicide, Blacks have higher death rates than whites (Kung et al. 2008).

- These elevated death rates exist across the life-course with Blacks and American Indians having higher age-specific mortality rates than whites from birth through the retirement years (Williams 2005).

- Experiencing racist treatment affects health. Experience of interpersonal racism has been observed as a mechanism that partially explains differences between Aboriginal and non-Aboriginal peoples' health (Larson et al 2007).

- There is a long history of unfair diagnoses of psychological conditions in minorities and women (Gard et al 1997).

Bias in Healthcare AI: History

- One of the earliest examples (1970s) of algorithmic discrimination comes an algorithm used by St. George's Hospital Medical School in the UK. It was discriminating admittance decisions based on race and gender.

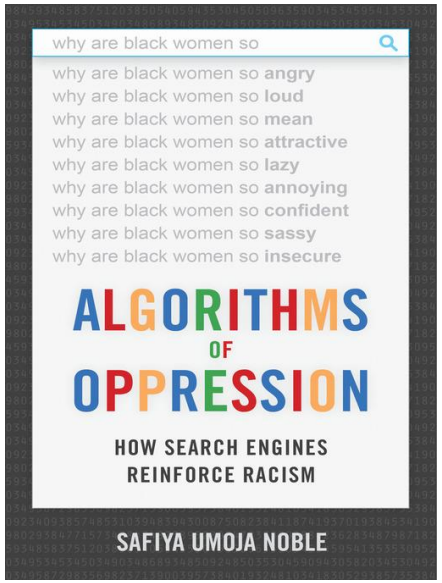
- In 1976 Joseph Weizenbaum raised the question of algorithmic bias, one of the first computer scientists to do so (Weizenbaum 1976).

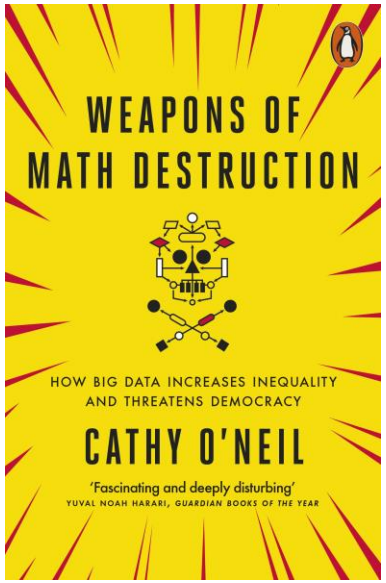
- Clinicians are more likely to believe AI that supports current practices and thus perpetuate implicit biases (Parikh 2019).

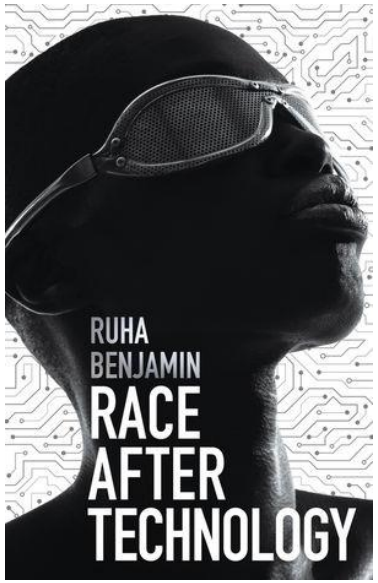
- Among women with breast cancer, Black women have a lower likelihood of being tested for high-risk mutations. An AI model that uses genetic tests as a predictor is more likely to mischaracterize the risk of breast cancer across groups defined by race (Parikh 2019).

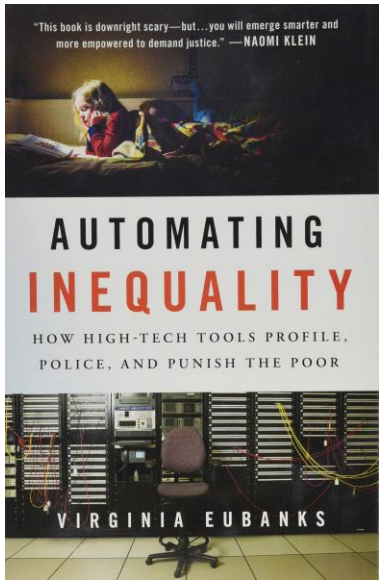
Bias in Healthcare AI: History

- Idahoans with cognitive or learning disabilities had their healthcare benefits reduced by $20,000—30,000 based on an AI algorithm without any explanation. This led to a lawsuit by ACLU that revealed the underlying algorithm (Stanley 2017).



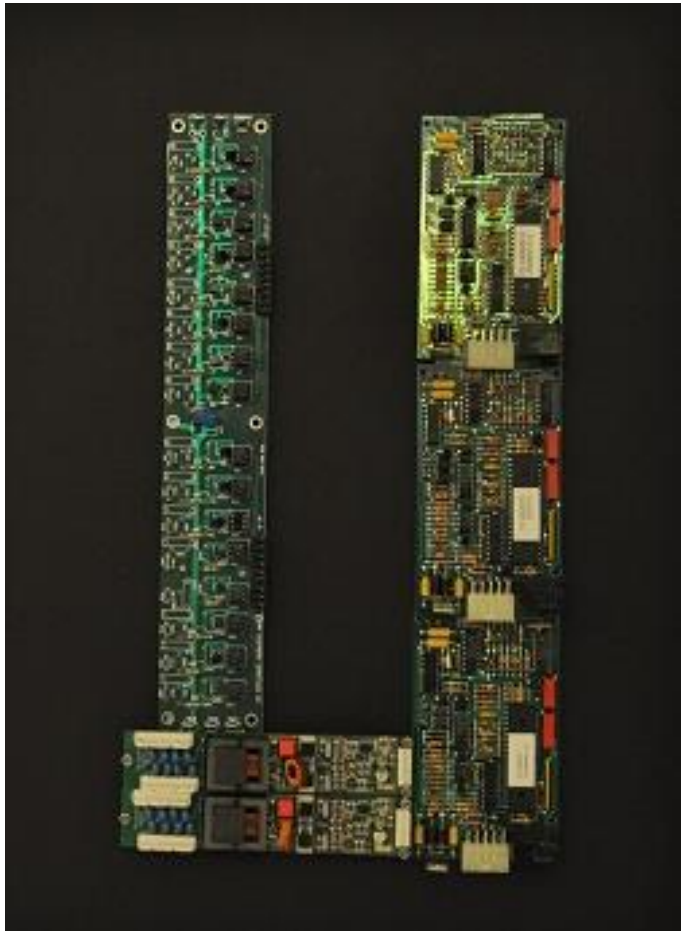




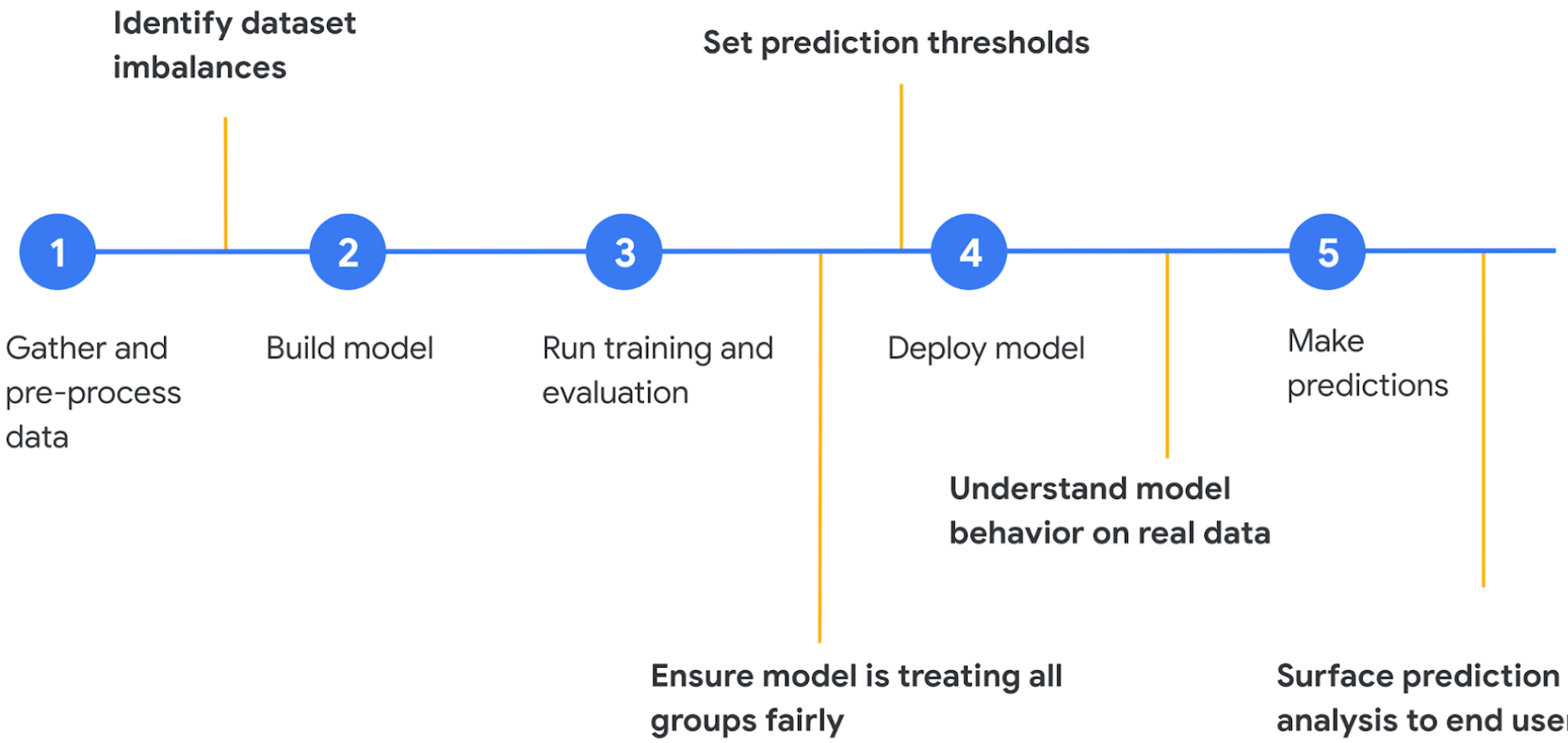


Making AI Fair in an Unfair World?

- Humans are biased and discrimination is a universal phenomenon
- People have implicit and explicit biases which permeate socio-technical systems
- Data points in healthcare represent human lives that may be affected by algorithmic decisions
- Thoughtful development and implementation of AI and models is required before AI in healthcare becomes pervasive



Fairness in Machine Learning is more than imbalanced datasets!



[Google Cloud 2018]

Bias in Healthcare AI: Is it just a data problem?

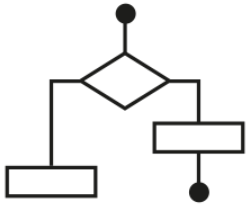
Generalizability and representativeness are also important considerations in healthcare. The generalizability of AI algorithms across subgroups is critically dependent on factors like representativeness of included populations, missing data, and outliers.

- EHRs are observational databases, the data reflects not just the health of the patients but also their interactions with the healthcare system e.g., the date of a code for diabetes is when the physician made the diagnosis, not when the patient first developed the disease (Agniel 2018).
- The billing code used for an office visit may be influenced by reimbursement policies in addition to the original reason for the visit.
- Practices regarding data capture may change over time e.g., reporting patient falls, opioid prescribing increased from 2005-12, but at rates that differed by practice and patient population (McLintock 2019).
- Data as a signal. Lab tests are ordered more often for sick patients (Agniel 2018).

Fairness in ML as a Systems Problem

ML Solution

Data



Algorithm

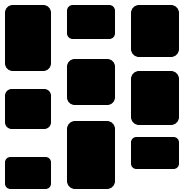
Model

Intervention

User

Domain Knowledge

Use Case



Context

Each constituent element can contribute towards making the solution unfair or biased

Legal Protected Classes in the US

- Race (Civil Rights Act of 1964)
- Color (Civil Rights Act of 1964)
- Sex (Equal Pay Act of 1963; Civil Rights Act of 1964)
- Religion (Civil Rights Act of 1964)
- National origin (Civil Rights Act of 1964)
- Citizenship status (Immigration Reform and Control Act 1965)
- Age (Age Discrimination in Employment Act of 1967)
- Familial status (Civil Rights Act of 1968)
- Disability status (Rehabilitation Act of 1973; Americans with Disabilities Act of 1990)
- Pregnancy (Pregnancy Discrimination Act 1978)
- Veteran status (Vietnam Era Veterans' Readjustment Assistance Act of 1974; Uniformed Services Employment and Reemployment Rights Act of 1994)
- Genetic information (Genetic Information Nondiscrimination Act of 2008)

[Barocas and Hardt 2017]

Legal Protected Classes Globally

Additional protected classes in other countries

- **Nigeria:** Constitution prohibits discrimination on grounds of political affinity and ethnic or tribal group
- **Portugal:** Ancestry, gender reassignment, economic status, education, social origin or status, genetic heritage, reduced working capacity, chronic disease, nationality, territory origin, ideological beliefs, union membership and maternity
- **Uganda:** HIV Status
- **Vietnam:** discrimination against outsourced employees is prohibited
- **Israel:** participation in military service (including military reserve duty)
- **India:** Scheduled castes and OBCs (Other Backward Classes) people is prohibited
- **Pakistan:** Discrimination against transgender people is prohibited

Fairness in the Age of COVID-19

Healthcare rationing: Due to stresses caused by the COVID-19 pandemic on national healthcare systems globally

- When limited resources in acute medical settings cannot be accessed by all patients who need them

Scenarios

- **ICU:** What happens when ICU demand exceeds the critical care facilities available? How should doctors decide between which patients to treat?
- **COVID-19:** Real world scenarios with COVID-19, insufficient knowledge about efficacy with insufficient supply of medications [White 2020]
- Randomized Centralized Lottery Allocation: Solves the problem of (i) Unfair Allocation (ii) Learn new knowledge about the underlying condition
- Vaccine allocation and historically marginalized populations

Fairness is Stakeholder Dependent

Physician: Of the patients that are labeled high risk of dying from COVID-19, how many are likely to be high risk?

Patient: What is the probability that I will be incorrectly labeled as low risk? Given that I am from a protected class, will I be given the same clinical services according to the best evidence

Societal (Group Fairness): Are the risks balanced across all protected classes?

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

[Narayanan 2018]

Dimensions of Fairness in Healthcare AI

- Computational
  - Data Bias
  - Model Bias
  - Loss Function Bias
  - Post-Hoc optimization
- Social/Institutional
  - Structural Bias
  - Cultural Practices
  - Embedded Practices
- Cognitive
  - Automation Bias
  - Automation Complacency
  - Delivery Bias

Bias in Computer Systems

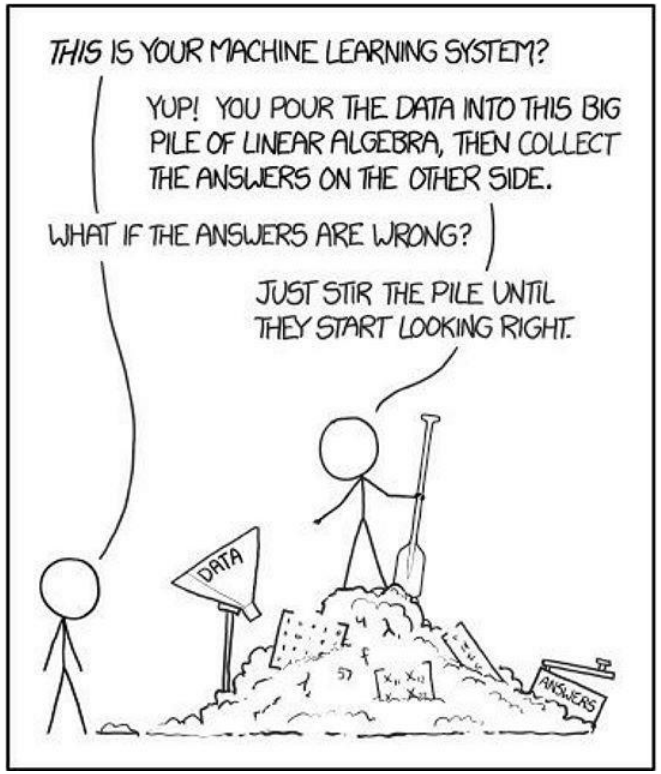
- Technical Bias
- Pre-existing Bias
- Emergent Bias

[Friedman & Nissenbaum, 1996]

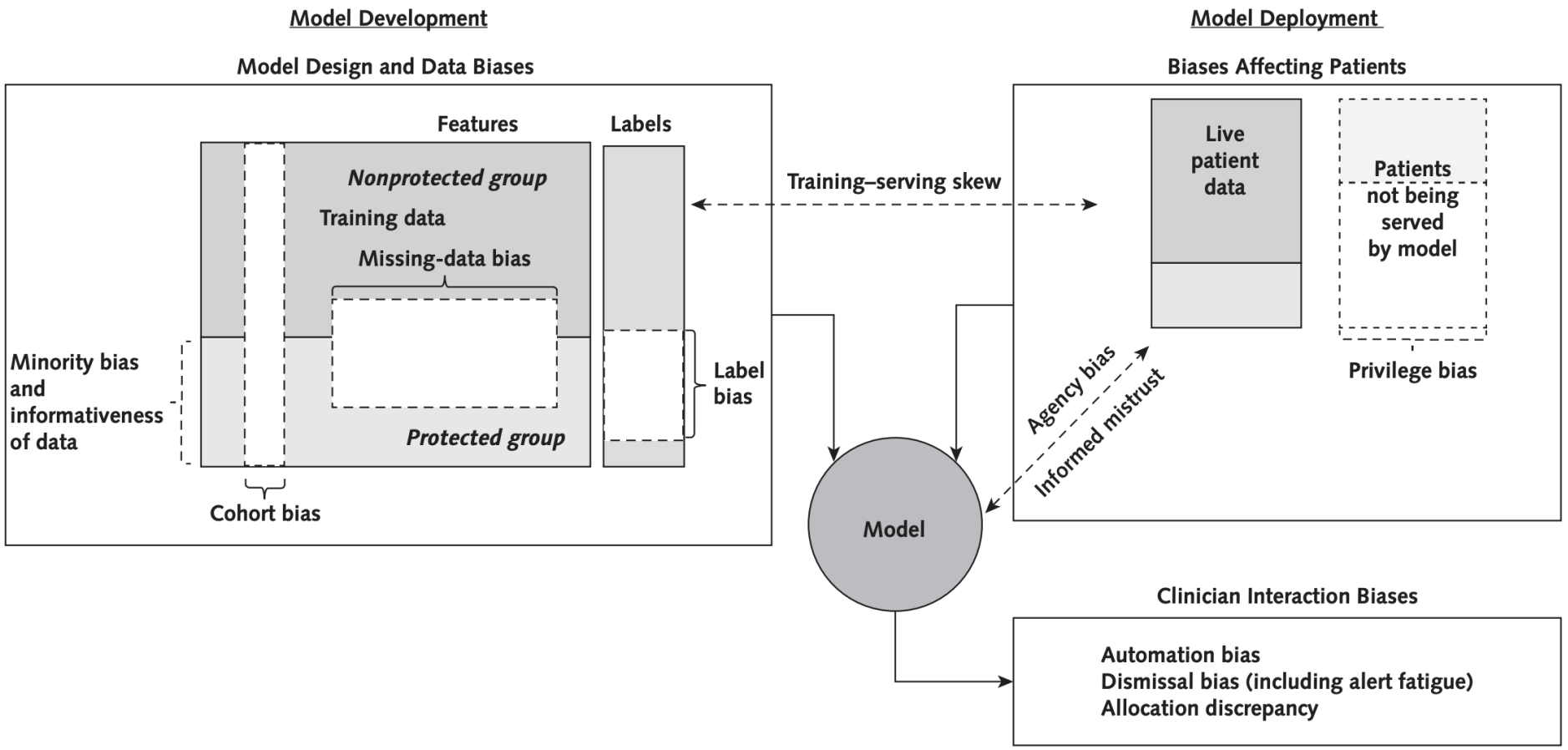
Weapons of Math Distraction ~ Cathy O'Neil

Dangers of making fairness problems as *only* technical problems to solve [Moritz Hardt]

- "Technical work without understanding social context"
- "Thinking we're more rigorous than social scientists"
- "Justifying an approach by the math it entails"
- "Big Data processes codify the past. They do not invent the future." [O'Neil 2016]



How Biases in Healthcare are Interrelated

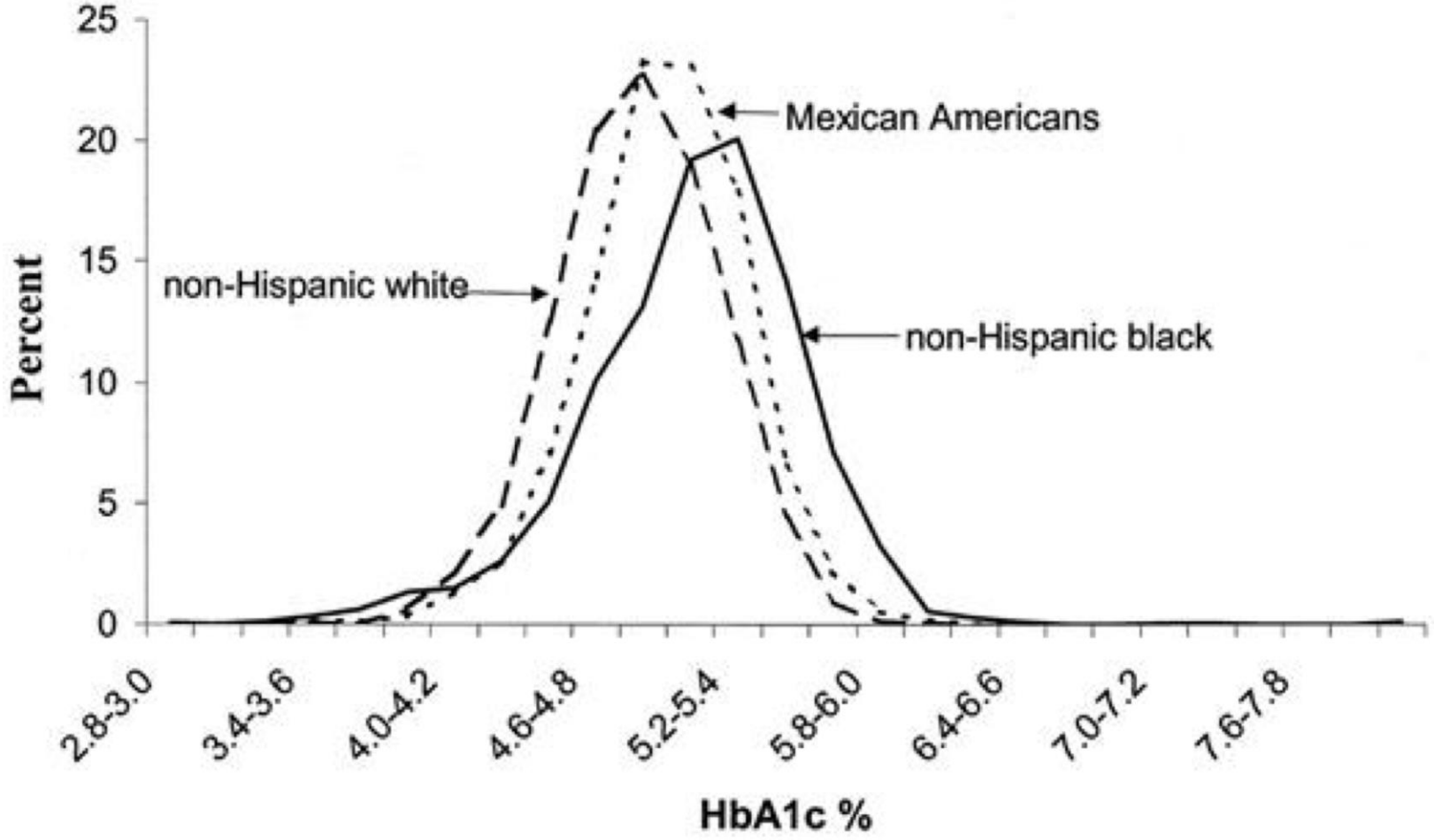


[Rajkomar et al 2018]

