Lecture 15: Fairness Definitions

CIS 7000: Trustworthy Machine Learning

Spring 2024

Homework 2

- Due tonight at 8pm

Bias in Machine Learning

- ML models may be biased against minorities

Bias in Machine Learning

- ML models may be biased against minorities

Gender stereotype *she-he* analogies.

sewing-carpentry	register-nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	hairdresser-barber

Gender appropriate *she-he* analogies.

queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

Bolukbasi et al. 2016 : https://arxiv.org/abs/1607.06520
Image from: https://www.analyticsvidhya.com/blog/2017/06/word-embeddings-count-word2veec/

Agenda

- Sources of bias
- Naïve fairness definitions
- Group fairness
- Other fairness definitions

Sources of Bias

- **Data representation:** Distribution of inputs $p(x)$
- **Tainted labels:** Distribution of label assignments $p(y \mid x)$
- **Sensitive features:** Selecting what features to include for each sample (e.g., whether to include sensitive attributes such as race and gender)

Data Representation

- Less data from minority groups → Higher error on minority groups
- **Example:** Many clinical trials historically recruited largely white males, leading to biases in understanding outcomes and side effects
- **Example:** Focus on easily accessible data (e.g. recent tweets, or easily measured features of people) can lead to biased datasets
- Need to be careful to gather representative datasets

Tainted Labels

- **Example:** Amazon hiring bias
  - Amazon's ML resume screening tool to predict hiring decisions based on 10 years of historical applicant data; but found it was biased against women
  - Labels tainted by historical bias
  - https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

- **Similar example**
  - Company filters hires by predicting how long they will stay at the company
  - But how long someone stays depends on how they were treated

Tainted Labels

- **Example:** Predictive policing
  - “PredPol” predictive policing system employed in some police departments
  - Suppose that crime happens equally everywhere
  - Some areas more policed → More crime found in those areas
  - ML learns to predict crime in neighborhoods that were more policed

Tainted Labels

- Need to be careful that labels are unbiased

- However, can be very hard to unbias data!
  - “We should strive to avoid giving women lower salaries”
  - **ML model:** “women” = “lower salaries”

Sensitive Attributes as Features

- When should sensitive attributes be used as features?

- **Example:** Predicting diabetes risk
  - Race is a sensitive attribute that may not cause diabetes, but may be correlated with unrecorded features that cause diabetes
  - What if an insurance company decides that people of some races are at higher risk and should pay higher premium?

- Omitting sensitive attributes is not enough!
  - Other features such as current income may be correlated with race/gender

Sources of Bias

- Need to gather representative sample
- Need to ensure labels are unbiased
- Need to think carefully about whether to include sensitive attributes

Agenda

- Sources of bias
- Naïve fairness definitions
- Group fairness
- Other fairness definitions

Fairness and ML

- What does it mean to be fair?

Accuracy and Fairness

- Low accuracy can result in unfairness
  - **Example:** Strong student as likely as weak student to be admitted into college
  - But highest accuracy model is not necessarily the most fair

- **Example:** Poor student less likely to be admitted into college because they could not afford SAT preparation
  - https://www.nytimes.com/interactive/2023/10/23/upshot/sat-inequality.html
  - https://www.nytimes.com/2024/01/07/briefing/the-misguided-war-on-the-sat.html
  - https://www.nytimes.com/interactive/2024/03/09/upshot/affirmative-action-alternatives.html

Blind Fairness

- Predictive model should ignore sensitive attributes
- **Problem:** Other attributes may be correlated with sensitive attributes
  - Race is correlated with poverty
- **Problem:** It is “fair” to randomly predict for one subgroup as long as sensitive attributes are omitted

Blind Fairness

- **Legally Protected Attributes**
  - Race, sex, color, religion, national origin (Civil Rights Act of 1964, Equal Pay Act of 1963)
  - Age (Discrimination in Employment Act of 1967)
  - Citizenship (Immigration Reform and Control Act)
  - Pregnancy (Pregnancy Discrimination Act)
  - Familial status (Civil Rights Act of 1968)
  - Disability (Rehabilitation Act of 1973; Americans with Disabilities Act of 1990)
  - Veteran status (Vietnam Era Veterans' Readjustment Assistance Act of 1974; Uniformed Services Employment and Reemployment Rights Act)
  - Genetic information (Genetic Information Nondiscrimination Act)