Sources of Bias in Healthcare AI

Bias in the ML Cycle	Data Bias	Non-Data Biases • **Model Bias** • **Loss Function Bias** • **Post-Hoc optimization**	Bias in Delivery • Cognitive Biases • Social Biases
Sources of Bias	• Selection/sample bias • Response bias • Publication bias • Prejudicial bias • Measurement bias • Hawthorne effect • Social desirability bias • Self-reporting bias	• Algorithmic Bias • Loss Function Bias • Post-Hoc Optimization	• Outcome Fairness • Lack of Understanding • Explainability • Lack of understanding/ Assume model is fair • Don't care
Bias Mitigation	• Equal representation	• Bias mitigation Algorithms • Fairness metrics • Explainable AI	• Acknowledgement & Explanation of bias during model delivery

AI needs FAT



HbA1c distribution by ethnicity in U.S. children and young adults ages 5–24 yr (NHANES-3, 1988–1994) [Saaddine et al., 2002]

Hemoglobin A1c (HbA1c): widely used as a measure of risk for the development of diabetic complications [Herman et al., 2012, Edelman et al., 2004, McCarter et al.. 2004]

Measurement & Mismeasurement of Fairness

Measurement & Mismeasurement of Fairness

Foundational Aspects of Fairness Measurement

How we are categorized through data affects how we will be treated

- Frank Pasquale in *The Black Box Society*

Discrimination: Treatment vs. Impact

Discrimination: The unjust or prejudicial treatment of different categories of people or things, especially on the grounds of race, age, or sex (Oxford Dictionary)

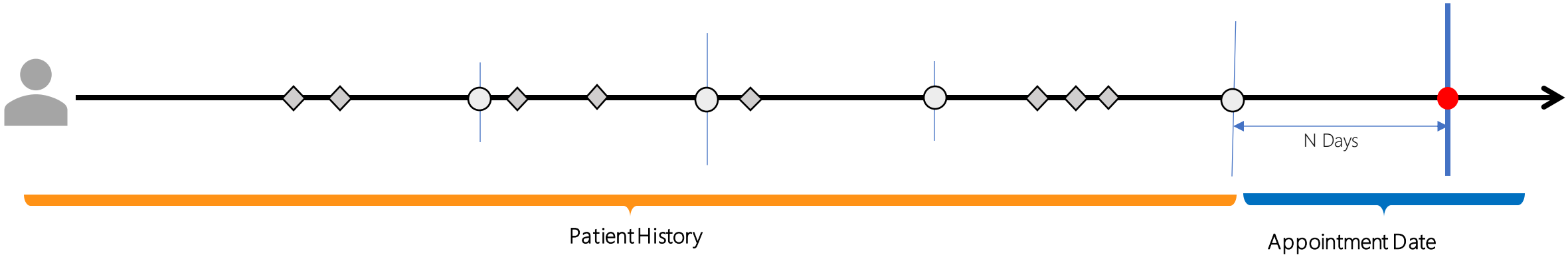
Disparate Treatment: The treatment depends on class membership
Example: implicit bias leading to differences in treatment in acute coronary syndrome

Disparate Impact: The treatment appears to be neutral, but it impacts the protected class
Example: hospital relocation and access to care for minority classes

[Barocas, S. and Selbst 2016]

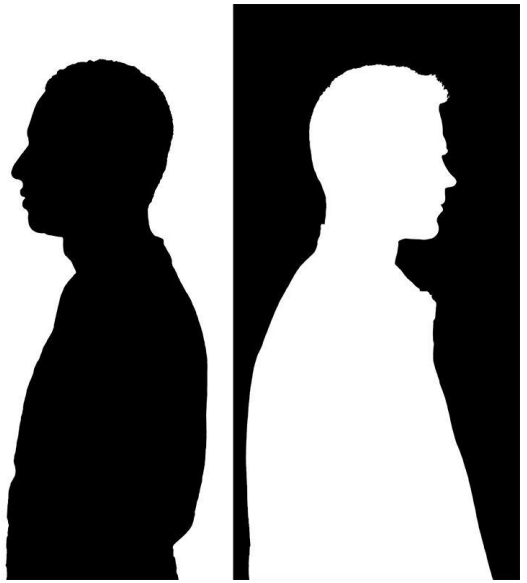
Tensions between disparate treatment and disparate impact

- Different groups may need to be treated differently to maintain fairness
- Humans (clinicians) deal with it on case-by-case basis. But this is not scalable for algorithmic decision making [Narayanan 2018]
- There is an element of subjectivity across clinicians while making such assessments
- Example: Patient "no show" prediction



Example: Differential Treatment by Race

- James, a 65-year-old Black male and David, a 65-year-old white male, both have coronary artery disease. They experience chest pain and shortness of breath and are rushed to the ED by their spouses.
- Both are seen by the same ED physician and are both diagnosed as having an acute myocardial infarction (a heart attack)*. Yet the clinical recommendations and interventions offered are different *and James is treated less aggressively*
- How do we determine if the two patients are treated fairly? [Arora et a. 2018]



For the purpose of this illustration, we are considering that all clinical factors are the same between these two patients.

Protected Classes & Proxy Variables

- Many variables of interest correlate with protected class.

Not all are considered illegitimate to use in decision making.
[e.g., educational qualifications in hiring decisions.]

- Many researchers have proposed methods to identify and mitigate “proxy discrimination”.

Protected Classes & Potential Surrogates
Age
Sex
Race / ethnicity
Insurance status
Disability, functional status
Zip code / census tract
Costs of care / utilization
Marital status
Disease conditions: HIV, mental health
Genetic results: BRCA

Model Performance and Fairness

Differences in performance

- Limited features
- Skewed distributions
- Limited data availability

Distribution of Error across sub-populations

- Different models with the same reported accuracy can have a very different distribution of error across population

Understanding disparities in predicted outcome

- Skewed Proxies
- External processes not captured in data

Fairness & Performance

Fairness Measurement
What are the different ways to measure Fairness

Predictive Performance
How well is the model performing

Calibration
How good is the model calibration

Intervention & Allocation
How are the insights from the model being used to intervene

Measurement & Mismeasurement of Fairness

Defining and Measuring Fairness

Fairness through Unawareness

- A predictor achieves fairness through unawareness if protected attributes are not explicitly used in the model
- There may still be many variables in the data that are proxies for protected variables e.g., sex, age etc.
- Inclusion of protected variables *may be necessary* to avoid discrimination
- Discrimination may be needed in order not to discriminate
- "There is no such thing as fairness through unawareness." - Moritz Hardt

Demographic Parity

Foundation

- Proportion of each protected class should receive the positive outcome at equal rates
- Inspired from the four-fifths rule
- Used to audit models for Disparate impact

When to use Demographic Parity

- Change the state of our current world to improve it
- Remedy historical biases that may have affected the quality of our data
- Prevent the reinforcement of historical biases

Notation

$X \in R^d$	Non-protected Features
$A \in \{0, 1\}$	A (binary) protected attribute
C	Binary classifier: $c(X,A) \in \{0,1\}$
$Y \in \{0, 1\}$	Target Variable
$(X, A, Y) \sim D$	Underlying distribution D
$P_0 [c]$	$P [c \mid A= 0]$

Formulation

$P_0 [C = c] = P_1 [C = c] \; \forall \; c \in \{0,1\}$

Alternate Nomenclature

Independence, Statistical Parity

Equalized Odds

Foundation

- C is independent of A conditional on Y:
- Since the definition is restrictive the following relaxed version is often used
- $P_0 [C = 1 | Y = 1] = P_1 [C = 1 | Y = 1]$ which is called Equality of Opportunity

When to use Equalized Odds

- When ensuring that accuracy is equally high in all demographics even punishing models that perform well only on the majority

Notation

$X \in R^d$	Non-protected Features
$A \in \{0, 1\}$	A (binary) protected attribute
C	Binary classifier: $c(X,A) \in \{0,1\}$
$Y \in \{0, 1\}$	Target Variable
$(X, A, Y) \sim D$	Underlying distribution D
$P_0 [c]$	$P [c \mid A= 0]$

Formulation

$P_0 [C = r \mid Y = y] = P_1 [C = r \mid Y = y] \; \forall \; r, y$

Alternate Nomenclature

Separation, Positive Rate Parity

Predictive Rate Parity

Foundation

- Y is independent of A conditional on C:
- This is equivalent to satisfying both
  $P_0 [Y = 1 | C = 1] = P_1 [Y = 1 | C = 1]$ and
  $P_0 [Y = 0 | C = 0] = P_1 [Y = 0 | C = 0]$

When to use Predictive Rate Parity

- When it is less important to balance across different demographic groups
- Use cases where differences in model performance across groups does not lead to discrimination

Notation

$X \in R^d$	Non-protected Features
$A \in \{0, 1\}$	A (binary) protected attribute
C	Binary classifier: $c(X,A) \in \{0,1\}$
$Y \in \{0, 1\}$	Target Variable
$(X, A, Y) \sim D$	Underlying distribution D
$P_0 [c]$	$P [c \mid A = 0]$

Formulation

$P_0 [Y = y | C = c] = P_1 [Y = y | C = c] \; \forall \; y, c \in \{0,1\}$

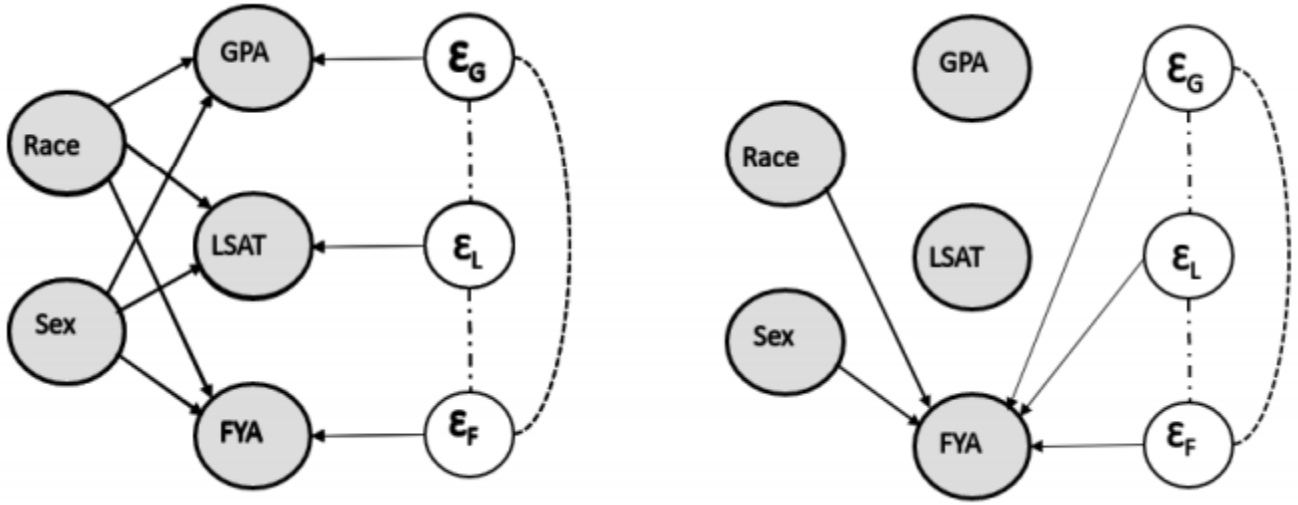
Alternate Nomenclature

Sufficiency

Counterfactual Fairness

- Measure fairness from the perspective of causes of bias

  $P[C_{\{A\leftarrow 0\}}=c \mid X, A=a]=P[C_{\{A\leftarrow 1\}}=c \mid X, A=a]$
- A counterfactual value replaces the original value of the sensitive attribute which propagates through the causal graph
- In practice it is difficult to determine what the causal graph should look like and or which features to use in the graph



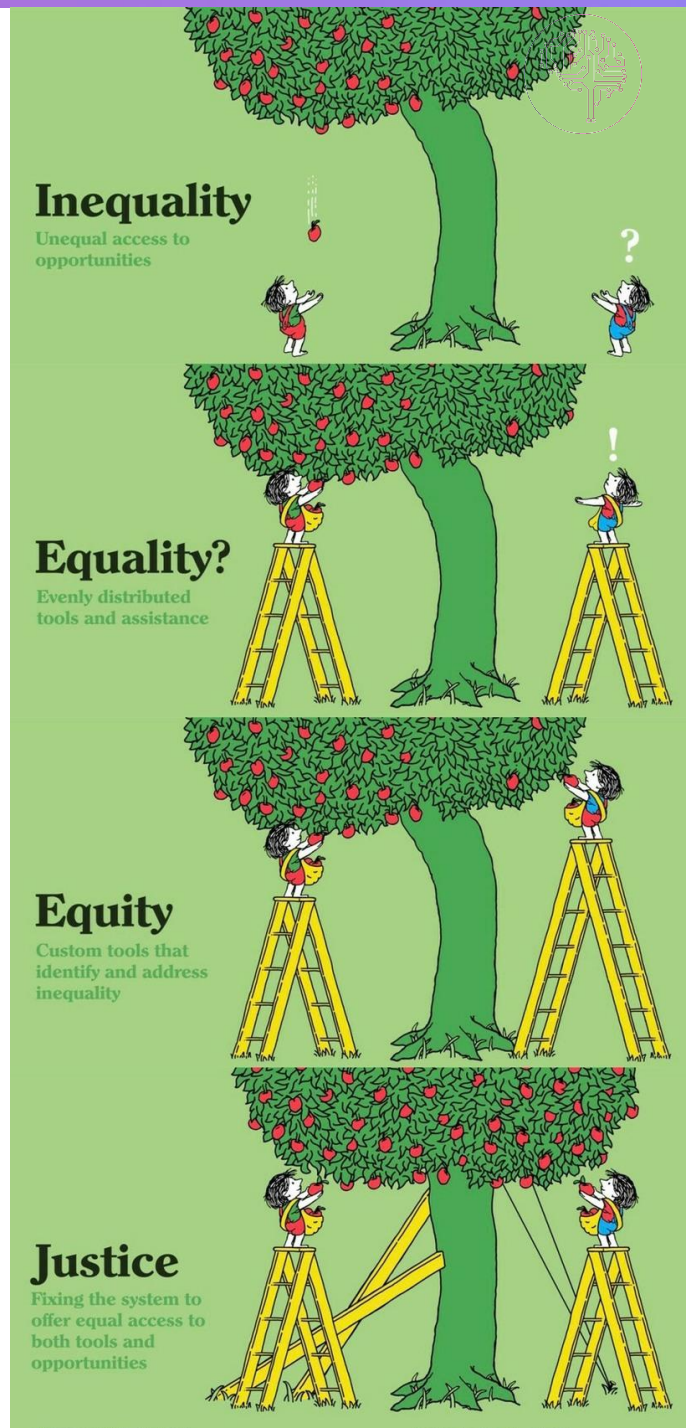
Individual Fairness

Formulation

- Similar individuals should be treated similarly
- Let Δ(O) to be the space of the distribution over measure space O. Let M:X→ Δ(O) maps each individual to a distribution of outcomes D(M(X),M(X'))≤ d(X,X')

Limitations

- Hard to determine what is an appropriate metric function to measure the similarity of two inputs
- Hard to determine which features to use to determine similarity



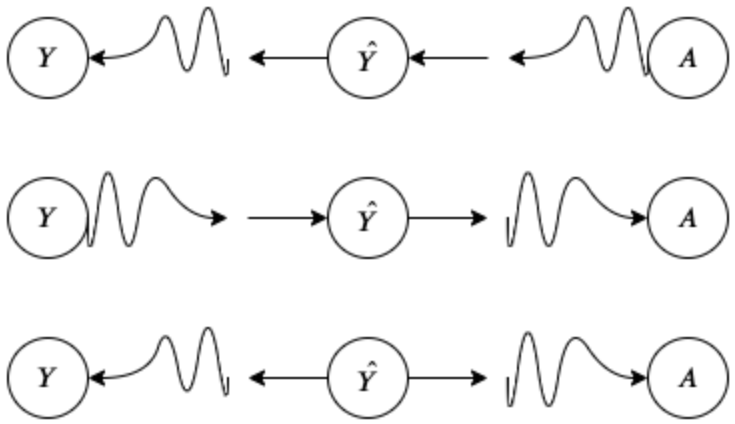
Type	Description	Formulation	Motivation	Flaws
Unawareness	Do not include the sensitive attribute as a feature in the training data	C=c(x, A) = c(X)	Intuitive, easy to use and legal support (disparate treatment)	There can be many highly correlated features(e.g. neighborhood) that are proxies of the sensitive attribute(e.g. race)
Demographic Parity / Independence / Statistical Parity	The outcomes must be equal		Legal Support: "four-fifth rule" prescribes that a selection rate for any disadvantaged group that is less than four-fifths of that for the group with the highest rate.	Ignores any possible correlation between Y and A e.g., rules out perfect predictor C=Y when base rates are different (i.e. $P_0[Y=1] \neq P_1[Y=1]$) laziness: if we hire the qualified from one group and random people from the other group, we can still achieve parity
Equalized odds / Separation / Positive Rate Parity	Different groups deal with similar odds	C is independent of A conditional on Y: $P_0[C = r \mid Y = y] = P_1[C = r \mid Y = y] \; \forall \; r, y$	Optimality compatibility: C=Y is allowed. Penalize laziness: it provides incentive to reduce errors uniformly in all groups.	It may not help closing the gap between two groups
Predictive Rate Parity / Sufficiency	The performance of the predictive model should be the same for different groups	Y is independent of A conditional on C: $P_0[Y = y \mid C = c] = P_1[Y = y \mid C = c] \; \forall \; y, c \in \{0,1$	Optimality compatibility: C=Y satisfies Predictive Rate Parity. Equal chance of success(Y=1) given acceptance(C=1)	It may not help closing the gap between two groups
Individual Fairness	similar individuals should be treated similarly	$D(M(X),M(X')) \leq d(X,X')$	Rather than focusing on group, as individuals, we tend to care more about the individuals. Besides, individual fairness is more fine-grained than any group-notion fairness	It is hard to determine what is an appropriate metric function to measure the similarity of two inputs
Counterfactual Fairness	How do the outcome change if the values of the sensitive variables change	P[C_{A← 0}=c|X, A=a]=P[C_{A← 1}=c|X, A=a]	Counterfactual fairness provides a way to check the possible impact of replacing only the sensitive attribute	The idea is very ideal. In practice, it is hard to reach a consensus in terms of what the causal graph should look like and it is even harder to decide which features to use

Measurement & Mismeasurement of Fairness

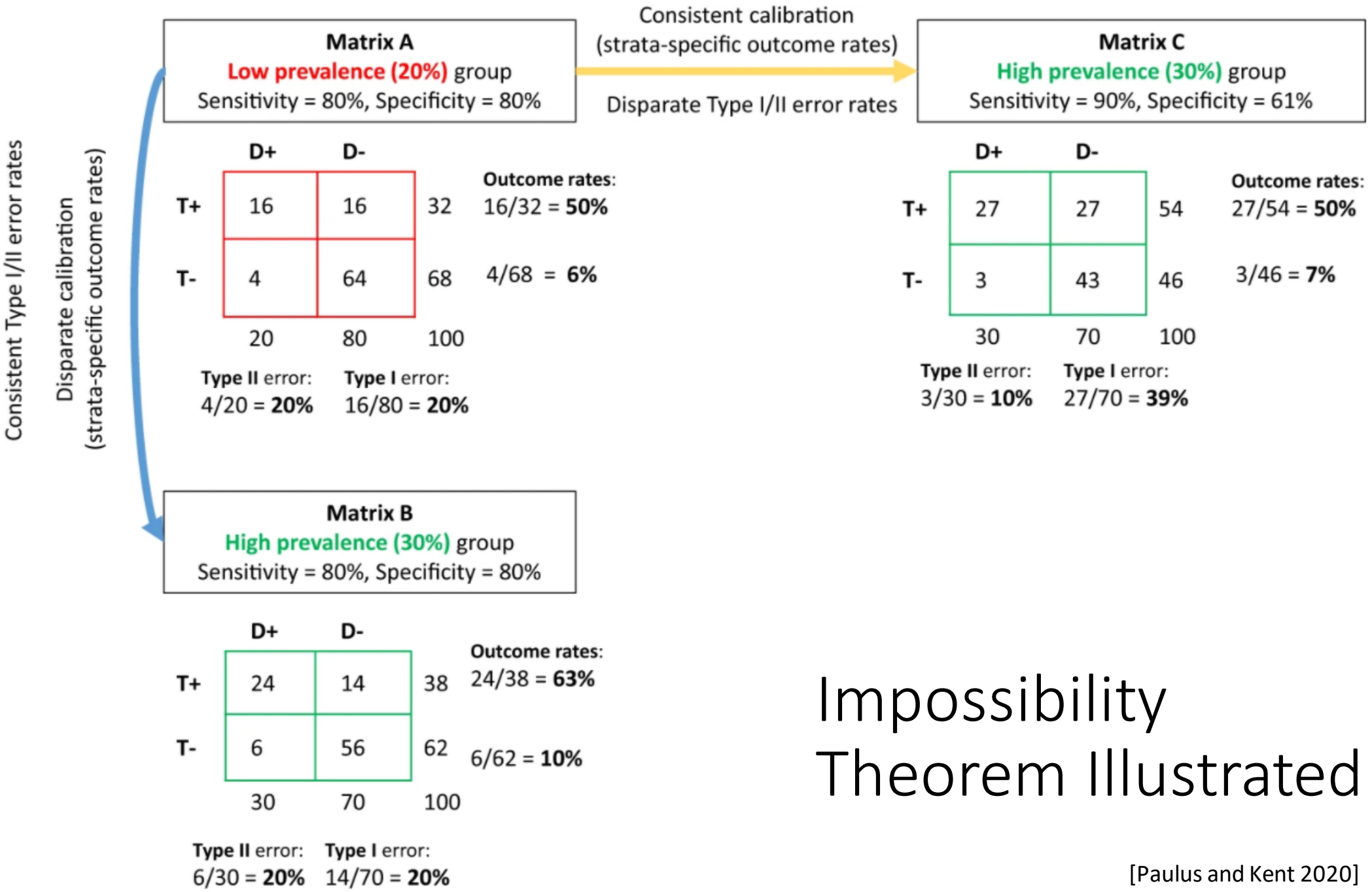
Impossibility Theorem(s) of Fairness

Impossibility Theorem of Fairness (in ML)