Case Study: Criminal Justice

- Software by Northpointe to predict **recidivism** for defendants
  - I.e., risk of committing future crimes

- Used to help make bail, sentencing, and parole decisions

Case Study: Criminal Justice

- **Features:** 137 questions answered by defendants or criminal records:
  - “Was one of your parents ever sent to jail or prison?”
  - “How many of your friends/acquaintances are taking drugs illegally?”
  - “How often did you get in fights while at school?”
  - Agree or disagree? “A hungry person has a right to steal”
  - Agree or disagree? “If people make me angry or lose my temper, I can be dangerous.”

- Exact algorithm and model is a trade secret

Case Study: Criminal Justice

- Race is **not** a feature

- **Problem:** Correlated features
  - E.g., poverty, joblessness and social marginalization
  - One of the developers of the system said it is difficult to construct a score that doesn't include items that can be correlated with race
  - "If those are omitted from your risk assessment, accuracy goes down"

- Similar to Amazon hiring bias example

Individual Fairness

- "Similar" individuals (differing only on sensitive attributes) should receive "similar" outcomes
  - The prediction function $f: X \to Y$ should be Lipschitz continuous:

$$\|x - x'\| \leq \epsilon \Rightarrow |f(x) - f(x')| \leq \epsilon'$$

- **Problem:** How to define "similar"?
  - What if we include someone's accent or attire as a feature?
  - Accent may be correlated with race, in which case $\|x - x'\|$ is always large for two individuals of different race, even if race is not included as a feature

- **Problem:** Scales poorly to high-dimensional spaces

Agenda

- Sources of bias
- Naïve fairness definitions
- Group fairness
- Other fairness definitions

Group Fairness

- Equalize “fairness metrics” across “subgroups”

- **Remaining challenges**
  - Need to define “subgroups” (e.g., ethnicity, gender, etc.)
  - Need to define “fairness metrics” (e.g., rate of positive outcomes, false positive/negative rates, etc.)

Group Fairness

- **Problem setup**
  - Sensitive attribute A
  - ML model R mapping input features X to prediction $\hat{Y} = R(X)$
  - True outcome Y (typically binary, and $Y = 1$ is the "good" outcome)

- **Group fairness:** Account for performance on subgroups

$$\text{Fairness metric} = F\big(L(f; X_1), \ldots, L(f; X_k)\big)$$

- **Example:** Insurance risk prediction
  - $A = \text{age}$, $R = \text{predicted cost}$, $Y = \text{true cost}$

Group Fairness: Independence

- Risk score distribution should be equal across ages:

$$P(\text{risk score} \mid \text{age}) = P(\text{risk score})$$

  - E.g., equal proportion of low risk customers for young vs. old people
  - Often called **demographic parity**

- **Intuition:** Subgroups should receive positive outcomes at similar rates
  - “Four-fifths rule” for assessing discrimination in hiring
  - https://www.eeoc.gov/laws/guidance/select-issues-assessing-adverse-impact-software-algorithms-and-artificial

Group Fairness: Independence

- **Problem:** Can assign randomly for one subgroup, no guarantee on quality of predictions across subgroups
  - Not a problem when **incentives are aligned** (i.e., decision-maker wants to choose best candidates within each subgroup)
  - But this is often not the case!

- **Problem:** What if the base rates are not equal?
  - In our insurance example, what if drivers in lower age groups in fact behave more riskily?