- No more than one of the three fairness metrics of demographic parity, predictive parity and equalized odds can hold at the same time for a well calibrated classifier and a sensitive attribute [Kleinberg et al 2016; Chouldechova 2017]



- It is impossible for a single data generation processes to satisfy all three group fairness metrics [Karthik S 2020]



Impossibility Theorem Illustrated

[Paulus and Kent 2020]

Impossibility Theorem of Privacy & Fairness

- It is not possible for a classifier to have differential privacy and satisfy group fairness conditions
- Kuppam et al. empirically show privacy-fairness. For multiple settings and using census data to which noise has been added to demonstrate how adding noise to achieve differential privacy can adversely affects fairness for minority groups
- “Even under a very simple binary classification setting no learning algorithm that is -differentially private (for any ≥ 0) and that is guaranteed to output a fair classifier (for any reasonable notion of fairness) can have non-trivial accuracy”

[Cummings et al 2019; Kuppam et al 2019; Agarwal 2020]

Impossibility of Fairness and Calibration

It is not possible to satisfy any major conditions of calibration and fairness simultaneously. Major notions of Calibration and Fairness

- **Group Calibration:** For each group *t*, and each bin *b* with associated score v_b, the expected number of people from group t in b who belong to the positive class should be a v_b fraction of the expected number of people from group t assigned to *b*
- **Negative Class Balance:** Requires that the average score assigned to people across groups belonging to the negative class should be the same
- **Positive Class Balance:** Requires that the average score assigned to people across groups belonging to the negative class should be the same

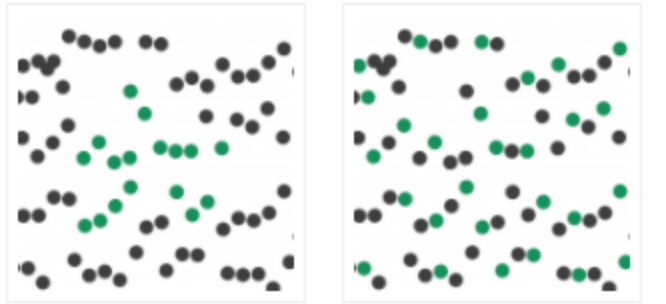
[Kleinberg et al 2016]

Measurement & Mismeasurement of Fairness

Measurement and Biases

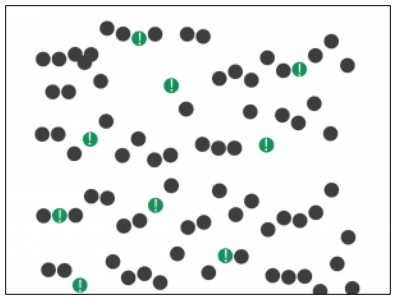
Data Biases: Statistical Biases

Selection Bias

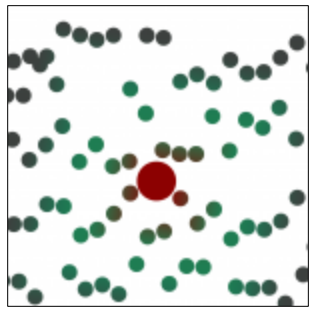


selection bias *proper random sampling*

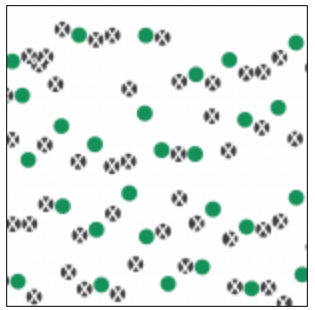
Self-Selection Bias



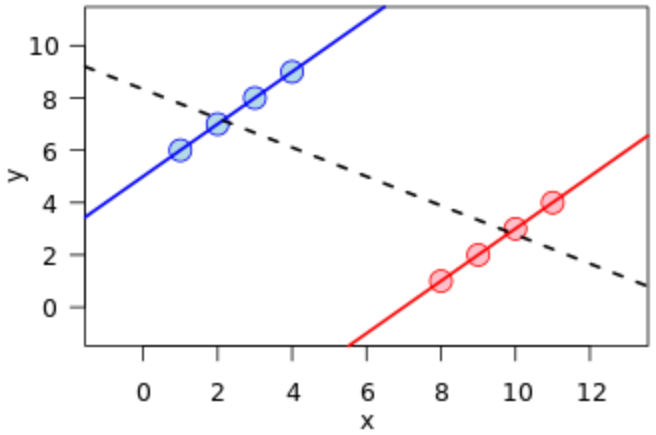
Observer Bias



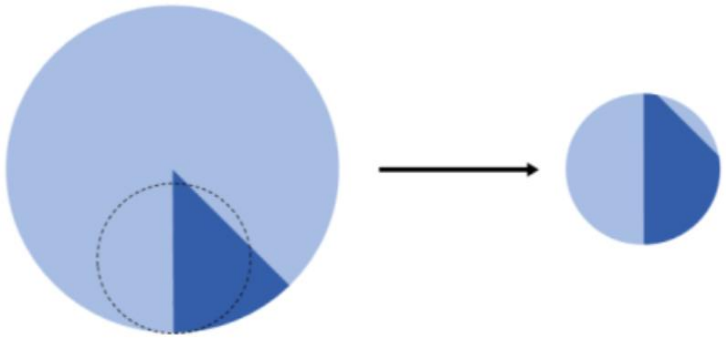
Survivorship Bias



Simpson's Paradox



Sampling Bias

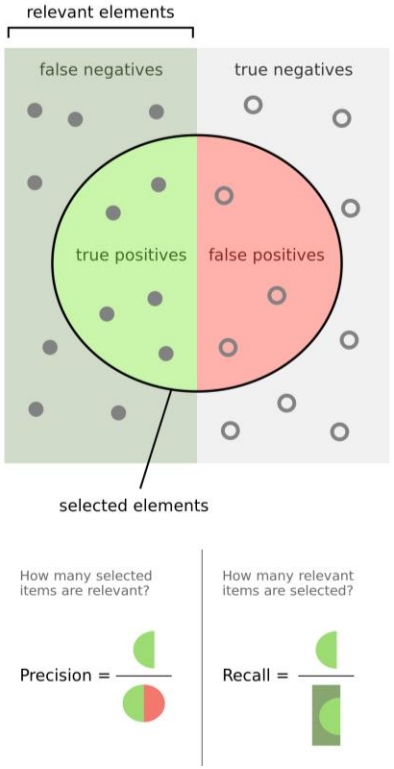


Cause Effect Bias

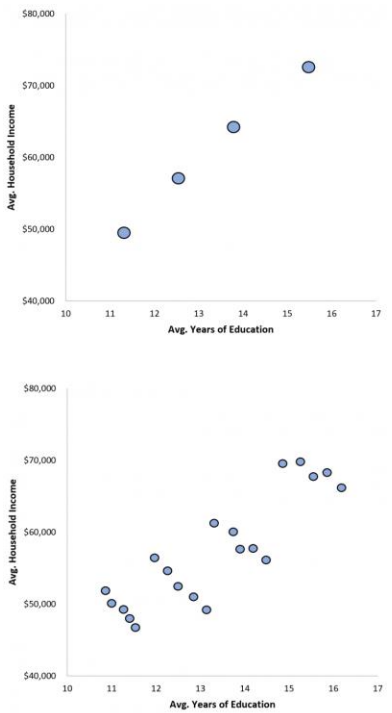
[Crawford 2017; Gebru et al 2018]

Data Biases: Statistical Biases

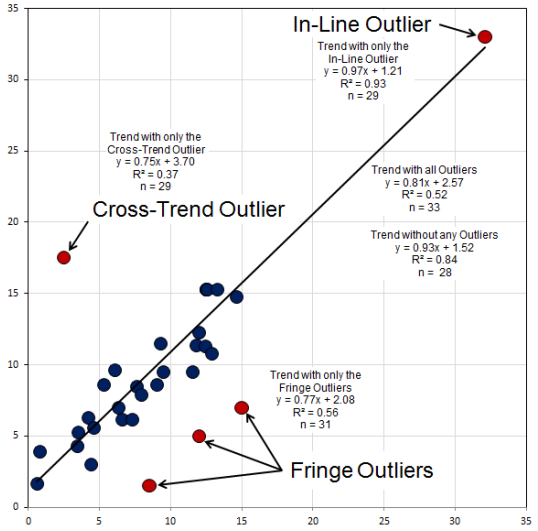
Evaluation Bias



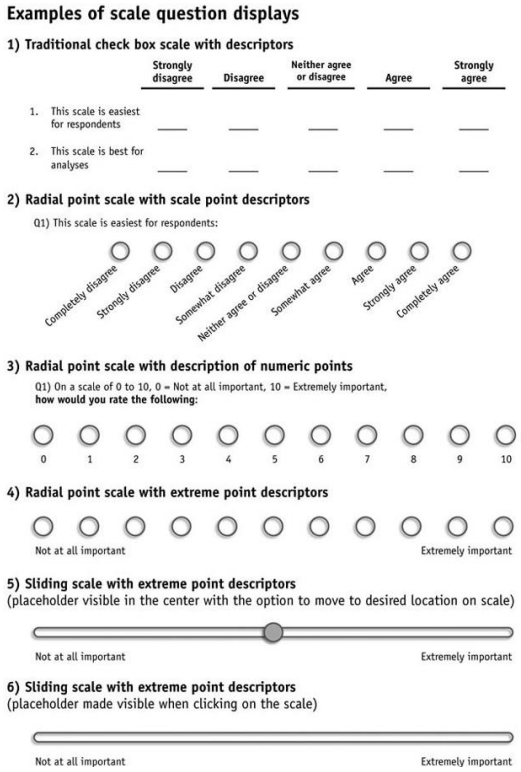
Aggregation Bias



Outlier Bias



Response Bias



Cognitive Biases

1. Anchoring bias.

People are **over-reliant** on the first piece of information they hear. In a salary negotiation, whoever makes the first offer establishes a range of reasonable possibilities in each person's mind.

2. Availability heuristic.

People **overestimate the importance** of information that is available to them. A person might argue that smoking is not unhealthy because they know someone who lived to 100 and smoked three packs a day.

3. Bandwagon effect.

The probability of one person adopting a belief increases based on the number of people who hold that belief. This is a powerful form of **groupthink** and is reason why meetings are often unproductive.

4. Blind-spot bias.

Failing to recognize your own cognitive biases is a bias in itself. People notice cognitive and motivational biases much more in others than in themselves.

13. Placebo effect.

When **simply believing** that something will have a certain effect on you causes it to have that effect. In medicine, people given fake pills often experience the same physiological effects as people given the real thing.

14. Pro-innovation bias.

When a proponent of an innovation tends to **overvalue its usefulness** and undervalue its limitations. Sound familiar, Silicon Valley?

15. Recency.

The tendency to weigh the **latest information** more heavily than older data. Investors often think the market will always look the way it looks today and make unwise decisions.

16. Salience.

Our tendency to focus on the **most easily recognizable features** of a person or concept. When you think about dying, you might worry about being mauled by a lion, as opposed to what is statistically more likely, like dying in a car accident.

5. Choice-supportive bias.

When you choose something, you tend to feel positive about it, even if that **choice has flaws**. Like how you think your dog is awesome — even if it bites people every once in a while.

6. Clustering illusion.

This is the tendency to **see patterns in random events**. It is key to various gambling fallacies, like the idea that red is more or less likely to turn up on a roulette table after a string of reds.

7. Confirmation bias.

We tend to listen only to information that confirms our **preconceptions** — one of the many reasons it's so hard to have an intelligent conversation about climate change.

8. Conservatism bias.

Where people favor prior evidence over new evidence or information that has emerged. People were **slow to accept** that the Earth was round because they maintained their earlier understanding that the planet was flat.

17. Selective perception.

Allowing our expectations to **influence how we perceive** the world. An experiment involving a football game between students from two universities showed that one team saw the opposing team commit more infractions.

18. Stereotyping.

Expecting a group or person to have certain qualities without having real information about the person. It allows us to quickly identify strangers as friends or enemies, but people tend to **overuse and abuse** it.

19. Survivorship bias.

An error that comes from focusing only on surviving examples, causing us to **misjudge a situation**. For instance, we might think that being an entrepreneur is easy because we haven't heard of all those who failed.

20. Zero-risk bias.

Sociologists have found that **we love certainty** — even if it's counterproductive. Eliminating risk entirely means there is no chance of harm being caused.

9. Information bias.

The tendency to **seek information when it does not affect action**. More information is not always better. With less information, people can often make more accurate predictions.

10. Ostrich effect.

The decision to **ignore dangerous or negative information** by "burying" one's head in the sand, like an ostrich. Research suggests that investors check the value of their holdings significantly less often during bad markets.

11. Outcome bias.

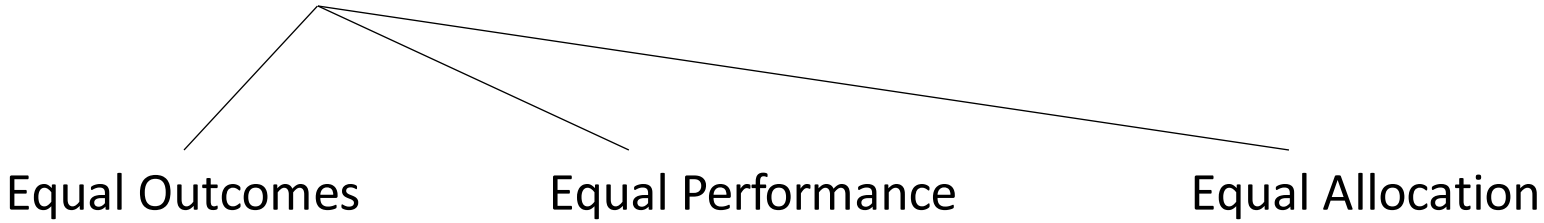
Judging a decision based on the **outcome** — rather than how exactly the decision was made in the moment. Just because you won a lot in Vegas doesn't mean gambling your money was a smart decision.

12. Overconfidence.

Some of us are **too confident about our abilities**, and this causes us to take greater risks in our daily lives. Experts are more prone to this bias than laypeople, since they are more convinced that they are right.

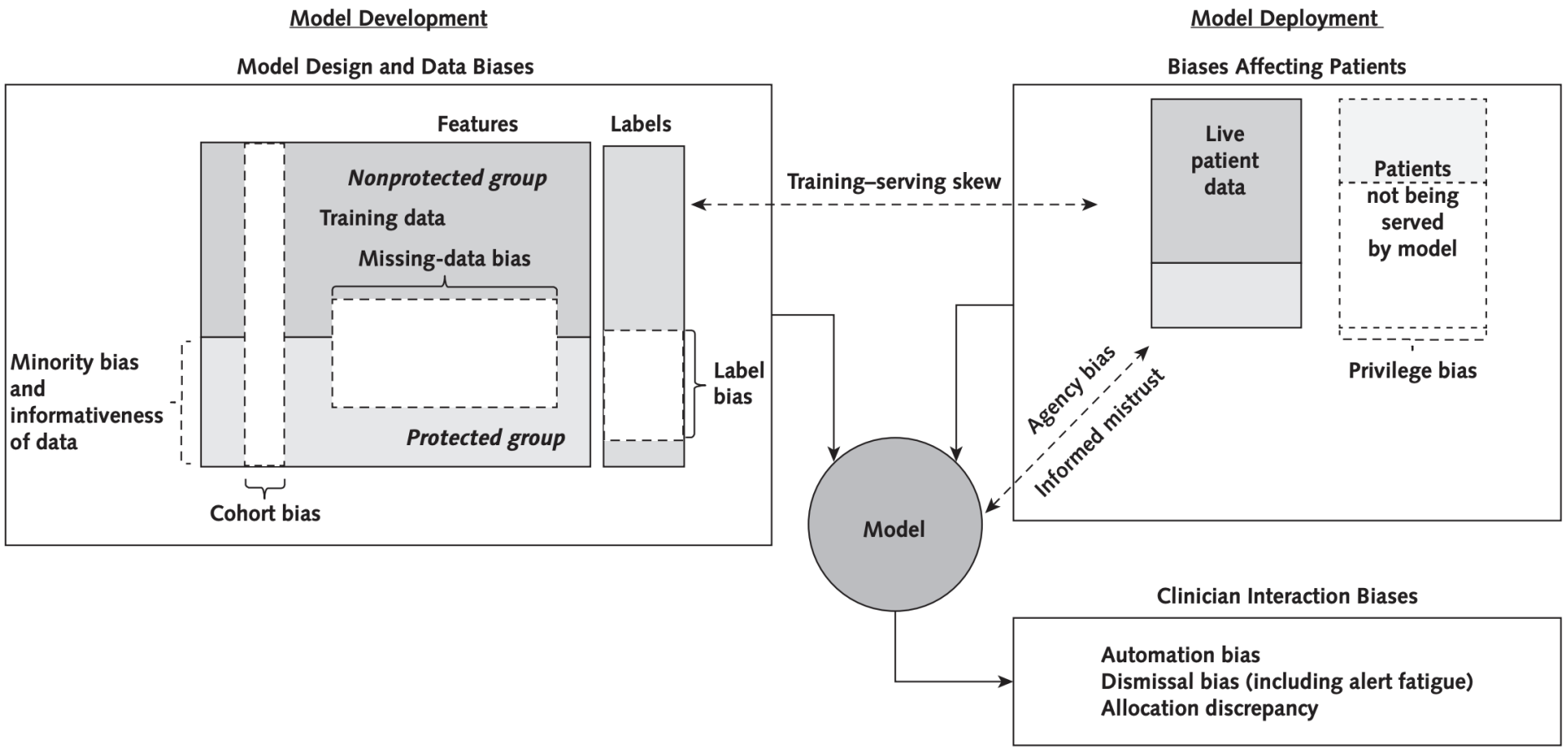
Bias: Outcomes & Clinical Perspective

- Model Performance
- Allocation of Services
- Clinical Outcomes



[Rajkomar et al 2018]

How Biases in Healthcare are Interrelated



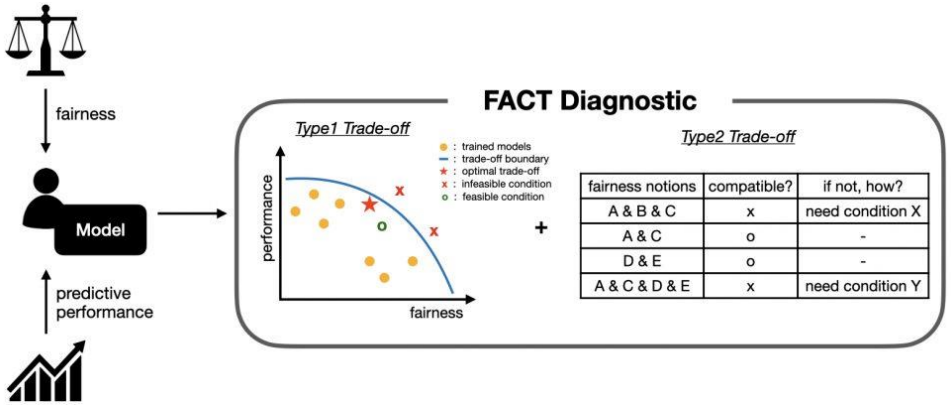
[Rajkomar et al 2018]

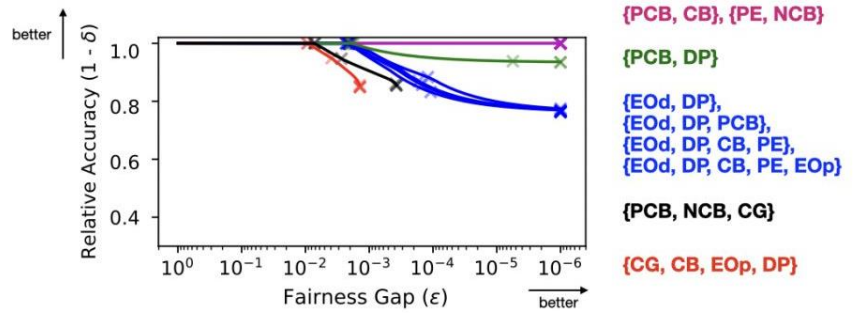
Measurement & Mismeasurement of Fairness

Tradeoffs of Applying Fairness in Healthcare ML

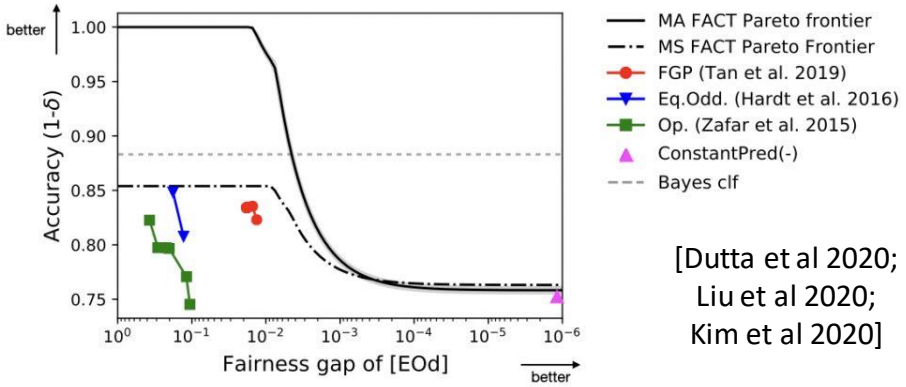
Fairness vs. Performance Trade-off

- Predictive performance of a model depends on data, the algorithm, and selected hyperparameters
- In general, fairness negatively impacts performance because it diverts the objective from accuracy *only* to both accuracy and fairness
- Trade-off are also present between the different notions of group fairness in conjunction with model performance
- We can define these trade-off as an optimization problem so that theoretical results for limits of trade-off are possible





PCB: Positive Class Balance, NCB: Negative Class Balance, CB: Class Balance,
PE: Predictive Equality, DP: Demographic Parity, EOd: Equalized Odds,



[Dutta et al 2020;
Liu et al 2020;
Kim et al 2020]

Fairness vs. Explainability Trade-off

- The relationship between interpretability and fairness is complex
- Follow four different trends depending on the correlations between protected, non-protected attributes and class labels
- Interpretability-fairness trade-offs do not depend on group imbalance
- Global Shapley values can be interpreted as each feature's marginal contribution to the overall demographic disparity in the model
- SHAP can be modified to explain fairness

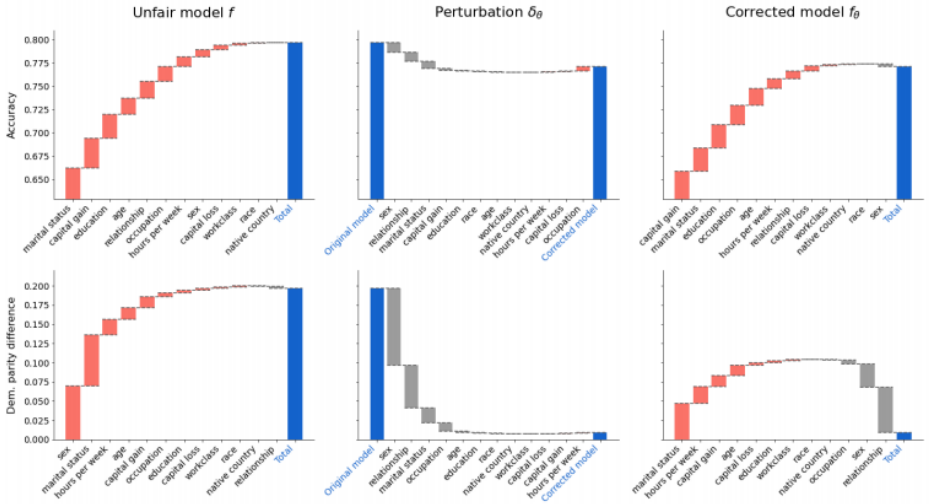


Figure 1: Explaining accuracy and unfairness (demographic parity) using Shapley values.