Group Fairness: Separation

- **Separation:** Risk score should be independent of age given outcome:

$$P(\text{risk score} \mid \text{age}, \text{true outcome}) = P(\text{risk score} \mid \text{true outcome})$$

  - Equivalently, true and false positive rates are equal across subgroups
  - Often called **equality of odds** (very similar to **equality of opportunity**)

- **Example:** Both of the following hold:
  - Fraction of young, low-insurance-usage people correctly identified as low-risk = Fraction of old low-insurance-usage people correctly identified as low-risk
  - Fraction of young high-insurance-usage people wrongly identified as low-risk = Fraction of old high-insurance-usage people wrongly identified as low-risk

Group Fairness: Separation

- **Separation:** Risk score should be independent of age given outcome:

$$P(\text{risk score} \mid \text{age}, \text{true outcome}) = P(\text{risk score} \mid \text{true outcome})$$

  - Equivalently, true and false positive rates are equal across subgroups
  - Often called **equality of odds** (very similar to **equality of opportunity**)

- **Intuition:** The predictive accuracy is equal across subgroups

Case Study: Criminal Justice

MACHINE BIAS

Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say

ProPublica's analysis of bias against black defendants in criminal risk scores has prompted research showing that the disparity can be addressed — if the algorithms focus on the fairness of outcomes.

by Julia Angwin and Jeff Larson, Dec. 30, 2016, 4:44 p.m. EST

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

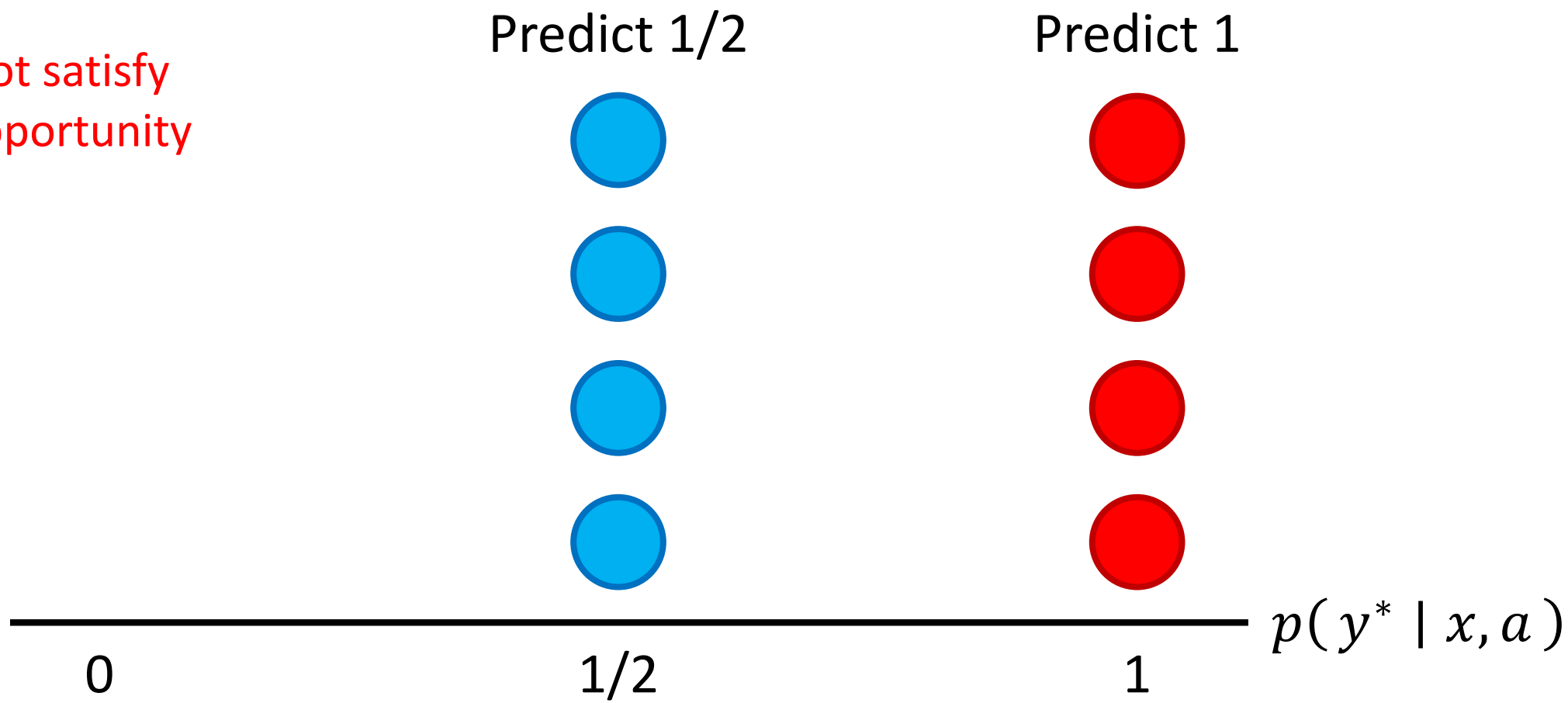
Why is this a good idea?