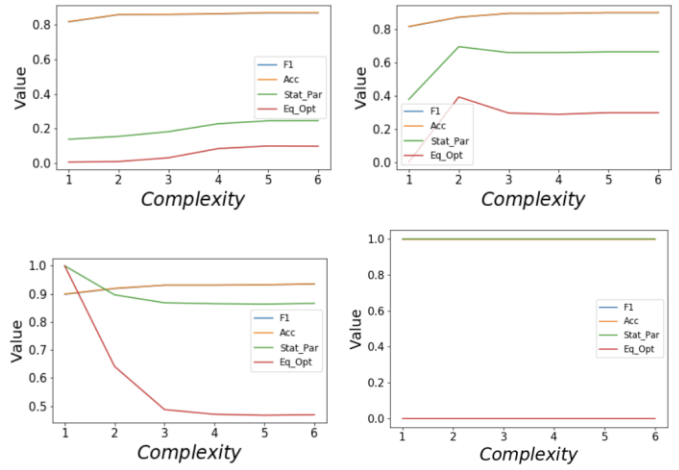
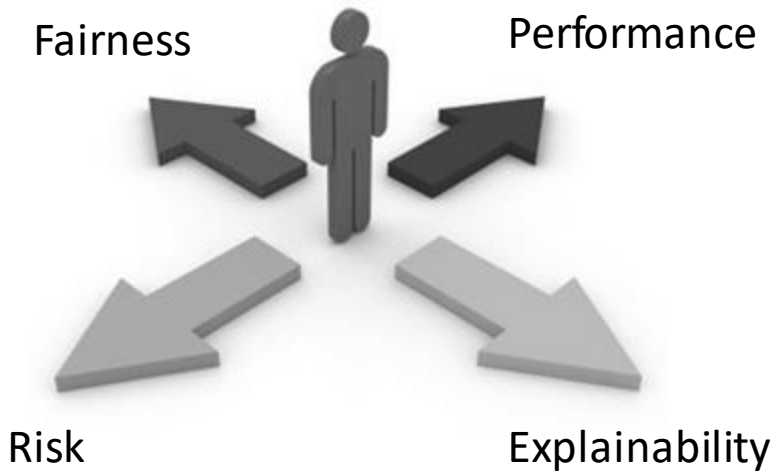


Figure 2. The effect of increasing the predictive power of the protected attribute p. C = logistic regression. $\sigma^2 = 10$ and $r = 2$. $p = 0.6$ (upper left), $p = 0.8$ (upper right), $p = 0.9$ (lower left) and $p = 0.999$ (lower right).

[Jabbari et al 2020; Begley et al 2020]

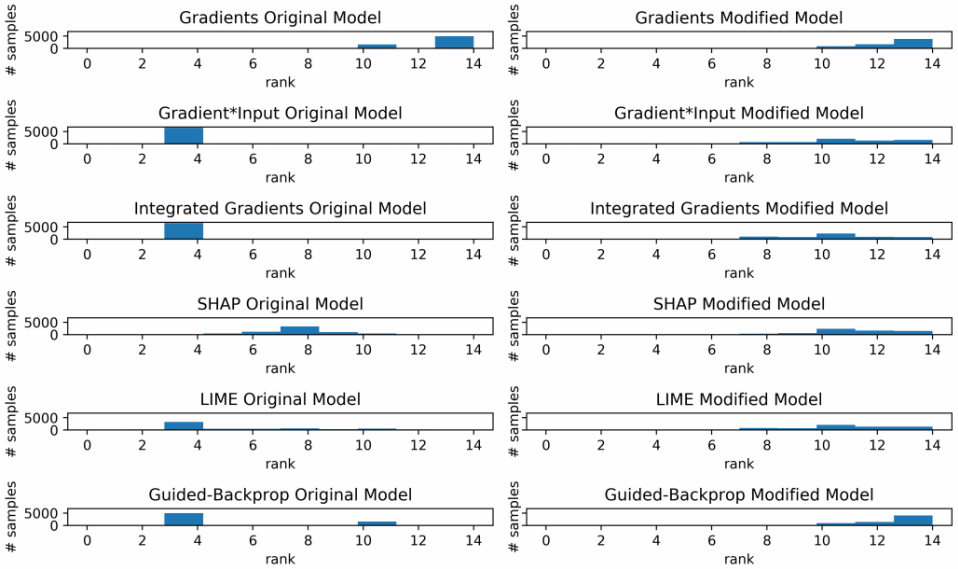
Fairness vs. Explainability Trade-off

- Explainability of ML models intends to bring about greater scrutiny of models and thus the possibility of fair and equitable models
- However, simplification of models may also bring about performance degradation as well as less fair models [Kleinberg and Mullainathan 2019]
- **The trade-off in healthcare is thus four way**: Fairness vs. Performance vs. Explainability vs. Risk
- Domain specific guidance should be used to help navigate these complex trade-offs



Limits of Fairness via Explanations

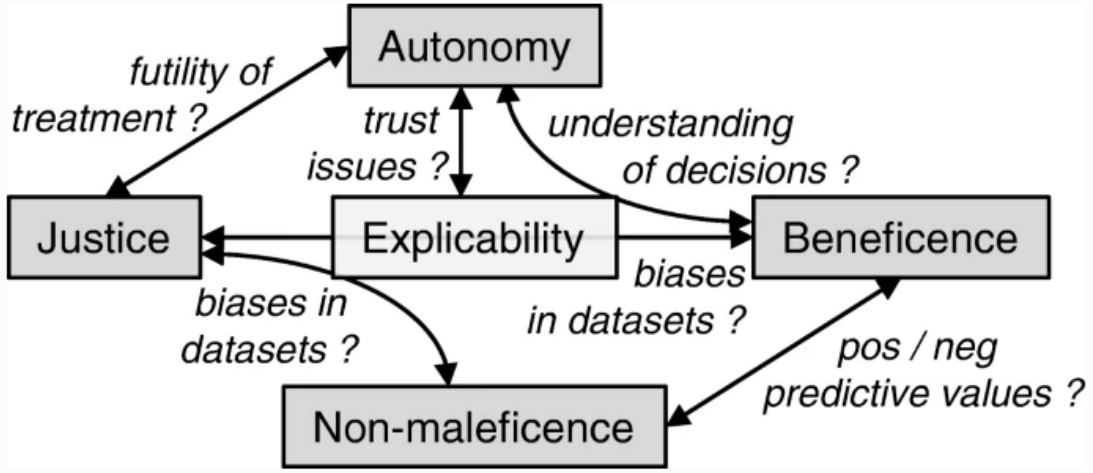
- Dimanov et al. showed that existing explainability models like LIME, SHAP etc. methods are not suited for fairness
- The authors retrained a model with an additional penalty term corresponding to the influence the protected attribute has on the output
- The explanation for that feature can be suppressed without substantially affecting the model predictions
- Thus, feature importance of the protected attribute is a poor measure of fairness. [Dimanov et al. 2020]



Importance ranking histograms for gender as the sensitive feature the original (left) and modified (right) models.

Fairness Trade-offs & Beneficence

- ***Beneficence:*** An ethical principle that providers must do everything they can to benefit the patient
- The removal/reduction of bias could possibly reduce predictive performance and undermining the principle of beneficence
- **Challenge:** How do we simultaneously reduce bias and maintain satisfactory model prediction performance (upholding beneficence)?



[Beil et al 2019]

Operationalizing Fairness in Healthcare ML

Operationalizing Fairness in Healthcare ML

Fairness from ML Pipeline Perspective

Being good is easy, what is difficult is being just.

- Victor Hugo

Bias Mitigation Strategies for ML Models

Pre-Processing	In-Processing	Post-Processing
• Reweighing • Optimized Preprocessing • Learning Fair Representations • Disparate Impact Remover	• Adversarial Debiasing • Prejudice Remover	• Equalized Odds Postprocessing • Calibrated Equalized Odds Postprocessing • Reject Option Classification

Fair ML via Pre-Processing

- **Reweighing**: Generate weights for the training data for each protected variable to ensure fairness before classification
- **Optimized Preprocessing**: Learn probabilistic transformations that minimally transforms the data while controlling for discrimination and limiting distortion in individual data samples
- **Learning Fair Representations**: Find latent representations that encodes the data well while obfuscating information about protected attributes
- **Disparate Impact Remover**: Edit features to increase group fairness while preserving rank-ordering within groups

Fair ML via In-Processing

- **Adversarial Debiasing**: Learn a classification model that simultaneously maximizes predictive performance and reduces an adversary's ability to determine the protected attribute from the predictions. Since the approach minimizes information that can be used to determine proxies for discrimination information w.r.t. an adversary and thus leads to a fair classifier
- **Prejudice removal via Regularization**: Add a discrimination-aware regularization term to the learning objective

Fair ML via Post-Processing

- **Equalized odds postprocessing**: Change target variables with certain probabilities (via linear programming) to optimize equalized odds
- **Calibrated equalized odds postprocessing**: Optimize calibrated classifier scores to find probabilities with which to change target variables with an equalized odds objective
- **Reject option classification**: For a confidence band around the decision boundary with the highest uncertainty, give favorable outcomes to unprivileged groups and unfavorable outcomes to privileged groups in

Operationalizing Fairness in Healthcare ML

Healthcare considerations in Operationalization

Outcome based fairness metrics revisited

- Fairness metrics like group fairness or even individual fairness optimize for immediate outcomes e.g., diagnosis, risk of readmission etc.
- Downstream effects are much harder to quantify
- The purpose of fairness metrics is to quantify the extent of the problem which is not equivalent to solving the problem
- Impossibly theorems do not imply that fairness in machine learning is not impossible but rather it is constrained by real world limitations

Cost Optimization is not need based optimization

- Algorithm scores are a key input to decisions about future enrollment in care coordination programs
- Less-healthy Blacks are scored at similar risk scores to more-healthy Whites which leads to disparities in program screening
- Algorithm's prediction on health needs is really a prediction on health costs
- At a given level of health (measured by number of chronic illnesses), Blacks generate lower costs than Whites on average ($1,801 less per year, holding constant the number of chronic illnesses) [Obermeyer et al 2019]

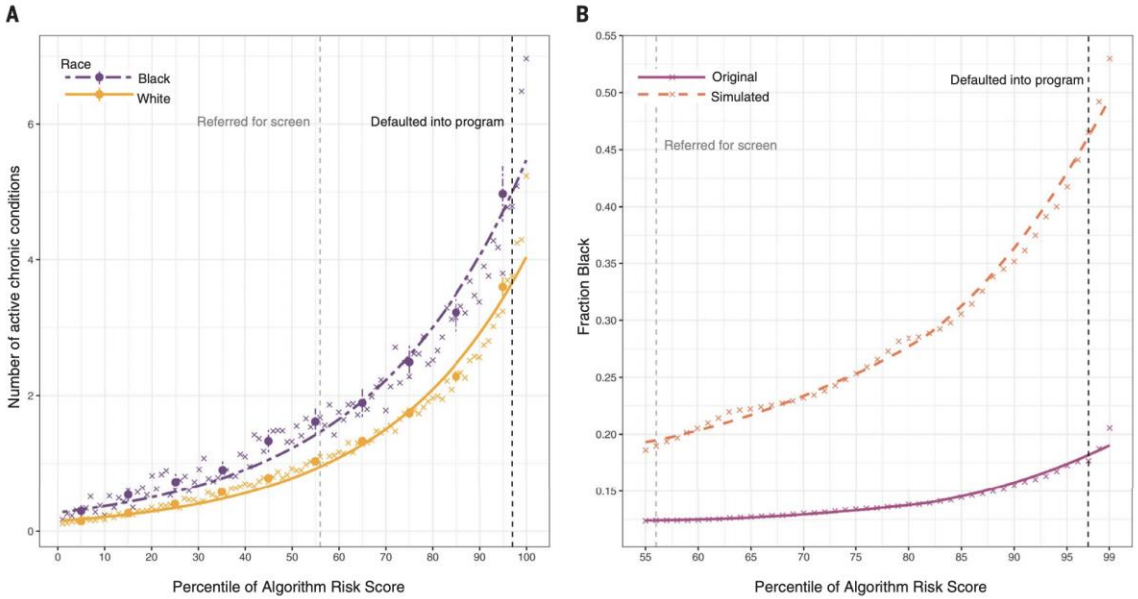
Healthcare Needs ≠ Healthcare Costs

- Black patients generate very different kinds of costs: Fewer inpatient surgical and outpatient specialist costs, and more costs related to emergency visits and dialysis
- “These results suggest that the driving force behind the bias we detect is that Black patients generate lesser medical expenses, conditional on health, even when we account for specific comorbidities”
- Result: Optimizing for costs does not always lead to fair outcomes

[Obermeyer et al 2019]

Treatment Effect is not monotonic

- The predicted risk of some future outcome e.g., healthcare needs is widely used to target policy interventions under the assumption that the treatment effect is monotonic
- However, this is however not always true
- At the same level of algorithm-predicted risk, Blacks have significantly more illness burden than Whites



Calculate an overall measure of health status, the number of active chronic conditions [or "comorbidity score," a metric used extensively in medical research to provide a comprehensive view of a patient's health] by race, conditional on algorithmic risk score.

[Obermeyer et al 2019]

Stakeholder Trade-off

- Fairness may also require trade-off *between* different stakeholders
- Optimizing for one sub-population may de-optimize for another population
- Even within the majority population there may be sub-groups which are not explicitly defined but which are vulnerable nonetheless

Example: No Show Prediction

- What a clinician wants to optimize for may be different from what staff planner may want to optimize for which may be different from what a patient from a marginalized group is optimizing for

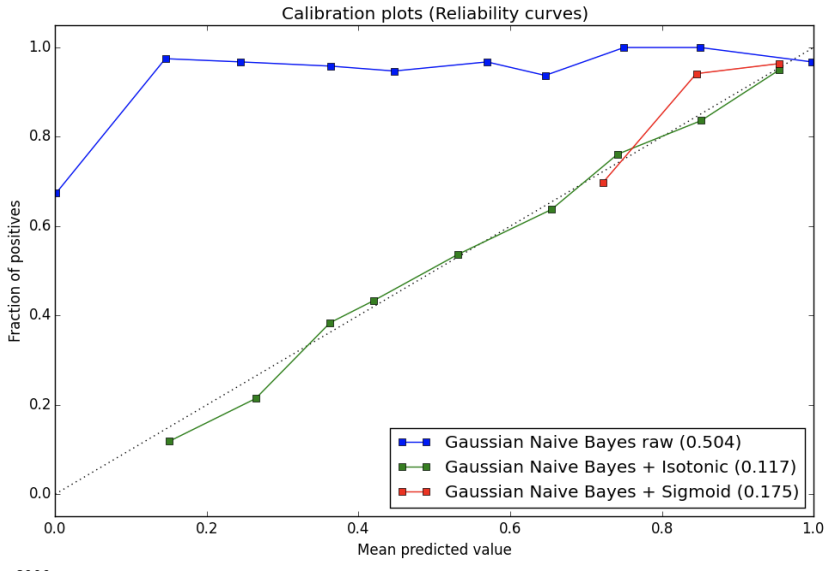
Operationalizing Fairness in Healthcare ML

Technical Challenges & Healthcare Impact

Fairness and Calibration

- Calibration: The set of people who receive a predicted probability of p, then fraction p of members of this set should be positive instances [Dawid 1982]
- Fairness between two groups G1 and G2 (e.g., Black and white patients) implies that this calibration condition to hold simultaneously for the set of people within each of these groups [Flores et al 2016]
- It is not feasible for certain notions of fairness

[Kleinberg et al 2016; Pleiss et al 2017]



Underspecification and Fairness

- Multiple ML models can roughly have the same performance in train-test but can vary wildly in performance when deployed
- "An ML pipeline is underspecified if there are many predictors *f* that a pipeline could return with similar predictive risk"

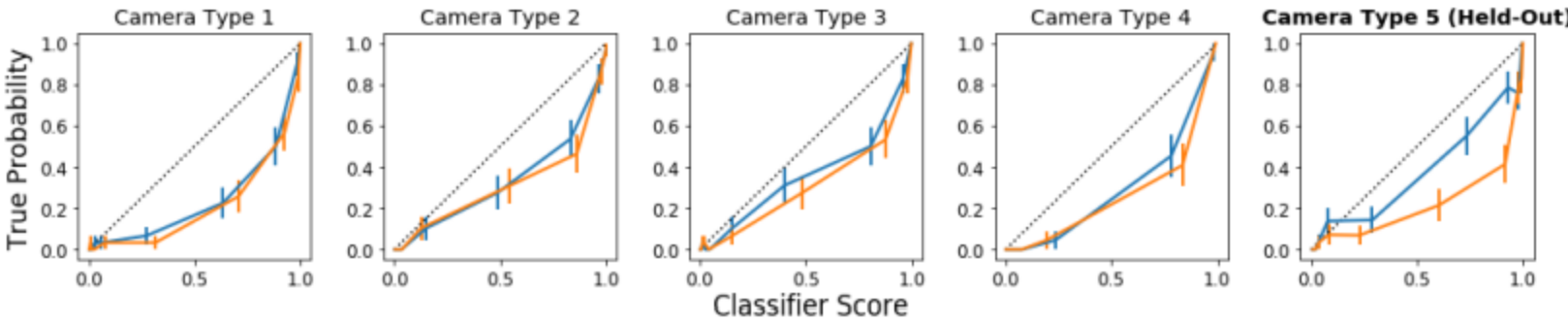


Figure 7: **Identically trained retinal imaging models show systematically different behavior on stress tests.** Calibration plots for two diabetic retinopathy classifiers (orange and blue) that differ only in random seed at fine-tuning. Calibration characteristics of the models are nearly identical for each in-distribution camera type 1–4, but are qualitatively different for the held-out camera type 5. Error bars are ±2 standard errors.

[D'Amour et al 2020; Marx et al 2020]

Underspecification and Fairness

- Stress tests (contrived testing conditions) can be done to determine how well will the models do in different scenarios
- Implications for Fairness: Predictive Multiplicity and underspecifiation can impact fairness and lead to unfair models

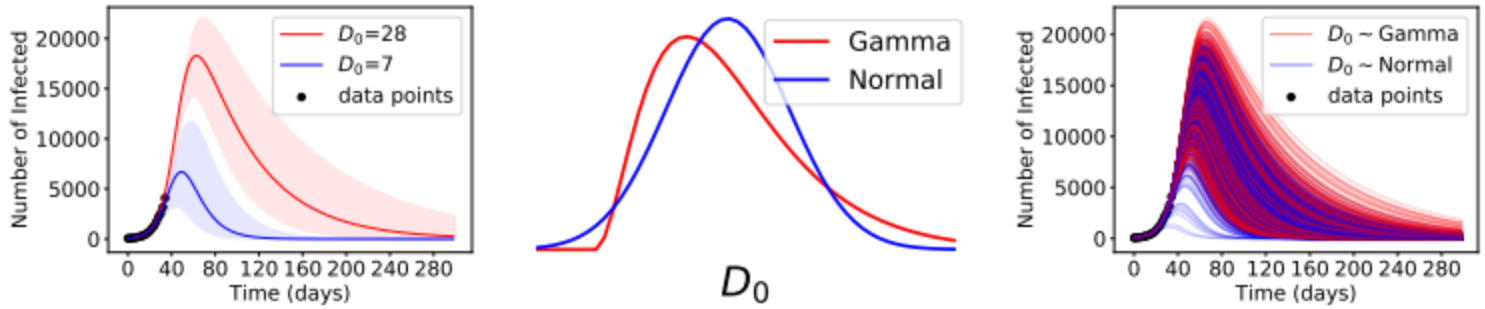
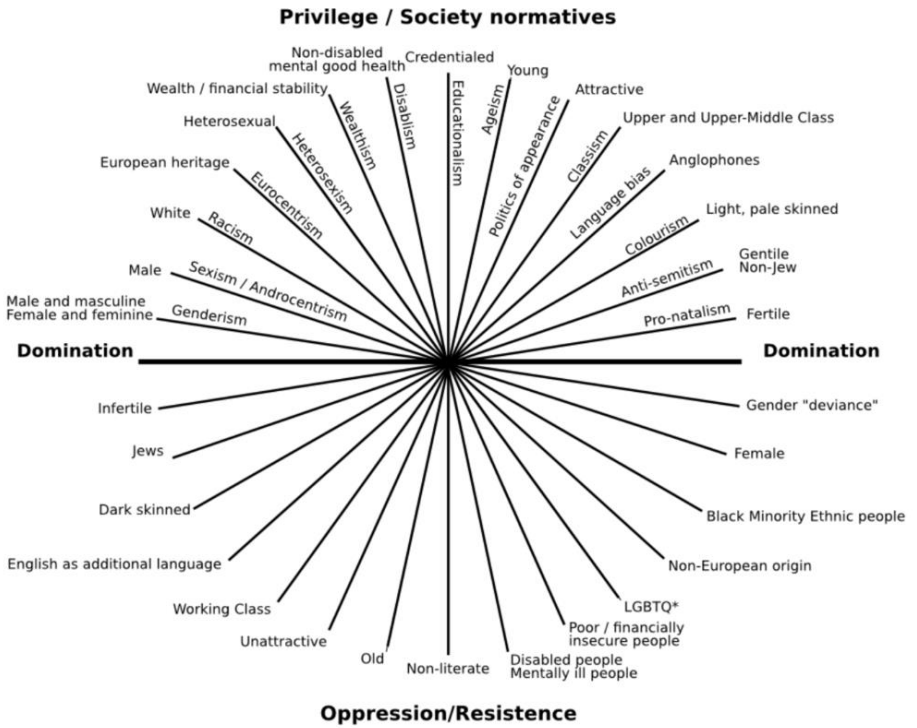


Figure 1: **Underspecification in a simple epidemiological model.** A training pipeline that only minimizes predictive risk on early stages of the epidemic leaves key parameters underspecified, making key behaviors of the model sensitive to arbitrary training choices. Because many parameter values are equivalently compatible with fitting data from early in the epidemic, the trajectory returned by a given training run depends on where it was initialized, and different initialization distributions result in different distributions of predicted trajectories.

[D'Amour et al 2020; Marx et al 2020]

Fairness Gerrymandering (Intersectionality)

- **Intersectionality:** the interconnected nature of social categorizations such as race, class, and gender as they apply to a given individual or group, regarded as creating overlapping and interdependent systems of discrimination or disadvantage [Oxford Dictionary]
- Intersectionality is susceptible to (intentional or inadvertent) **fairness gerrymandering** where a classifier appears to be fair on each individual group, but not for subgroups



Intersectionality in Healthcare ML

- Statistical notions of fairness across exponentially (or infinitely) many subgroups, defined by a structured class of functions over the protected attributes
- This interpolates between statistical definitions of fairness, and recently proposed individual notions of fairness, but it raises several computational challenges. It is no longer clear how to even check or audit a fixed classifier to see if it satisfies such a strong definition of fairness
- The Computational problem of auditing subgroup fairness for both equality of false positive rates and statistical parity is equivalent to the problem of weak agnostic learning (Computationally hard in the worst case)
- However, it also suggests that common heuristics for learning can be applied to successfully solve the auditing problem in practice [Kearns et al 2017]

Intersectionality: Differential Fairness

- **Differential Fairness:** A mechanism M(x) is ε-differentially fair (DF) with respect to (A, Θ) if for all θ ∈ Θ with x ~ θ, and y ∈ Range(M),

$$e^{-\epsilon} \leq \frac{P_{M,\theta}(M(\mathbf{x}) = y|\mathbf{s}_i, \theta)}{P_{M,\theta}(M(\mathbf{x}) = y|\mathbf{s}_j, \theta)} \leq e^{\epsilon}$$

Where, s_i , s_j ∈ A are tuples of all protected attribute values, Θ is a set of distributions θ which could plausibly generate each instance x i.e., regardless of the combination of protected attributes, the probabilities of the outcomes will be similar [Foulds et al 2019]

- Work on intersectional fairness in ML is relatively scarce

Intersectionality: Multiaccuracy

- **Multiaccuracy:** A strong notion of subgroup fairness. Models should be unbiased, overall as well as on but on every identifiable subpopulation
- Given: Black-box access to a classier *C*, and a relatively small validation set drawn from some representative distribution D
- Audit *C* to determine whether the predictor satisfies multiaccuracy.
- If auditing reveals that the predictor does not satisfy multiaccuracy, one could aim to post-process *C* to produce a new classier *C'* that is multiaccurate, without adversely affecting the subpopulations where *C* was already accurate [Kim et al 2019]

Intersectionality: Multiaccuracy

- Generate a synthetic disease outcome for each subgroup, divide the data set into subgroups (Gender & Age)
- For each subgroup, create synthetic binary labels using a different polynomial function of the input features with different levels of difficulty

	All	**F**	**M**	**O**	**Y**	**OF**	**OM**	**YF**	**YM**
$\mathcal{D}$	100	39.6	60.4	34.6	65.4	15.0	19.7	24.6	40.7
f_0	18.9	29.4	12.2	21.9	17.3	36.8	10.9	24.9	12.8
MA	16.0	24.1	10.7	16.4	15.7	26.5	9.0	22.7	11.6
SS	19.5	32.4	11.0	22.1	18.1	37.6	10.3	29.3	11.3

Table 5: **Results for UK Biobank semi-synthetic data set.** $\mathcal{D}$ denotes the percentages of each population in the data distribution; f_0 denotes the classification error (%) of the initial predictor; MA denotes the classification error (%) of the model after post-processing with MULTIACCURACY BOOST; SS denotes the classification error (%) of the subgroup-specific models trained separately for each population.

Process Fairness vs. Outcome Fairness

- Process Fairness is ensuring that the process is fair and not just the outcome
- One way to measure it is by estimating the degree to which people consider the usage various features to be fair in a model (intuitive moral sense)
- Let *U* denote the set of all members of society, and *F* denote the set of all possible features that might be used in the decision-making process
- **Feature-Apriori Fairness:** Without a priori knowledge of how feature usage affects outcomes

$$\text{feature-apriori fairness of } \mathcal{C}_{\mathcal{F}'} := \frac{\left|\bigcap_{f_i \in \mathcal{F}'} \mathcal{U}_{f_i}\right|}{|\mathcal{U}|}.$$

[Grgic-Hlaca et al 2016]

Process Fairness vs. Outcome Fairness

- **Feature-Accuracy Fairness:** Fair to use if it increases the accuracy of the classifier

$$\text{feature-accuracy fairness of } \mathcal{C}_{\mathcal{F}'} := \frac{|\bigcap_{f_i \in \mathcal{F}'} Condition(\mathcal{U}_{f_i}, \mathcal{U}_{f_i}^{Acc})|}{|\mathcal{U}|},$$

where

$$Condition(\mathcal{U}_{f_i}, \mathcal{U}_{f_i}^{Acc}) = \begin{cases} \mathcal{U}_{f_i} \cup \mathcal{U}_{f_i}^{Acc}, & \text{if } Acc(\mathcal{C}_{\mathcal{F}'}) > Acc(\mathcal{C}_{\mathcal{F}' \setminus \{f_i\}}) \\ \mathcal{U}_{f_i}, & \text{if } Acc(\mathcal{C}_{\mathcal{F}'}) \leq Acc(\mathcal{C}_{\mathcal{F}' \setminus \{f_i\}}). \end{cases}$$

- **Feature-Disparity Fairness:** Fair to use even if it increases a measure of disparity (i.e., disparate impact or disparate mistreatment) of the classifier

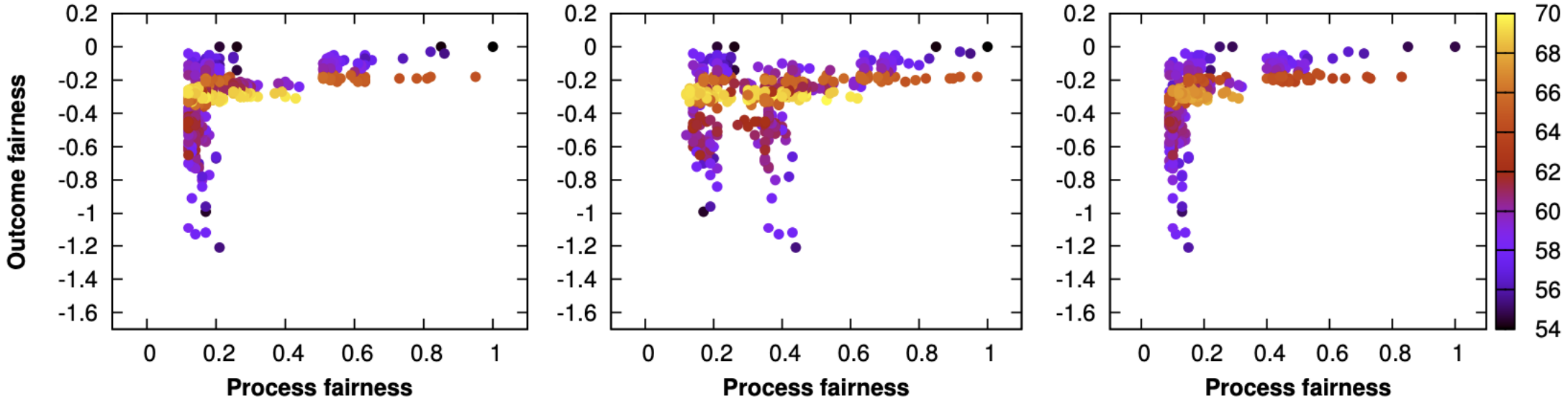
$$\text{feature-disparity fairness of } \mathcal{C}_{\mathcal{F}'} := \frac{|\bigcap_{f_i \in \mathcal{F}'} Condition(\mathcal{U}_{f_i}, \mathcal{U}_{f_i}^{Disp})|}{|\mathcal{U}|},$$

where

$$Condition(\mathcal{U}_{f_i}, \mathcal{U}_{f_i}^{Disp}) = \begin{cases} \mathcal{U}_{f_i} \cup \mathcal{U}_{f_i}^{Disp}, & \text{if } Disp(\mathcal{C}_{\mathcal{F}'}) \leq Disp(\mathcal{C}_{\mathcal{F}' \setminus \{f_i\}}) \\ \mathcal{U}_{f_i}^{Disp}, & \text{if } Disp(\mathcal{C}_{\mathcal{F}'}) > Disp(\mathcal{C}_{\mathcal{F}' \setminus \{f_i\}}). \end{cases}$$

[Grgic-Hlaca et al 2016]

Process Fairness vs. Outcome Fairness



(a) Feature-apriori fairness (b) Feature-accuracy fairness (c) Feature-disparity fairness

Figure 2: Outcome fairness, measured as disparity in mistreatment, vs. different measures of process fairness for different classifiers. The color intensity of each point represents the accuracy of the corresponding classifier.

Process Fairness also exhibits Performance-Fairness trade-off

Decoupled Classifiers

- **Problem with Classification:** Models that ignores group membership impose heterogenous trade-offs across groups
- **Decoupled Classifiers:** Train a classifier for each group using data from it
- **Preference Guarantees:** Decoupling would recover the most accurate classifier for each group and a set of decoupled classifiers satisfies rationality if each group is assigned a model that is at least as accurate as the pooled classifier

[Ustun et al 2018; Herbert-Johnson et al 2018]

	Group A			Group B			Pooled		
(x_1, x_2)	n^+	n^-	h_A^*	n^+	n^-	h_B^*	n^+	n^-	$\hat{h}_0$
(0, 0)	50	101	-	100	50	+	150	151	-
(0, 1)	101	50	+	50	100	-	151	150	+
(1, 0)	101	50	+	50	100	-	151	150	+
(1, 1)	101	50	+	50	100	-	151	150	+

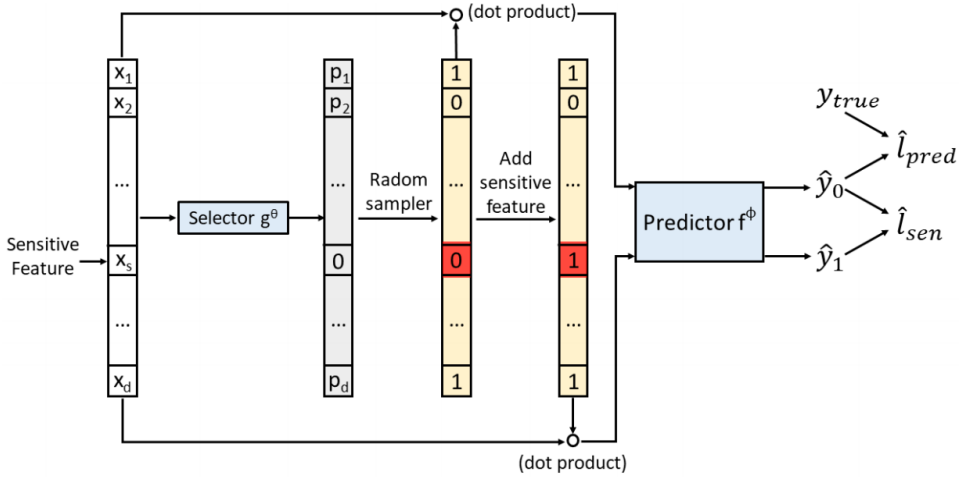
Figure 1. Training a pooled classifier that ignores group membership may impose unavoidable trade-offs between groups. We are given data from two groups $z \in \{A, B\}$ with heterogeneous data distributions $\mathbb{P}(y = +1|\boldsymbol{x}, A) = \mathbb{P}(y = -1|\boldsymbol{x}, B)$. Here, n^+ and n^- denote the number of training examples with $y = +1$ and $y = -1$. Decoupled training produces the best classifier for each group $\hat{h}_A = h_A^*$ and $\hat{h}_B = h_B^*$, both of which have an error rate of 33%. In contrast, pooled training produces a classifier $\hat{h}_0$ with disparate impact due to a *tyranny of the majority*: the data contains slightly more samples from A so that empirical risk minimization outputs the *best* classifier for A which is the *worst* classifier for B. Pooled training with a parity constraint such as equal accuracy between A and B would fix the performance gap, but achieve an error rate of 50% for each group, missing the opportunity to provide better accuracy.

Group A				Group B				Pooled with z			
x_1	n^+	n^-	h_A^*	x_1	n^+	n^-	h_B^*	(x_1, z)	n^+	n^-	h_0^*
0	50	0	-	0	0	50	+	(0,0)	0	50	+
1	0	50	+	1	50	0	-	(1,0)	50	0	-
								(0,1)	50	0	-
								(1,1)	0	50	+

Figure 2. A pooled classifier that encodes group membership may not perform as well as a pair of decoupled classifier when we fit classifiers from a hypothesis class that cannot represent the heterogeneity between groups. Here, we consider training linear classifiers using data from heterogeneous groups $z \in \{A, B\}$. A linear classifier trained separately for each group has zero error. However, there does not exist a linear, pooled classifier with zero error due to the XOR structure.

Fairness & Adversarial Debiasing

- Simultaneously reduce data bias and model bias via adversarial networks
- Sample features and reformulate input with only non-sensitive features
- Minimizing the marginal contribution of the sensitive feature to strengthen model robustness towards the sensitive feature
- Results: Adding sensitive information does not influence prediction results. Improves fairness as well as prediction performance
- Adversarial machine learning widely used in learning fair representations



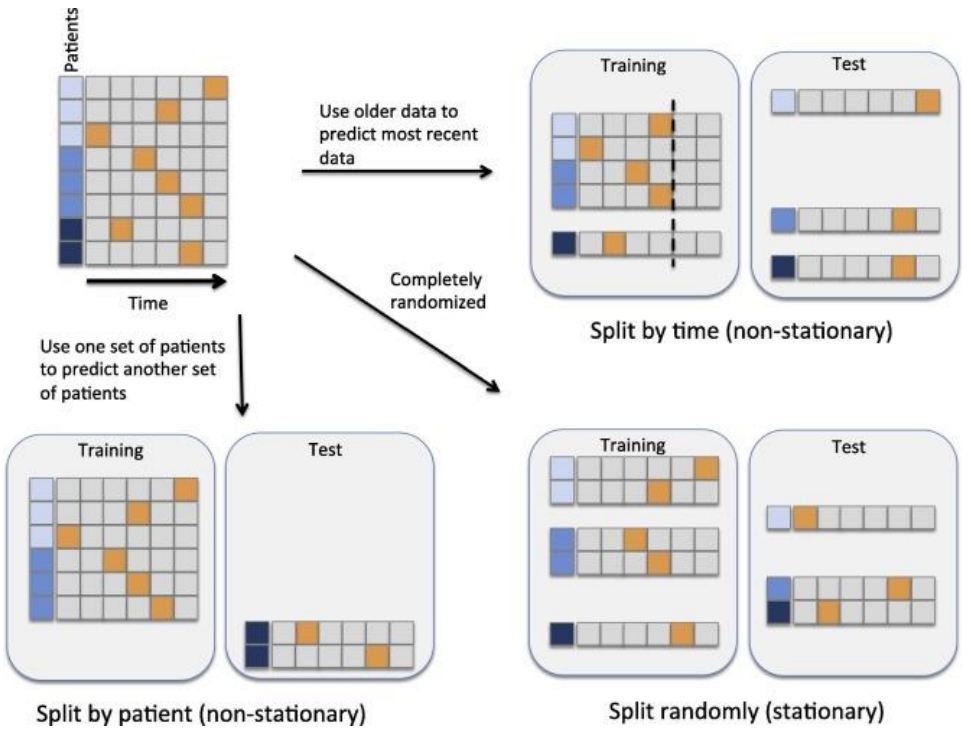
[Beutel et al 2017; Zhang et al 2018; Wang et al 2019]

Operationalizing Fairness in Healthcare ML

Post-Deployment Issues

The Problem of Non-Stationarity

- Real world data is characterized by non-stationarity. [Jung et al 2015]
- Extended versions of ML algorithms can deal with concept drift while being fair
- Use older data to predict most recent data
  - Split by time (non-stationary)
  - Split randomly (stationary)
- Use of one set of patients to predict another set of patients
  - Split by patient (non-stationary)
- If one is interested in “using the model on future data, then it appears that a substantially simpler model is best” [Jung and Shah 2015]



Fairness and Non-Stationarity

- **Fairness Gain:** The fairness gain of an attribute A relative to instances *D* can be defined as the discrimination reduction in D due to splitting on *A* where D_v $v \in dom(A)$ are the partitions induced by *A*

$$FG(D, A) = |Disc(D)| - \sum_{v \in dom(A)} \frac{|D_v|}{|D|} |Disc(D_v)|$$

$$FIG(D, A) = \begin{cases} IG(D, A), & if\ FG(D, A) = 0 \\ IG(D, A) \times FG(D, A), & otherwise \end{cases}$$

[Zhang et al 2020]

Ethics Gone Wrong: Exploration & Exploitation in Medicine

The New York Times

Syphilis Victims in U.S. Study Went Untreated for 40 Years

By JEAN HELLER
The Associated Press

WASHINGTON, July 25—For 40 years the United States Public Health Service has conducted a study in which human beings with syphilis, who were induced to serve as guinea pigs, have gone without medical treatment for the disease and a few have died of its late effects, even though an effective therapy was eventually discovered.

The study was conducted to determine from autopsies what the disease does to the human body.

Officials of the health service who initiated the experiment have long since retired. Current officials, who say they have serious doubts about the morality of the study, also say that it is too late to treat the syphilis in any surviving participants.

Doctors in the service say they are now rendering whatever other medical services they can give to the survivors while the study of the disease's effects continues.

Dr. Merlin K. DuVal, Assistant Secretary of Health, Education and Welfare for Health and Scientific Affairs, expressed shock on learning of the study. He said that he was making an immediate investigation.

The experiment, called the Tuskegee Study, began in 1932 with about 600 black men,

MEMORANDUM

DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
OFFICE OF THE SECRETARY

TO : Director
Center for Disease Control
THROUGH: Administrator, HS

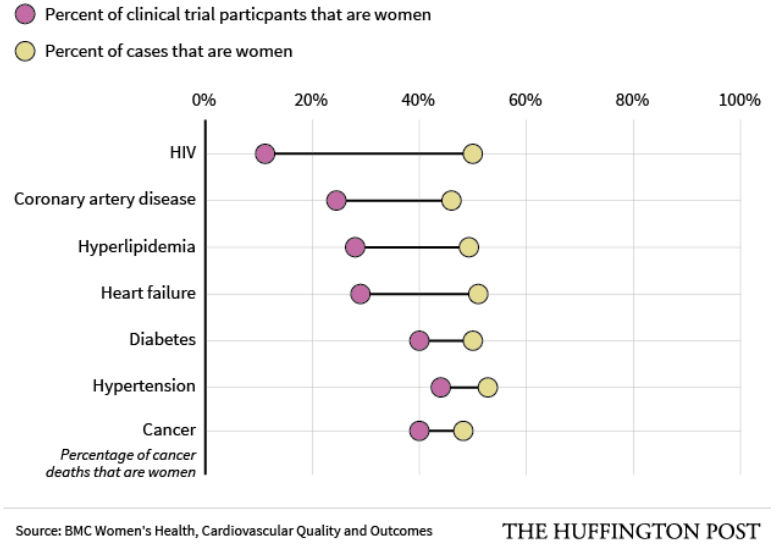
DATE: NOV 16 1972

FROM : Assistant Secretary for Health

SUBJECT: Termination of USPHS Study of Untreated Syphilis (the Tuskegee Study)

As recommended by the Tuskegee Syphilis Study Ad Hoc Advisory Panel, I have decided that the "Tuskegee Study" as a study of untreated syphilis must be terminated. I will advise you of the necessary steps to be taken to assure that appropriate medical care be given to all remaining participants in the "Tuskegee Study" as a part of the close-out phase of the project.

Merlin K. DuVal, M.D.



MEDICAL MALAISE

If you're not a white male, artificial intelligence's use in healthcare could be dangerous

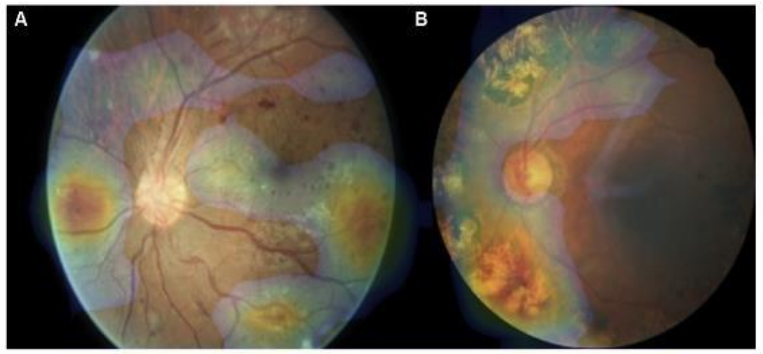
AI Fairness Obfuscating Fair outcomes?