- Focus on **equality of opportunity**
  - Similar to separation, but only require that false negative rates are equalized
  - **Intuition:** Only care about being fair to qualified individuals

- **Algorithm:** If accuracy for one subgroup is lower, simply give positive outcomes to more individuals in that subgroup

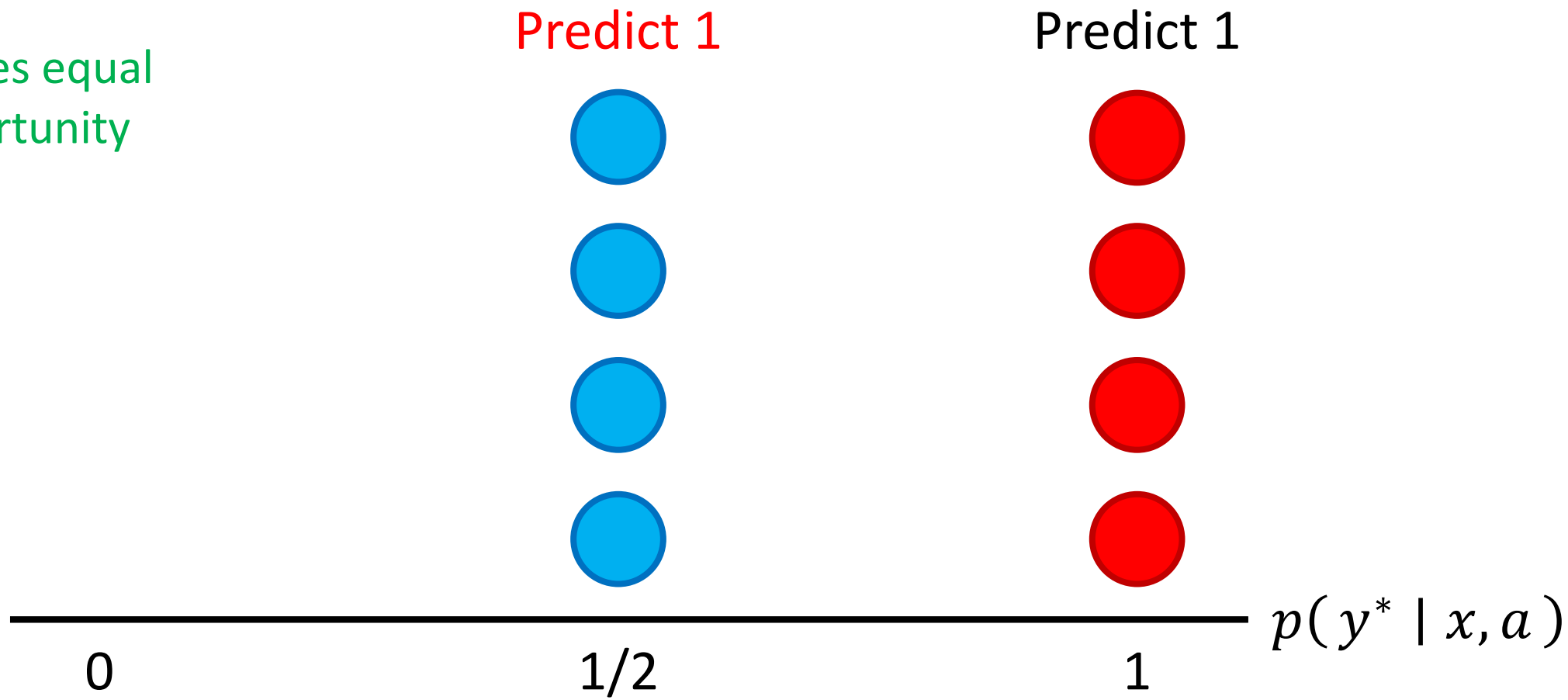
Why is this a good idea?