- Due to complexity of real-world healthcare use cases the predictors may not always map onto real world causal relationships
- Fairness in predictions is not the same as guaranteeing fair outcomes with respect to a given health condition
- If a model's performance is assessed w.r.t true events (and not just train-test) most notions of model performance suffer
- "These discrepancies would only be evident longer term and not at the point-of-care where decisions must be made concerning the care management of patients."
- Fairness as measured by output metrics is insufficient. Real-world downstream effects of decision-making must be carefully considered

[McCradden et al 2020]

From High resource to Low resource environments

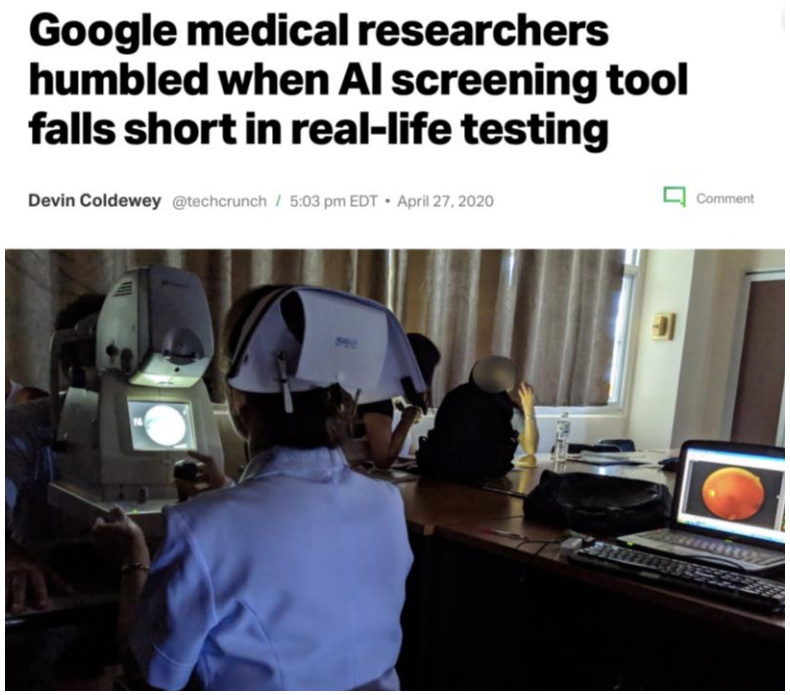
- A practical limitation of deploying machine learning models is the shift in performance observed when moving from high resource to low resource environments
- What are the regulatory obligations and resource support needed to ensure that translation of technology across high-to-low resource contexts happens
- Most aspects of responsible AI like fairness, explainability, and performance usually break down



[Gargeya 2017]

Generalization problems in moving from High resource to low resource environments

- Deployment of Google's Retinopathy tool in low resource environments
- Clinics in Google's study often experienced slower and unreliable connections
- Example: In one clinic the internet went out for two hours during eye screening, reducing the number of patients screened by half (from 200 to 100)
- Fewer people in this case received treatment because of an attempt to leverage this technology



Google medical researchers humbled when AI screening tool falls short in real-life testing

Devin Coldewey @techcrunch / 5:03 pm EDT • April 27, 2020 Comment

Operationalizing Fairness in Healthcare ML

Societal impacts of Fair ML in Healthcare

Long Term Impact of Fair ML

- Under what circumstances does fairness criteria promote the long-term well-being of protected groups over time
- In standard classification setting such scenarios are not considered
- Even in one-step feedback models, common fairness criteria do not promote improvement over time
- In many scenarios in fact cause harm in cases where an unconstrained objective would not
- Most models that aim to predict long term impacts of fairness are brittle w.r.t specific modeling assumptions

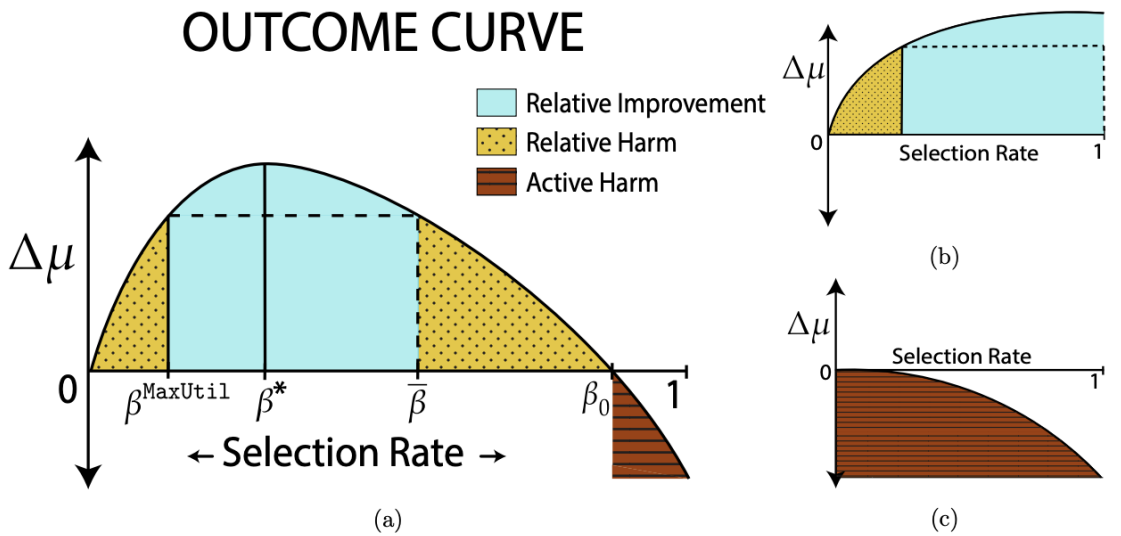
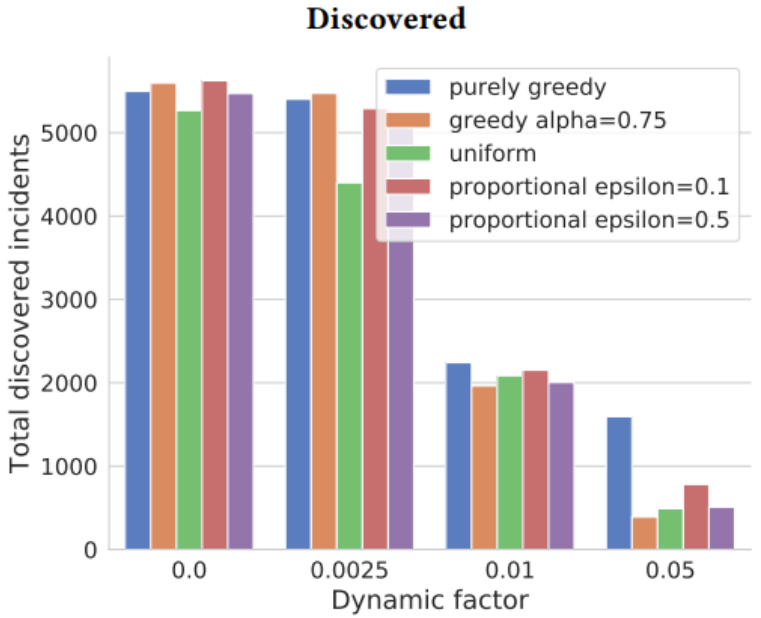


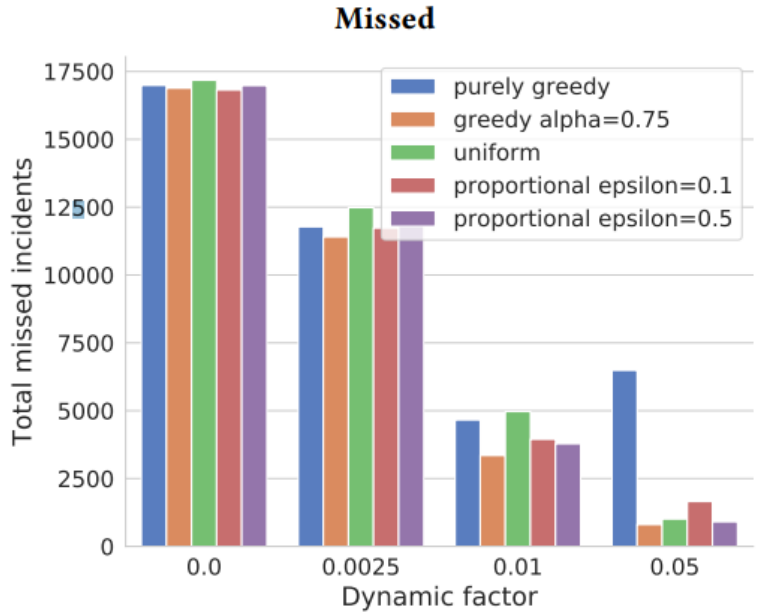
Figure 1: The above figure shows the *outcome curve*. The horizontal axis represents the selection rate for the population; the vertical axis represents the mean change in score. (a) depicts the full spectrum of outcome regimes, and colors indicate regions of active harm, relative harm, and no harm. In (b): a group that has much potential for gain, in (c): a group that has no potential for gain.

[Liu et al 2020]

Long Term aspects of Fair ML

- Long term dynamics in any system are hard to assess
- The use of simulation for studying fairness has been proposed as a way to address this structurally
- Work by D'Amour et al showed that "the long-term results offered by simulation supports qualitatively different (though not incompatible) fairness conclusions from those obtained before"
- Similar studies for healthcare-based scenarios are lacking

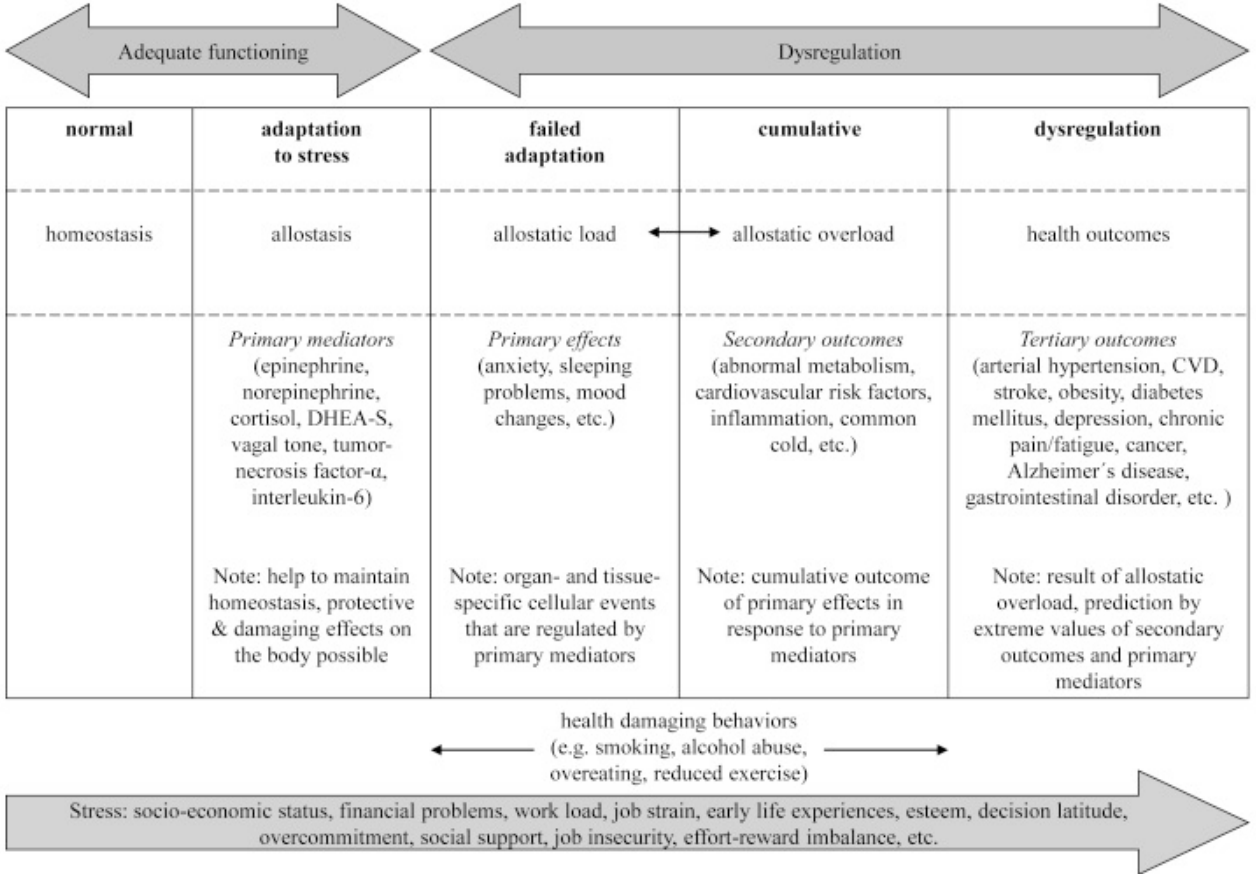




[D'Amour et al. 2020]

A Lifecycle view of Inequity

- Long term consequences of (un)fair decisions can accumulate over time
- What is the cumulative effect of discrimination faced by an individual over the course of a lifetime?
- What are the effects of discrimination at multiple points of care over the course of decades?
- What are the physiologic effects of chronic stressors related to inequity?



[Mauss et al. 2015]

Data: Minority Class, Fairness and Privacy

- **Problem:** For better predictive performance, it is critical to gather data that include minorities and to ensure that these data are not completely subsumed by data from presumed “normative” populations
- **Conundrum:** Data collection in this context would also lead to issues related to confidentiality and privacy e.g., potentially dangerous for subjects
- **Multi-dimensional Problem:** Minority status has multiple dimensions, varies in intensity and impact, and varies changes over time. The simple protected vs. other class framework may not suffice

Data: Minority Classes & Generalization

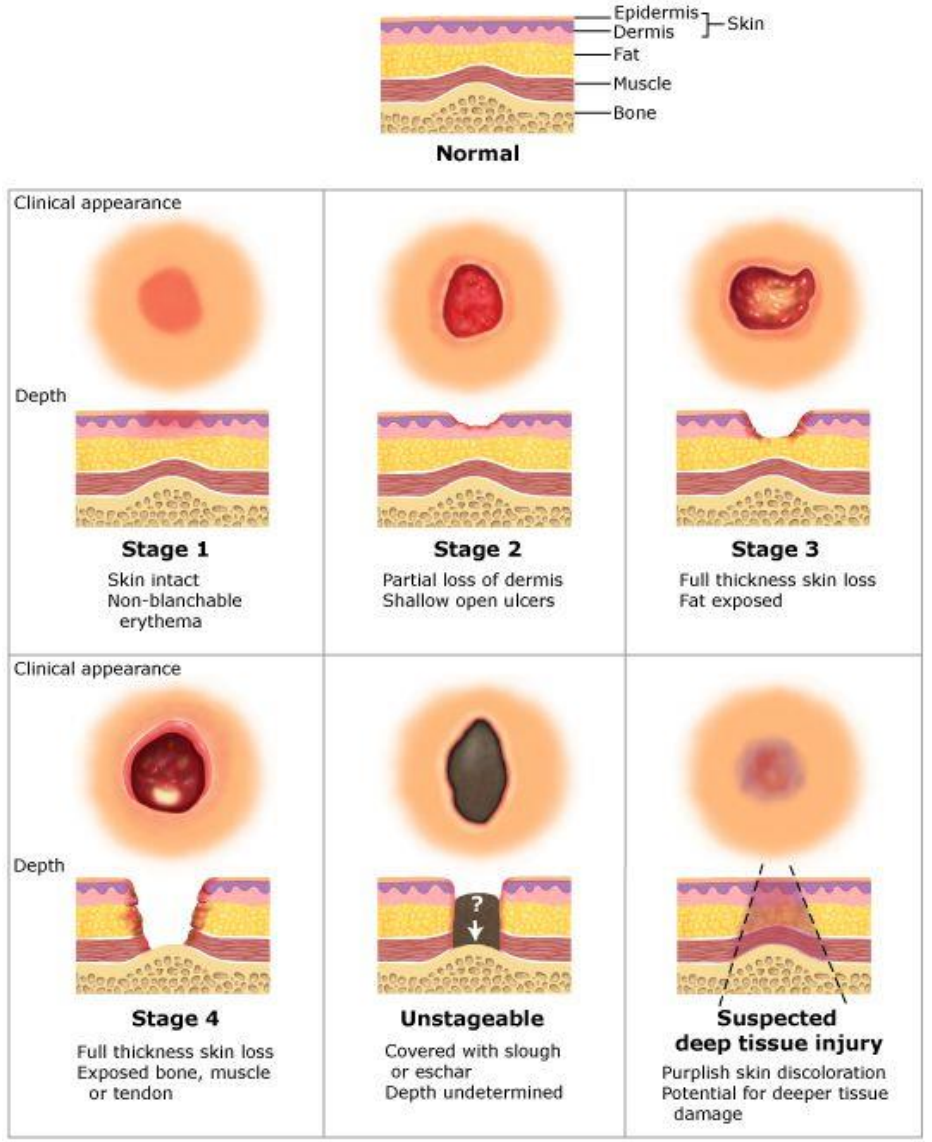
- In sub-Saharan Africa, women are diagnosed with breast cancer younger, on average, than are their peers in developed countries, and their disease is more advanced at diagnosis. Diagnostic AI tools trained on mammograms from European women are primed to identify disease in its early stages in older women [Nordling 2019]
- Data security and access concerns have been raised about allowing developers to access such data from low-income countries

Data: Trouble with Labels

- Data labels are generally derived from the Electronic Health Records (EHRs)
- Labeling incentives may change as healthcare processes change (e.g., Meaningful Use criteria, HEDIS measures)
- Definitions of an outcome may change over time e.g., thresholds that define a disease
- Mental health evaluations, psychological assessments, pain assessment by clinicians vs. patients, patient reported outcomes may vary
- Racial and sex/gender biased disparities have been observed for pain assessment across multiple studies
- Optimizing for the wrong label can lead to biased outcomes

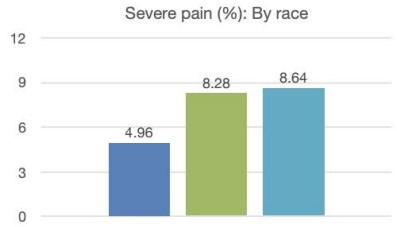
Data: Trouble with Labels: Pressure Injury

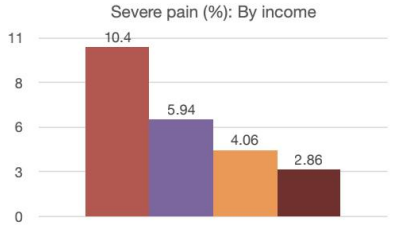
- “Localized damage to the skin and underlying soft tissue, usually over a bony prominence or related to a medical or other device.” NPUAP
- Multiple risk assessment exist for pressure injury (PI). The **Braden Scale** is the most widely used scale
- All measurement scales for PI are highly subjective in nature

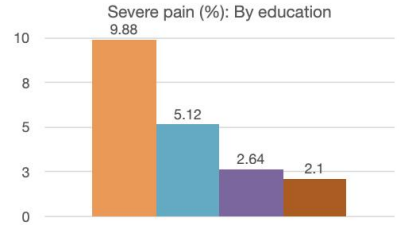


Data: Trouble with Labels: Knee Pain

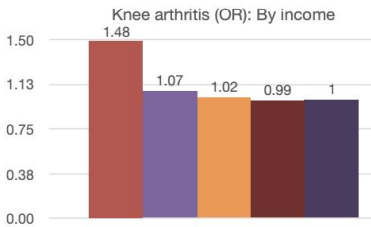
- Large pain gradients
  - Race
  - Income
  - Education

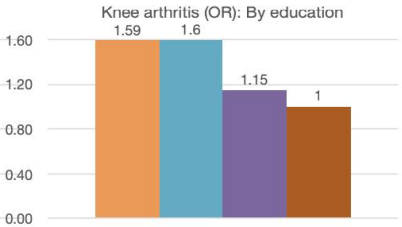






- Higher prevalence of painful conditions
  - By income
  - By education

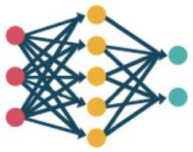




What if instead of learning from the radiologist…

We trained the algorithm to listen to the patient?

Simulation: Who would get surgery…
if the algorithm were in charge, not the doctor?



- Identify patients with severe pain and
  - High disease severity according to human
  - High disease severity according to algorithm

More - Black knees eligible for surgery

Less - Black knees, severe pain but ineligible for surgery

*Severe pain + no surgery + high algorithm score = most likely to be on oral pain medicine **incl. opiates***

Slide courtesy - Ziad Obermeyer from https://blogs.worldbank.org/impactevaluations/machine-learning-pain-relief

Grol-Prokopczyk, *Pain* 2017, Baldassari et al., *Osteoarthritis and Cartilage* 2014
Pierson, Emma, et al. "Using machine learning to understand racial and socioeconomic differences in knee pain" Under Review at JAMA 2019.

Normative Moral Principles and ML Fairness

- Since it is not possible to satisfy every fairness metric, moral and political philosophy are required for design choices **Consequentialist Approaches** focuses on each potential distribution and its effects
- **Deontological Approaches** evaluate distributions based on rights
- An Egalitarian approach confers equal rights, and thus equal shares, to every member of the population

Utilitarian Principle (Maxisum):

$$\mathrm{UT}(d_i) = \arg\max_{x_i} \sum_{i=1}^{n} w_i u(d_i, v_i)$$

Prioritarian Principle (Maximin):

$$\mathrm{PR}(d_i) = \arg\max_{x_i} \min_{x_i} w_i u(d_i, v_i)$$

Egalitarian Principle:

$$\mathrm{EG}(d_i) = \arg\min \sum_{i=1}^{n} \ell(d_i, \tfrac{1}{n} S)$$

Libertarian Principle:

$$\mathrm{LB}(d_i) = \arg\min \sum_{i=1}^{n} \ell(d_i, \tfrac{R_x}{\sum R_i} S)$$

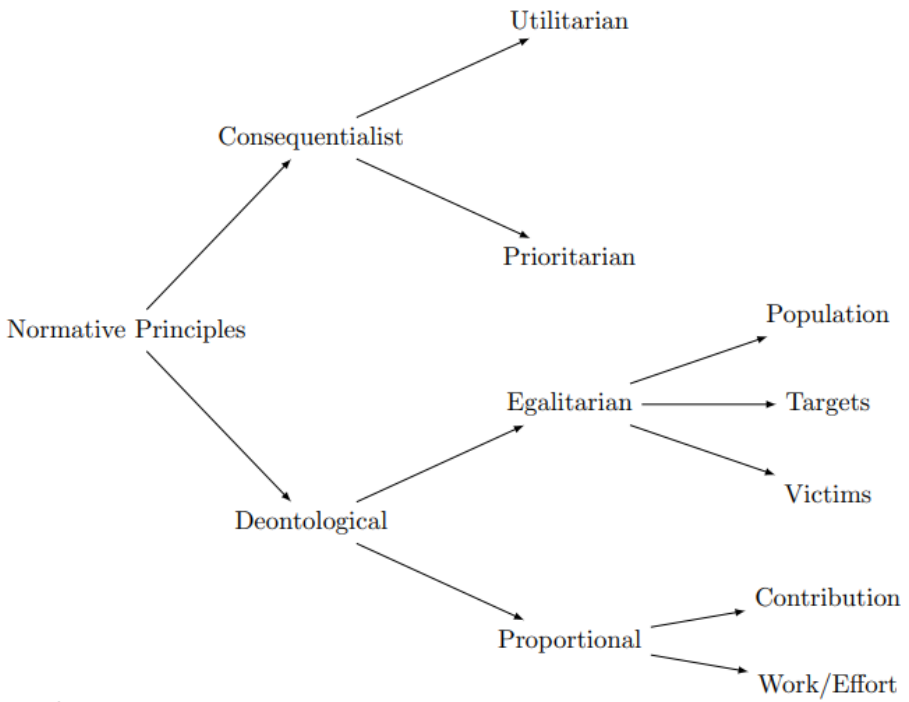
Desert Principle:

$$\mathrm{DS}(d_i) = \arg\min \sum_{i=1}^{n} \ell(d_i, \tfrac{W_x}{\sum W_i} S)$$

[Binns 2018; Leben 2020]

Deontic Justice & Fair ML in Healthcare

- **Deontic Justice:** It is not just the state of affairs of unfairness that matters but also what were the conditions that led to that state [Binns 2018]
- This requires integrating a perspective from philosophy, history, economics, sociology etc.
- Once identified, where should the locus of responsibility be? focus on improving outcomes
- When is a particular mistreatment of a protected group worse than the mistreatment of the protected group?
- Modeling strongly coupled complex systems is hard!