Does not satisfy equal opportunity



Why is this a good idea?

Satisfies equal opportunity

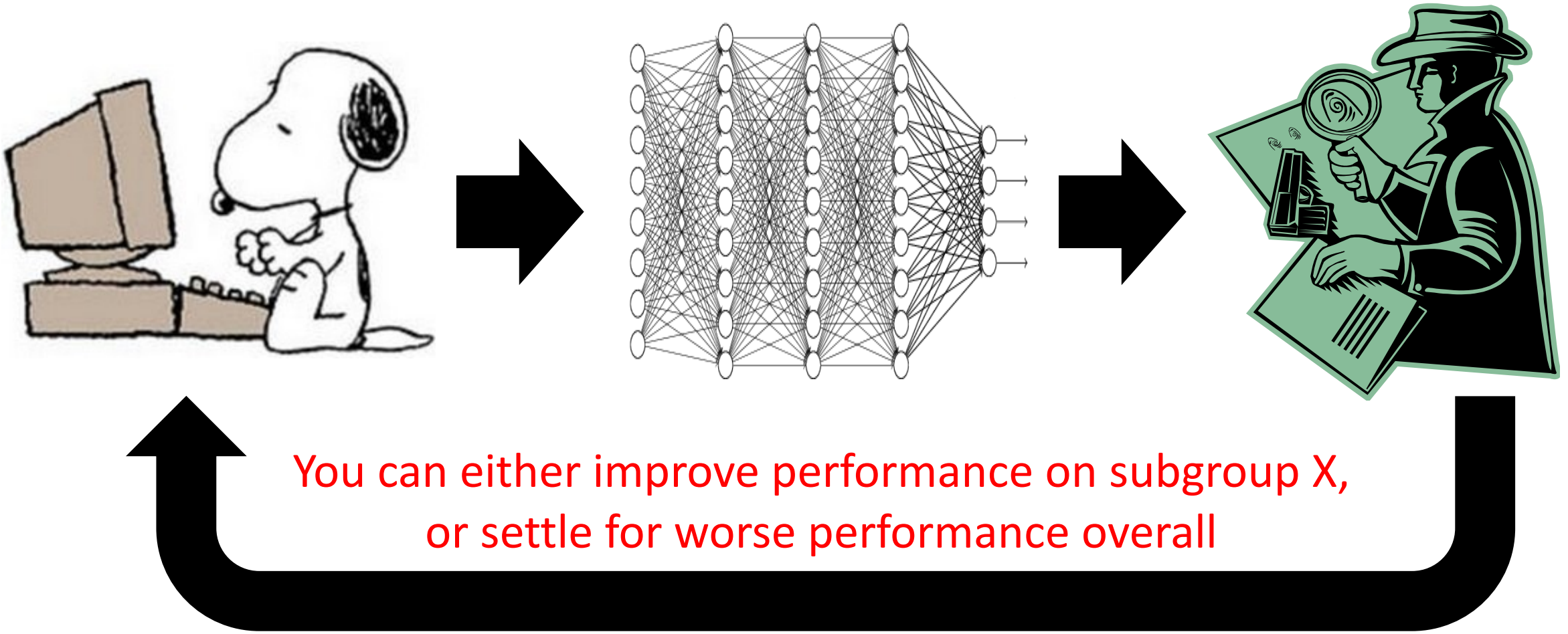


Why is this a good idea?

- Suppose the decision-maker doesn't want to give positive outcomes to too many individuals

- Then, they need to improve model performance on all subgroups
  - Epistemic uncertainty → gather more data
  - Aleatoric uncertainty → maybe just need better features

Why is this a good idea?