Luck Egalitarianism & Fair ML in Healthcare

- What type of inequalities are acceptable?
- **Luck Egalitarianism:** Allow inequality in cases which result from people's efforts and risk taking and do not allow it in cases where it is because of brute luck (skin color, born with debilitating health condition) [Arneson 1989]
- **Coupled nature of social units:** People choices may be limited because they choose to take of sick, elderly, young family members
- Free choice are not always free; need to audit systems to determine how stakeholders are being affected?

Race Correction and Fair ML IN Healthcare

- Race corrections i.e., different risk scoring adjustments based on race, are made in many healthcare models which also seep into ML models
- Example: Spirometer is widely used across the world for the diagnosis and management of many respiratory diseases
- The notion that black and white lungs differ goes back to Thomas Jefferson who posited differences between slaves and freeman's lung functions [Braun 2015]
- Spirometer based studies were used to justify differences between Europeans, Asians and Africans by Plantation owners and colonial administrators [Braun 2014]
- Such race corrections are now part of computer models (ML or otherwise)


Viewpoint

July 29, 2020

Black Kidney Function Matters

Use or Misuse of Race?

Neil R. Powe, MD, MPH, MBA[1]

Library Demo

FairMLHealth Library

- Vision
  - An extensible Python library dedicated to fairness in machine learning specifically tailored for healthcare with domain knowledge integration
- Future Goals & Milestones
  - Measurement of Fairness in Healthcare Applications
  - Comparison of classifiers for Fairness and Performance Trade-offs
  - Arbitrary comparison of protected classes and intersectional classes
- Current Release: FairMLHealth 0.1: Alpha Release
  - Demonstration of measurement and comparison of fairness metrics for a publicly available dataset (MIMIC-3)

FairMLHealth Library

Measuring Fairness in Healthcare ML for Scikit-Compatible Models

Overview

This tutorial introduces methods and libraries for measuring fairness and bias in machine learning models as as they relate to problems in healthcare. Through the tutorial you will first learn some basic background about fairness and bias in machine learning. You will then generate a simple baseline model predicting Length of Stay (LOS) using data from the MIMIC-III database, which you will use as an example to understand the most prominent fairness measures. You will also gain familiarity with the Scikit-Learn-compatible tools available in AIF360 and FairLearn, two of the most comprehensive and flexible Python libraries for measuring and addressing bias in machine learning models.

Tutorial Contents

Part 0 - Metrics of Fairness

Part 1 - Model Setup

Part 2 - Metrics of Fairness in AIF360

Part 3 - Comparing Against a Second Model: Evaluating Unawarenes

Part 4 - Testing Other Sensitive Attributes

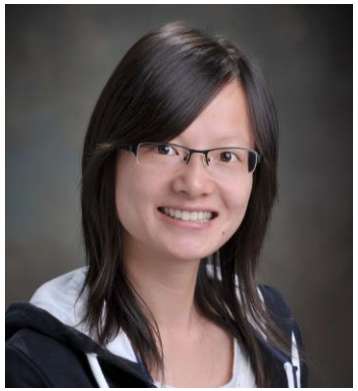
Part 5 - Comparison to FairLearn

Tutorial Requirements

This tutorial assumes basic knowledge of machine learning implementation in Python. Before starting, please install AIF360 and FairLearn. Also, ensure that you have installed the Scipy, Pandas, Numpy, Scikit, and XGBOOST libraries.

The tutorial also uses data from the MIMIC III Critical Care database, a freely accessible source of Electronic Health Records from Beth Israel Deaconess Medical Center in Boston. To download the MIMIC III data, please use this link: Access to MIMIC III. Please save the data with the default directory name ("MIMIC"). No further action is required beyond remembering the download location: you do not need to unzip any files.

KenSci



Best Practices

Stakeholders in the Machine Learning Cycle

Model Design	• Review the goal of the ML model with diverse and representative stakeholders • Evaluate current initiatives which may be interacting with proposed model / workflow and consider what downstream consequences may be • Discuss ethical concerns about model use and / or misuse
Data Collection	• Determine which features and which patient groups should be considered protected or surrogates of protected • Evaluate relationship between surrogate and outcome • Assess if the protected groups are adequately represented.
Training	• Apply pre-processing, in-processing and post-processing techniques to make the model more fair
Evaluation	• Engage with stakeholders, patient groups representatives, data scientists, machine learning experts to determine what are appropriate evaluation metrics given the use case • Determine what processes will be affected by different model outputs (allocation)
Deployment and Review	• Monitor the results and periodically check with the stakeholders regarding how the model is affecting them

AI Audits in Responsible ML

- Model risk management involves periodic look backs to evaluate the performance and the consequences of ML models in deployment
- Internal processes vs external processes
- Ad-hoc vs. Domain specific standard auditing process
- Scalability and comprehensiveness
- Challenges with race, ethnicity data that is either commonly missing or too coarse for meaningful evaluation

Impossibility of Fairness in the real world

- Unfair practices do not exist in a vacuum but are embedded in the larger context of historical, social and political realities [Glymour et al 2019; Herington 2020]
- Measures of algorithmic bias assume that an algorithm which is fair in the abstract will be fair in the world
- Centuries of injustice continue to permeate society and continue to be responsible for race- and gender-based inequality
- Implicit vs explicit biases can be difficult and / or impossible to adjust for and demand societal changes

Prediction and Policy

- Allocation of services, particularly those derived from outputs of machine learning models, must be continually evaluated for evidence of bias to ensure that services are delivered equally across protected groups
- The allocation of services will be determined by how clinicians or other end users interact with the model
  - Is there a disparate impact?
  - Is the clinical team subject to automation bias or dismissal bias? And how may that differentially affect patient groups?
  - Opportunity cost

Ethics Washing in Healthcare AI

Reductionism

"Doing the morally right thing is essentially the same as acting in a *fair* way. *(or: transparent, or egalitarian, or <substitute any other value>)*. So ethics is the *same* as fairness *(or transparency, or equality, etc.)*. If we're being fair, then we're being ethical."

Simplicity

"In order to make ethics practical and action-guiding, we need to distill our moral framework into a user-friendly compliance checklist. After we've decided on a particular path of action, we'll go through that checklist to make sure that we're being ethical."

Ethics Washing in Healthcare AI

Relativism

"We all disagree about what is morally valuable, so it's pointless to imagine that there is a *universal* baseline against which we can use in order to evaluate moral choices."

Dichotomy

"The goal of ethical reasoning is to 'be(come) ethical'.

Value Alignment

"If relativism is wrong there must be *one* morally right answer. We need to find that right answer, and ensure that everyone in our organization acts in alignment with that answer."

Ethics Washing in Healthcare AI

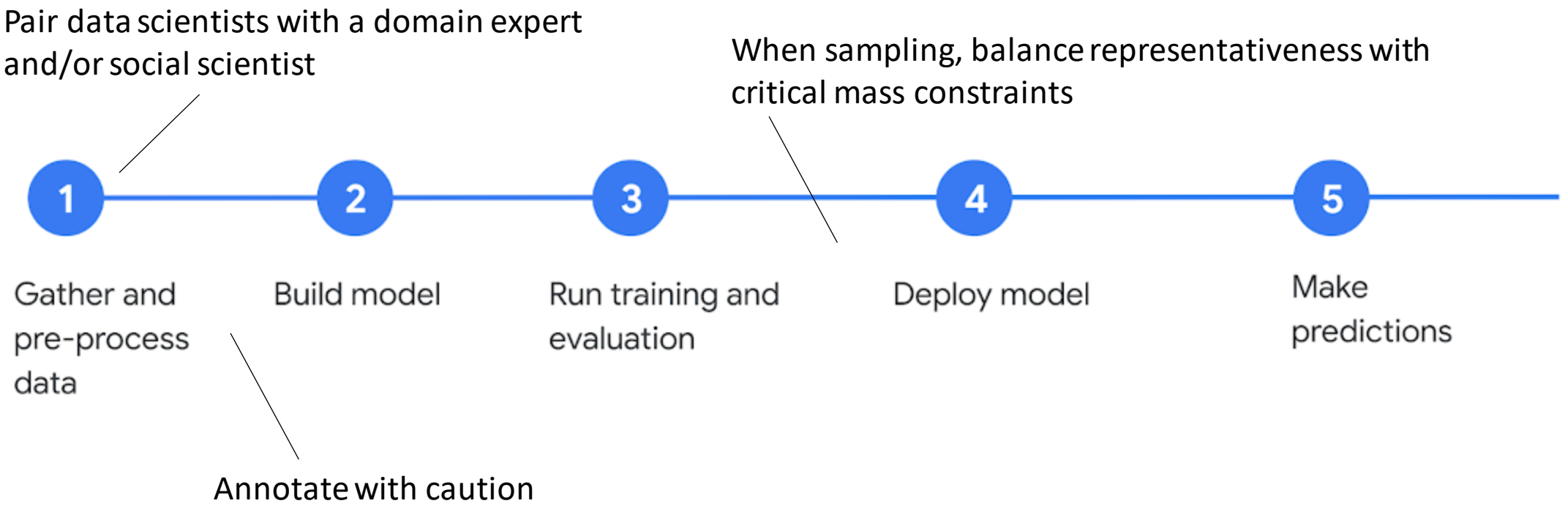
The myopia Trap

"The ethical trade-offs that we identify within one context are going to be the same ethical trade-offs that we are going to face in other contexts and moments in time, both with respect to the nature of the trade-off and with respect to the scope of the trade-off."

The rule of law Trap

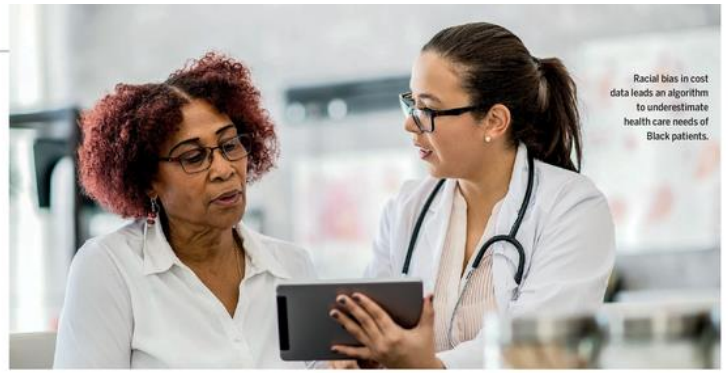
"Ethics is essentially the same as the rule of law. When we lack appropriate legal categories for the governance of AI, ethics is a good substitute. And when we do have sufficient legal frameworks, we don't need to think about ethics."

Fairness and Inclusion as design principles



Target Labels, Modeling Building and Intervention

- **Inference A:** i) Black patients cost less ii) Black patients' poor care is the result of patients' "non-compliance" & "lack of trust"
- **Inference B:** "Black patients are valued less, structural and interpersonal racism are persistent in the healthcare system" which are responsible for these outcomes
- Predictions will have the same semantic interpretation "if companies, institutions, and individuals provided the same level of care for Black patients" [Benjamin 2019]



SOCIAL SCIENCE

Assessing risk, automating racism

A health care algorithm reflects underlying racial bias in society

By Ruha Benjamin

As more organizations and industries adopt digital tools to identify risk and allocate resources, the automation of racial discrimination is a growing concern. Social scientists have been at the forefront of studying the historical, political, economic, and ethical dimensions of such tools (*1*–*3*). But most analysts do not have access to widely used proprietary algorithms and so cannot typically identify the precise mechanisms that produce disparate outcomes. On page 447 of this issue, Obermeyer *et al.* (*4*) report one of the first studies to examine the outputs and inputs of an algorithm that predicts health risk, and influences treatment, of millions of people. They found that because the tool was designed to predict the cost of care as a proxy for health needs, Black patients with the same risk score as White patients tend to be much sicker, because providers spend much less on their care overall. This study contributes greatly to a more socially conscious approach to technology development, demonstrating how a seemingly benign choice of label (that is, health cost) initiates a process with potentially life-threatening results. Whereas in a previous era, the intention to deepen racial inequities was more explicit, today coded inequity is perpetuated precisely because those who design and adopt such tools are not thinking carefully about systemic racism.

Obermeyer *et al.* gained access to the training data, algorithm, and contextual data for one of the largest commercial tools used by health insurers to assess the health profiles for millions of patients. The purpose of the tool is to identify a subset of patients who require additional attention for complex health needs before the situation becomes too dire and costly. Given increased pressure by the Affordable Care Act to minimize spending, most hospital systems now utilize predictive tools to decide how to invest resources. In addition to identifying the precise mechanism that produces biased predictions, Obermeyer *et al.* were able to quantify the racial disparity and create alternative algorithmic predictors.

Practically speaking, their finding means that if two people have the same risk score that indicates they do not need to be enrolled in a "high-risk management program," the health of the Black patient is likely much worse than that of their White counterpart. According to Obermeyer *et al.*, if the predictive tool were recalibrated to actual needs on the basis of the number and severity of active chronic illnesses, then twice as many Black patients would be identified for intervention. Notably, the researchers went well beyond the algorithm developers by constructing a more fine-grained measure of health outcomes, by extracting and cleaning data from electronic health records to determine the severity, not just the number, of conditions. Crucially, they found that so long as the tool remains effective at predicting costs, the outputs will continue to be racially biased by design, even as they may not explicitly attempt to take race into account. For this reason, Obermeyer *et al.* engage the literature on "problem formulation," which illustrates that depending on how one defines the problem to be solved—whether to lower health care costs or to increase access to care—the outcomes will vary considerably.

To grasp the broader implications of the study, consider this hypothetical: The year is 1951 and an African American mother of five, Henrietta Lacks, goes to Johns Hopkins Hospital with pain, bleeding, and a knot in her stomach. After Lacks is tested and treated with radium tubes, she is "digitally triaged" (*2*) using a new state-of-the-art risk assessment tool that suggests to hospital staff the next course of action. Because the tool assesses risk using the predicted cost of care, and because far less has commonly been spent on Black patients despite their actual needs, the automated system underestimates the level of attention Lacks needs. On the basis of the results, she is discharged, her health rapidly deteriorates,

Department of African American Studies, Princeton University, Princeton, NJ, USA. Email: ruha@princeton.edu

PHOTO: PATCHARA/GETTY IMAGES

Published by AAAS

Human Centered Design

- Engage a diverse population of potential users
- Employ a variety of different use-case scenarios
- Disclose data collection
- Favor user control over automation
- Prepare for potentially adverse (problematic) feedback

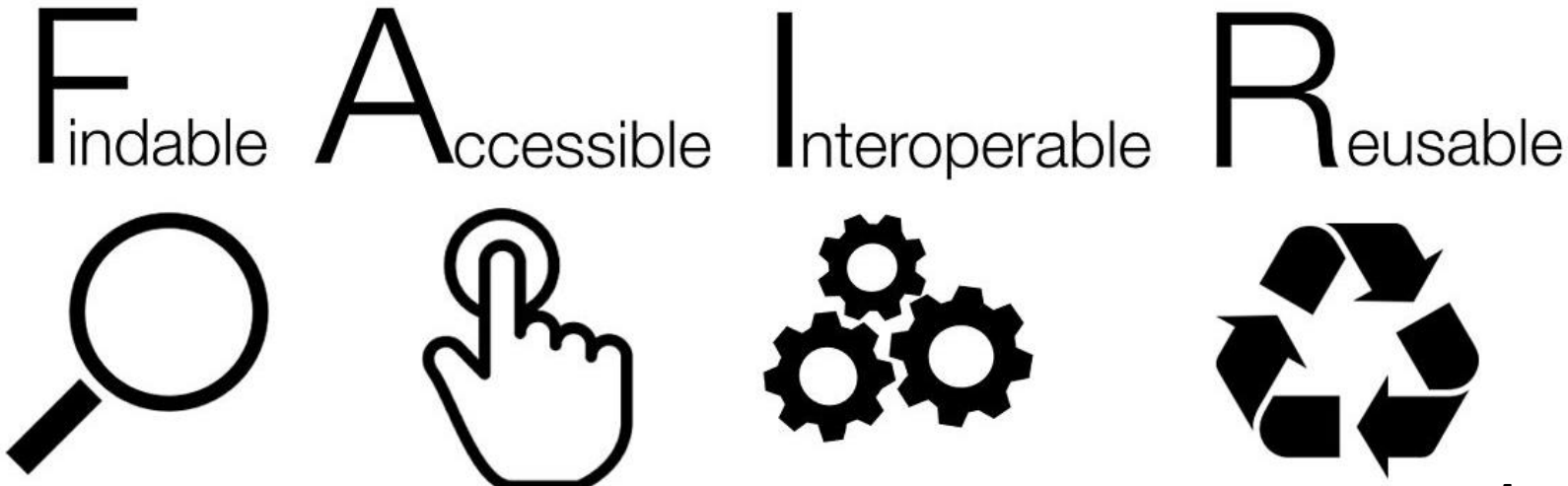
[Awad et al 2020]

Differences do not always entail inequality

- It some cases one can incorporate differences between group into the model e.g., “biological differences between genders can affect the efficacy of pharmacological compounds” [McCradden et al 2020]
- Biological difference may exist between groups but establishing those require rigorous studies since past science on racial difference is tainted by blatant racism
- “In many cases it is difficult to distinguish between acknowledging difference and propagating discrimination” [McCradden et al 2020; Powe 2020]

FAIR Principles in Fairness

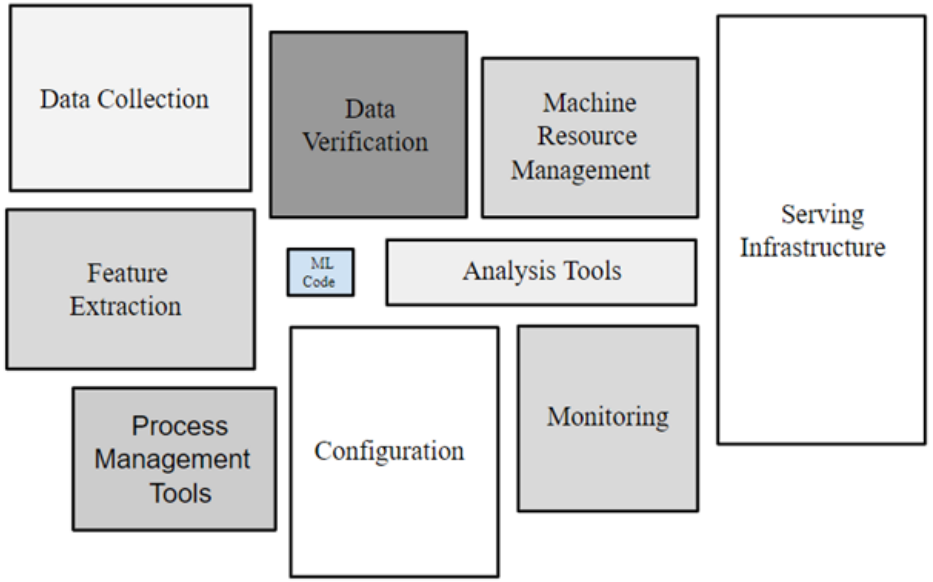
- **F**indability, **A**ccessibility, **I**nteroperability, and **R**euse (of data)
- Ensuring fairness often requires data provenance, accessibility and auditability of data
- Machine-actionability i.e., the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention



[Wilkinson et al 2016]

Fairness in ML Deployment

- Continually monitor the ML pipeline
  - Does the training data, test data and the data in deployment match
  - User reports with fairness metrics
  - Evolution of values of fairness metrics over time
  - Triggers when value go below given thresholds
- Having diverse stakeholders to audit system for fairness
- Tools and for determining data/label drift
- Periodically audit system for bias
- Database of user complaints



[Sculley et al 2015]

[Cramer et al 2019]

Comprehension, Cognition and AI Fairness

- Since fairness metrics and fair models are likely to be used by non-experts it is important to gauge an average person's understanding of fairness
- Studies have found that "comprehension is lower for equal opportunity, false negative rate than other definitions"
- Education is a strong predictor of comprehension of fairness metrics (Problem: Marginalized are more likely to be affected)
- Those with the "weakest comprehension of fairness metrics also express the least negative sentiment toward them"

[Saha et al 2020]

How should algorithms be regulated?

- **Side Effects:** Chemotherapy drugs can shrink a patient's tumor but they can also have devastating side effects
- **Different effect on different populations:** The blood thinner clopidogrel does not work in the 75% of Pacific Islanders as their bodies do not produce the enzyme for drug activation
- **Understanding who the product is for:** Clinical researchers are required to "clearly define a drug's target users so a prescribing clinician can have confidence that the drug has been successfully tested on similar patients."
- **Understanding how the product was developed:** Clinical trials have mandatory reporting requirements to ensure transparency and to hold product developers accountable

[Coravos et al 2019]

Regulation: An FDA for ML and Algorithms?