You can either improve performance on subgroup X,
or settle for worse performance overall

Group Fairness: Sufficiency

- Outcome should be independent of risk score given age:

$$P(\text{true outcome}, \text{age} \mid \text{risk score}) = P(\text{true outcome} \mid \text{risk score})$$

- Equivalently, calibrated conditional on each subgroup

Group Fairness

Non-discrimination criteria		
Independence	Separation	Sufficiency
$R \perp A$	$R \perp A \mid Y$	$Y \perp A \mid R$

Group Fairness

- Three notions are incompatible!

Proposition 2. *Assume that A and Y are not independent. Then sufficiency and independence cannot both hold.*

Proposition 3. *Assume Y is binary, A is not independent of Y, and R is not independent of Y. Then, independence and separation cannot both hold.*

Proposition 5. *Assume Y is not independent of A and assume $\hat{Y}$ is a binary classifier with nonzero false positive rate. Then, separation and sufficiency cannot both hold.*

- Thus, need carefully choose what kinds of fairness we ask for

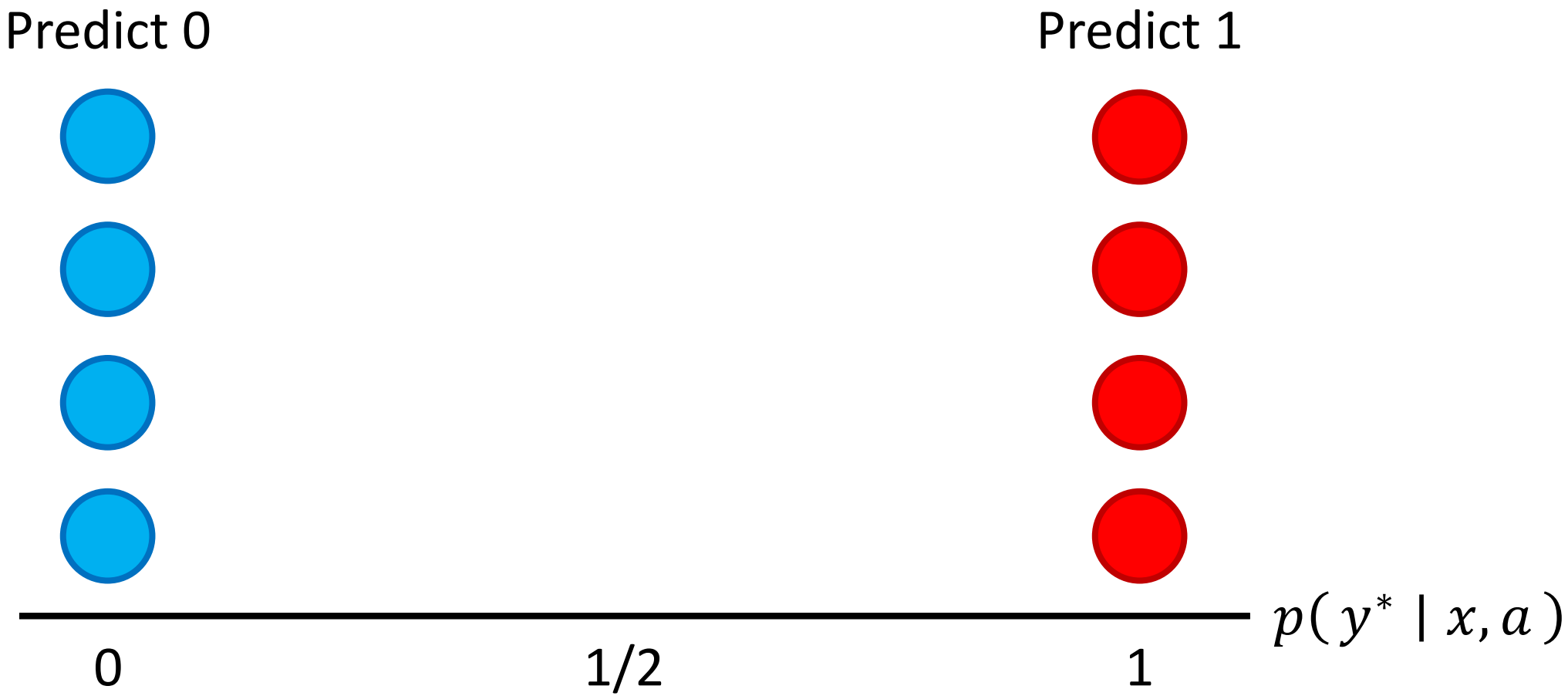
Solon Barocas, Moritz Hardt, Arvind Narayanan, “Fairness and Machine Learning”