- In drug development FDA enforces protocols for manufacturers to prove the safety and effectiveness of drug products before they go on the market
- Algorithms and ML models do no come with warning labels
- Adverse Events: Drugs carry the risk of adverse events (injury, hospitalization etc.) A well-documented public reporting structure for handling such mishaps exists (as given by FDA) where reports of serious events like death, serious injury can be tracked
- Ethical and quality control standards in healthcare: Good Clinical Practice (GCP) for clinical trials, Good Manufacturing Practice (GMP) for products, and Good Laboratory Practice (GLP) for research laboratories

[Coravos et al 2019]

Challenges & Open Questions

- Which measures of fairness are most appropriate in a given context?
- Which variables are legitimate grounds for differential treatment, and why?
- When is disparity between groups acceptable and why?
- Should fairness consist of maximizing equal probability of obtaining some benefit, or minimizing the harms to the least advantaged?
- In making such tradeoffs, should the decision-maker consider only the harms and benefits imposed within the decision-making context, or also those faced by decision-subjects in other contexts?

[Binns 2018]

Challenges & Open Questions

- Many aspects of fairness not captured by metrics or data; how do we address those?
- What relevance should past, future or inter-generational injustices have?
- How to deal with Fairness Gerrymandering where there is insufficient data for modeling?

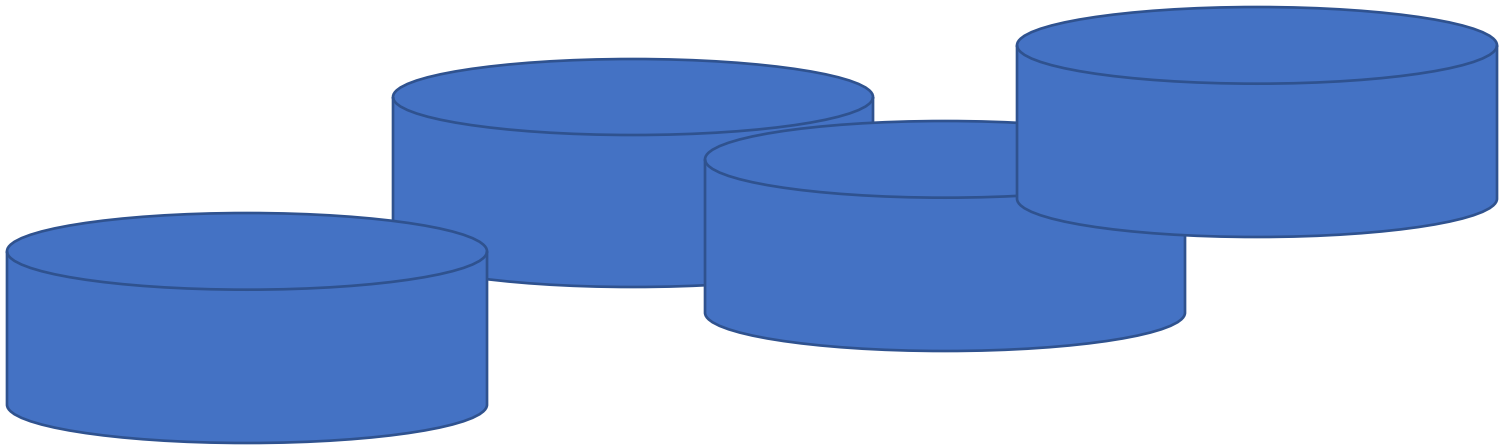
[Binns 2018]

Recap & Conclusion

Recap

- Foundations: Fairness in Healthcare ML
- Measurement & Mismeasurement of Fairness
- Operationalizing Fairness in Healthcare ML
- Best Practices
- Library Demo
- Conclusion

Call to Action: Datasets



- Deployment of Enterprise grade AI and ML models in healthcare at multiple locations in the US and internationally
- Fairness across multiple locations, settings and cohort

Call to Action: Partner with us!
It takes a village

- Deployment of Enterprise grade AI and ML models in healthcare at multiple locations in the US and internationally
- Fairness across multiple locations, settings and cohort

Call to Action: Resources: Websites

- AI Now Institute
- Algorithmic Justice League
- Berkman Klein Center for Internet and Society
- ML Healthcare Resources
- Partnership on AI

Call to Action: Resources: Libraries

Library	Creator	Metrics	Algorithms	Simulations
AIF 360	IBM	✓	✓	
Fairlearn	Hannah Walllach et al	✓	✓	
Fairness Comparison	Sorelle Friedler	✓		
Fairness Indicators	Tensorflow	✓		
ML Fairness Gym	Google			✓
Themis-ML	Niels Bantilan	✓	✓	

And now the **FairMLHealth** Library

References

- Byrd, W. Michael, and Linda A. Clayton. "Race, medicine, and health care in the United States: a historical survey." Journal of the National Medical Association 93, no. 3 Suppl (2001): 11S.
- Wailoo, Keith. "Sickle cell disease—a history of progress and peril." N Engl J Med 376, no. 9 (2017): 805-807.
- de Malave, Florita Z. Louis. Sterilization of Puerto Rican women: a selected, partially annotated bibliography. University of Wisconsin System, Women's Studies Librarian, 1999.
- Randall, Vernellia R. "Slavery, Segregation and Racism: Trusting the Health Care System Ain't Always Easy--An African American Perspective on Bioethics." . Louis U. Pub. L. Rev. 15 (1995): 191.
- King, M. L. Jr. (1966). National Convention for Medical Committee for Human Rights. Washington, DC. Excerpt from speech retrieved from http://www.goodreads.com/quotes/106932-of-all-the-forms-ofinequality-injustice-in-health-care
- Parikh, Ravi B., Stephanie Teeple, and Amol S. Navathe. "Addressing bias in artificial intelligence in health care." Jama 322, no. 24 (2019): 2377-2378.
- Parasuraman, Raja, and Dietrich H. Manzey. "Complacency and bias in human use of automation: An attentional integration." Human factors 52, no. 3 (2010): 381-410.
- Agniel, Denis, Isaac S. Kohane, and Griffin M. Weber. "Biases in electronic health record data due to processes within the healthcare system: retrospective observational study." Bmj 361 (2018).
- McLintock K, Russell AM, Alderson SL, et al. The effects of financial incentives for case finding for depression in patients with diabetes and coronary heart disease: interrupted time series analysis. BMJ Open2014;4:e005178. doi:10.1136/bmjopen-2014-005178. pmid:25142262
- Weizenbaum, Joseph (1976). Computer power and human reason : from judgment to calculation. San Francisco: W.H. Freeman
- Adelman, Larry. "Unnatural causes: Is inequality making us sick?." Preventing Chronic Disease 4, no. 4 (2007).
- Simkin RJ. Women's health: time for a redefinition. *CMAJ*. 1995;152(4):477-479.

References

- Almond, Amanda Lee. "Measuring racial microaggression in medical practice." Ethnicity & health 24, no. 6 (2019): 589-606.Barocas, Solon and Hardt, Moritz., Fairness in machine learning, NeurIPS Tutorial, 2017.
- Dovidio, John F., Susan Eggly, Terrance L. Albrecht, Nao Hagiwara, and Louis A. Penner. "Racial biases in medicine and healthcare disparities." TPM: Testing, Psychometrics, Methodology in Applied Psychology 23, no. 4 (2016).
- Agniel, Denis, Isaac S. Kohane, and Griffin M. Weber. "Biases in electronic health record data due to processes within the healthcare system: retrospective observational study." Bmj 361 (2018).
- Parikh, Ravi B., Stephanie Teeple, and Amol S. Navathe. "Addressing bias in artificial intelligence in health care." Jama 322, no. 24 (2019): 2377-2378.
- Binns, Reuben. "Fairness in machine learning: Lessons from political philosophy." arXiv preprint arXiv:1712.03586 (2017).
- Bierman, Arlene S. "Sex matters: gender disparities in quality and outcomes of care." Cmaj 177, no. 12 (2007): 1520-1521.
- Leben, Derek. "Normative Principles for Evaluating Fairness in Machine Learning." In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 86-92. 2020.
- Chen, Esther H., Frances S. Shofer, Anthony J. Dean, Judd E. Hollander, William G. Baxt, Jennifer L. Robey, Keara L. Sease, and Angela M. Mills. "Gender disparity in analgesic treatment of emergency department patients with acute abdominal pain." Academic Emergency Medicine 15, no. 5 (2008): 414-418.
- Chouldechova, Alexandra, and Aaron Roth. "The frontiers of fairness in machine learning." arXiv preprint arXiv:1810.08810 (2018).
- Parasuraman, Raja, and Dietrich H. Manzey. "Complacency and bias in human use of automation: An attentional integration." Human factors 52, no. 3 (2010): 381-410.

References

- McLintock K, Russell AM, Alderson SL, et al. The effects of financial incentives for case finding for depression in patients with diabetes and coronary heart disease: interrupted time series analysis. BMJ Open2014;4:e005178. doi:10.1136/bmjopen-2014-005178. pmid:25142262
- Weizenbaum, Joseph (1976). Computer power and human reason : from judgment to calculation. San Francisco: W.H. Freeman
- Adelman, Larry. "Unnatural causes: Is inequality making us sick?." Preventing Chronic Disease 4, no. 4 (2007).
- Almond, Amanda Lee. "Measuring racial microaggression in medical practice." Ethnicity & health 24, no. 6 (2019): 589-606.
- Barocas, Solon and Hardt, Moritz., Fairness in machine learning, NeurIPS Tutorial, 2017.
- Wailoo, Keith. "Sickle cell disease—a history of progress and peril." N Engl J Med 376, no. 9 (2017): 805-807.
- de Malave, Florita Z. Louis. Sterilization of Puerto Rican women: a selected, partially annotated bibliography. University of Wisconsin System, Women's Studies Librarian, 1999.
- Randall, Vernellia R. "Slavery, Segregation and Racism: Trusting the Health Care System Ain't Always Easy--An African American Perspective on Bioethics." . Louis U. Pub. L. Rev. 15 (1995): 191.
- King, M. L. Jr. (1966). National Convention for Medical Committee for Human Rights. Washington, DC. Excerpt from speech retrieved from Good reads
- Begley, Tom, Tobias Schwedes, Christopher Frye, and Ilya Feige. "Explainability for fair machine learning." arXiv preprint arXiv:2010.07389 (2020).
- Wilson, Kalpana. "In the name of reproductive rights: race, neoliberalism and the embodied violence of population policies." New Formations 91, no. 91 (2017): 50-68.wi

References

- Corbett-Davies, Sam., Goel, Sharad., Defining and Designing Fair Algorithms, Tutorials at EC 2018 and ICML 2018.
- Corbett-Davies, Sam, and Sharad Goel. "The measure and mismeasure of fairness: A critical review of fair machine learning." arXiv preprint arXiv:1808.00023 (2018).
- Cramer, Henriette., Holstein, Kenneth., Jennifer Wortman Vaughan et. al., Translation Tutorial: Challenges of incorporating algorithmic fairness into industry practice, FAT* Tutorial, 2019.
- Crawford, Kate The Trouble with Bias, NeurIPS Keynote, 2017.
- Dawes, Robyn M., David Faust, and Paul E. Meehl. "Clinical versus actuarial judgment." *Science* 243, no. 4899 (1989): 1668-1674.
- Dresser, Rebecca. "Wanted single, white male for medical research." The Hastings Center Report 22, no. 1 (1992): 24-29.
- Friedler, Sorelle A., Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. "A comparative study of fairness-enhancing interventions in machine learning." In Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 329-338. 2019.
- Friedler, Sorelle A., Carlos Scheidegger, and Suresh Venkatasubramanian. "On the (im) possibility of fairness." arXiv preprint arXiv:1609.07236 (2016).
- Garcia, M (2017): "Racist in the Machine: The Disturbing Implications of Algorithmic Bias" In World Policy Journal.
- Gajane, Pratik, and Mykola Pechenizkiy. "On formalizing fairness in prediction with machine learning." arXiv preprint arXiv:1710.03184 (2017).
- Bellamy, Rachel KE, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia et al. "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias." *IBM Journal of Research and Development* 63, no. 4/5 (2019): 4-1.

References

- Ghassemi, Marzyeh, Tristan Naumann, Peter Schulam, Andrew L. Beam, and Rajesh Ranganath. "Opportunities in machine learning for healthcare." arXiv preprint arXiv:1806.00388 (2018).
- Grote, Thomas, and Philipp Berens. "On the ethics of algorithmic decision-making in healthcare." Journal of Medical Ethics 46, no. 3 (2020): 205-211.
- Hajian, Sara., Bonchi, Francesco and Castillo, Carlos ., Algorithmic bias: From discrimination discovery to fairness-aware data mining, KDD Tutorial, 2016.
- Holstein, Kenneth, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. "Improving fairness in machine learning systems: What do industry practitioners need?." In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1-16. 2019.
- Hutchinson, Ben., and Mitchell, Margaret., "50 Years of Test (Un) fairness: Lessons for Machine Learning." In Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 49-58. 2019.
- Hutchinson, Ben., and Mitchell, Margaret., Translation Tutorial: A History of Quantitative Fairness in Testing, FAT* Tutorial, 2019.
- Jensen, Arthur R. "Bias in mental testing." (1980).
- Joseph, Matthew, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. "Rawlsian fairness for machine learning." arXiv preprint arXiv:1610.09559 1, no. 2 (2016).
- Kusner, Matt J., Joshua Loftus, Chris Russell, and Ricardo Silva. "Counterfactual fairness." In Advances in Neural Information Processing Systems, pp. 4066-4076. 2017
- Kamiran, Faisal, and Toon Calders. “Classifying Without Discriminating.” In *Proc. 2nd International Conference on Computer, Control and Communication*, 2009.
- Kilbertus, Niki, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. "Avoiding discrimination through causal reasoning." In *Advances in Neural Information Processing Systems*, pp. 656-666. 2017.
- Krieger, Nancy. "Discrimination and health inequities." *International Journal of Health Services* 44, no. 4 (2014): 643-710.

References

- Liu, Lydia T., Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. “Delayed Impact of Fair Machine Learning.” In *Proc. 35th ICML*, 3156–64, 2018.
- Miller, Tim. “Explanation in artificial intelligence: Insights from the social sciences.” arXiv preprint arXiv:1706.07269 (2017).
- Nabi, Razieh, and Ilya Shpitser. "Fair inference on outcomes." In *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.
- Narayanan, Arvind 21 fairness definitions and their politics, FAT* Tutorial, ACM Conference on Fairness, Accountability, and Transparency (ACM FAT*) 2018
- Rajkomar, Alvin, Michaela Hardt, Michael D. Howell, Greg Corrado, and Marshall H. Chin. "Ensuring fairness in machine learning to advance health equity." Annals of internal medicine 169, no. 12 (2018): 866-872.
- Rajkomar, Alvin, Jeffrey Dean, and Isaac Kohane. "Machine learning in medicine." New England Journal of Medicine 380, no. 14 (2019): 1347-1358.
- Rudin, Cynthia, and Berk Ustun. "Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice." Interfaces 48, no. 5 (2018): 449-466.
- Smedley BD, Stith AY, Nelson AR. Institute of medicine, committee on understanding and eliminating racial and ethnic disparities in health care. Unequal Treatment: Confronting Racial and Ethnic Disparities in Healthcare Washington, DC: National Academies Press; 2003.
- Vayena, Effy, Alessandro Blasimme, and I. Glenn Cohen. "Machine learning in medicine: addressing ethical challenges." PLoS medicine 15, no. 11 (2018).
- Tamayo-Sarver, Joshua H., Susan W. Hinze, Rita K. Cydulka, and David W. Baker. "Racial and ethnic disparities in emergency department analgesic prescription." American journal of public health 93, no. 12 (2003): 2067-2073.
- Thomas SB, Quinn SC. The Tuskegee Syphilis Study, 1932 to 1972: implications for HIV education and AIDS risk education programs in the black community. *Am J Public Health*. 1991;81(11):1498-1505. doi:10.2105/ajph.81.11.1498
- Zhang, Junzhe, and Elias Bareinboim. “Fairness in Decision-Making — the Causal Explanation Formula.” In *Proc. 32nd AAAI*, 2018.

References

- Dimanov, Botty, Umang Bhatt, Mateja Jamnik, and Adrian Weller. "You Shouldn't Trust Me: Learning Models Which Conceal Unfairness From Multiple Explanation Methods." In *SafeAI@ AAAI*, pp. 63-73. 2020.
- White, Douglas B., and Derek C. Angus. "A proposed lottery system to allocate scarce COVID-19 medications: promoting fairness and generating knowledge." *Jama* 324, no. 4 (2020): 329-330.
- Friedman, Batya, and Helen Nissenbaum. "Bias in computer systems." *ACM Transactions on Information Systems (TOIS)* 14, no. 3 (1996): 330-347.
- Drew Fudenberg, David K. Levine, Fairness, risk preferences and independence: Impossibility theorems, Journal of Economic Behavior & Organization, Volume 81, Issue 2, 2012, Pages 606-612.
- Dutta, Sanghamitra, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush Varshney. "Is there a trade-off between fairness and accuracy? A perspective using mismatched hypothesis testing." In International Conference on Machine Learning, pp. 2803-2813., 2020.
- Liu, Suyun, and Luis Nunes Vicente. "Accuracy and fairness trade-offs in machine learning: A stochastic multi-objective approach." *arXiv preprint arXiv:2008.01132* (2020)
- Kim, Joon Sik, Jiahao Chen, and Ameet Talwalkar. "Model-Agnostic Characterization of Fairness Trade-offs." arXiv preprint arXiv:2004.03424 (2020).
- Cummings, Rachel, Varun Gupta, Dhamma Kimpara, and Jamie Morgenstern. "On the compatibility of privacy and fairness." In Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization, pp. 309-315. 2019.
- Kuppam, Satya, Ryan McKenna, David Pujol, Michael Hay, Ashwin Machanavajjhala, and Gerome Miklau. "Fair decision making using privacy-protected data." arXiv preprint arXiv:1905.12744 (2019).
- Agarwal, Sushant Trade-offs Between Fairness, Interpretability, and Privacy in Machine Learning Master's Thesis 2020
- Powe, Neil R. "Black kidney function matters: use or misuse of race?." *Jama* 324, no. 8 (2020): 737-738.

References

- Beil, Michael, Ingo Proft, Daniel van Heerden, Sigal Sviri, and Peter Vernon van Heerden. "Ethical considerations about artificial intelligence for prognostication in intensive care." Intensive Care Medicine Experimental 7, no. 1 (2019): 70.
- Braun, Lundy. "Race, ethnicity and lung function: a brief history." Canadian journal of respiratory therapy: CJRT= Revue canadienne de la therapie respiratoire: RCTR 51, no. 4 (2015): 99.
- Karthik, S. "The Impossibility Theorem of Machine Fairness--A Causal Perspective." arXiv e-prints (2020): arXiv-2007.
- Chouldechova, Alexandra. "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments." Big data 5, no. 2 (2017): 153-163
- Creager, Elliot, David Madras, Jörn-Henrik Jacobsen, Marissa A. Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. "Flexibly fair representation learning by disentanglement." arXiv preprint arXiv:1906.02589 (2019).
- D'Amour, Alexander, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D. Sculley, and Yoni Halpern. "Fairness is not static: deeper understanding of long term fairness via simulation studies." In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 525-534. 2020.
- Du, Mengnan, Fan Yang, Na Zou, and Xia Hu. "Fairness in deep learning: A computational perspective." IEEE Intelligent Systems (2020).
- Bose, Avishek Joey, and William L. Hamilton. "Compositional fairness constraints for graph embeddings." arXiv preprint arXiv:1905.10674 (2019).
- Creager, Elliot, David Madras, Jörn-Henrik Jacobsen, Marissa A. Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. "Flexibly fair representation learning by disentanglement." arXiv preprint arXiv:1906.02589 (2019).
- Foulds, James R., Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. "An intersectional definition of fairness." In 2020 IEEE 36th International Conference on Data Engineering (ICDE), pp. 1918-1921. IEEE, 2020.

References

- Beutel, Alex, Jilin Chen, Zhe Zhao, and Ed H. Chi. "Data decisions and theoretical implications when adversarially learning fair representations." arXiv preprint arXiv:1707.00075 (2017).
- McCradden, Melissa D., Shalmali Joshi, Mjaye Mazwi, and James A. Anderson. "Ethical limitations of algorithmic fairness solutions in health care machine learning." The Lancet Digital Health 2, no. 5 (2020): e221-e223.
- Ustun, Berk, Yang Liu, and David Parkes. "Fairness without harm: Decoupled classifiers with preference guarantees." In International Conference on Machine Learning, pp. 6373-6382. 2019.
- Wang, Xiaoqian, and Heng Huang. "Approaching machine learning fairness through adversarial network." arXiv preprint arXiv:1909.03013 (2019).
- Zhang, Brian Hu, Blake Lemoine, and Margaret Mitchell. "Mitigating unwanted biases with adversarial learning." In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pp. 335-340. 2018.
- Awwad, Yazeed, Richard Fletcher, Daniel Frey, Amit Gandhi, Maryam Najafian, and Mike Teodorescu. Exploring fairness in Machine Learning for international development. CITE MIT D-Lab, 2020.
- Becker, Gary S. The economics of discrimination. University of Chicago press, 2010.
- Benjamin, Ruha. "Assessing risk, automating racism." Science 366, no. 6464 (2019): 421-422
- D'Amour, Alexander, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen et al. "Underspecification Presents Challenges for Credibility in Modern Machine Learning." *arXiv preprint arXiv:2011.03395* (2020).
- Jung, Kenneth, and Nigam H. Shah. "Implications of non-stationarity on predictive modeling using EHRs." Journal of biomedical informatics 58 (2015): 168-174.
- Marx, Charles, Flavio Calmon, and Berk Ustun. "Predictive multiplicity in classification." In International Conference on Machine Learning, pp. 6765-6774. PMLR, 2020.
- Zhang, Wenbin, and Liang Zhao. "Online Decision Trees with Fairness." *arXiv preprint arXiv:2010.08146* (2020).

References

- Coravos, Andy, Irene Chen, Ankit Gordhandas, and Ariel Dora Stern. "We should treat algorithms like prescription drugs." (2019).
- Sculley, David, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. "Hidden technical debt in machine learning systems." In Advances in neural information processing systems, pp. 2503-2511. 2015.
- Saha, Debjani, Candice Schumann, Duncan Mcelfresh, John Dickerson, Michelle Mazurek, and Michael Tschantz. "Measuring non-expert comprehension of machine learning fairness metrics." In International Conference on Machine Learning, pp. 8377-8387. PMLR, 2020.
- Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg et al. "The FAIR Guiding Principles for scientific data management and stewardship." Scientific data 3, no. 1 (2016): 1-9.
- O'neil, Cathy. Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books, 2016.
- Savitt, Todd Lee. Medicine and slavery: The diseases and health care of blacks in antebellum Virginia. Vol. 82. University of Illinois Press, 2002.