Examples

Does not satisfy independence / demographic parity

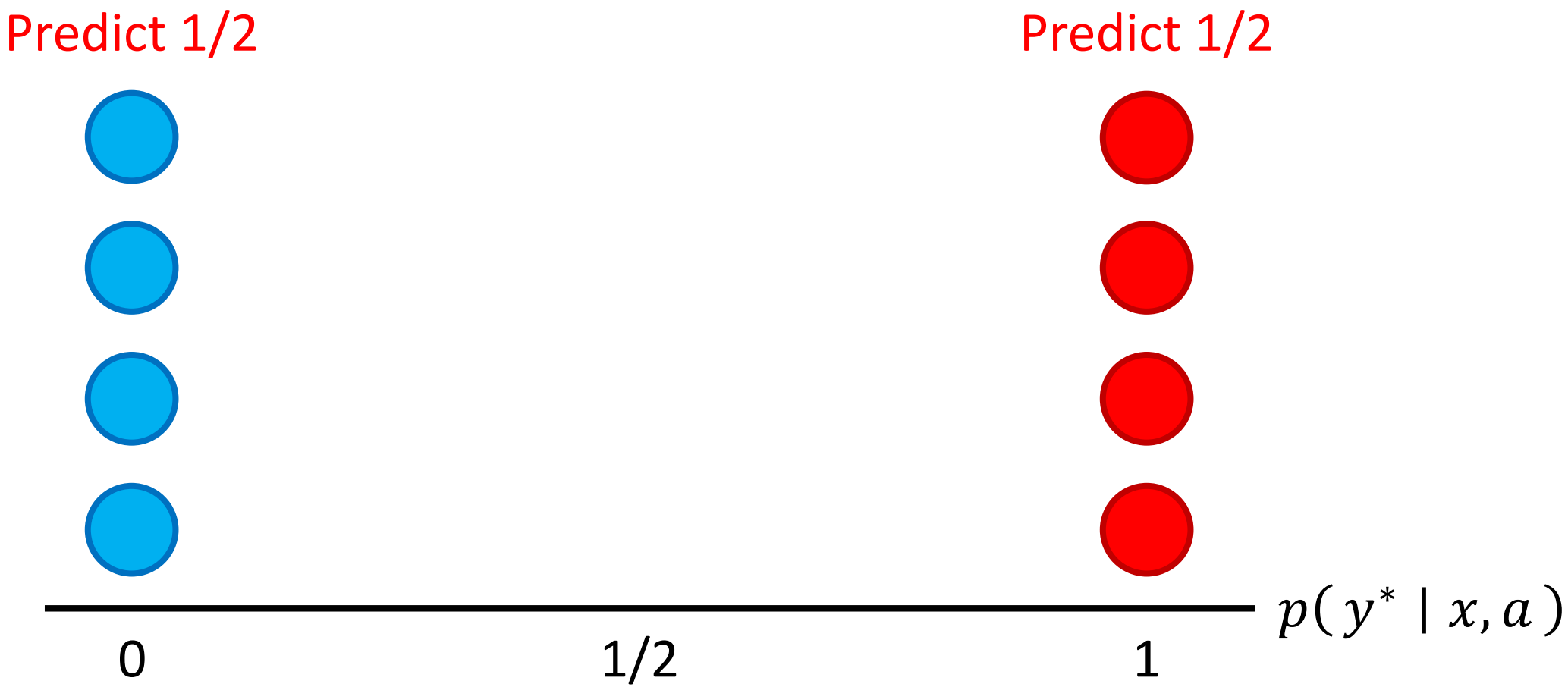
Satisfies separation / equal odds

Satisfies sufficiency / calibration



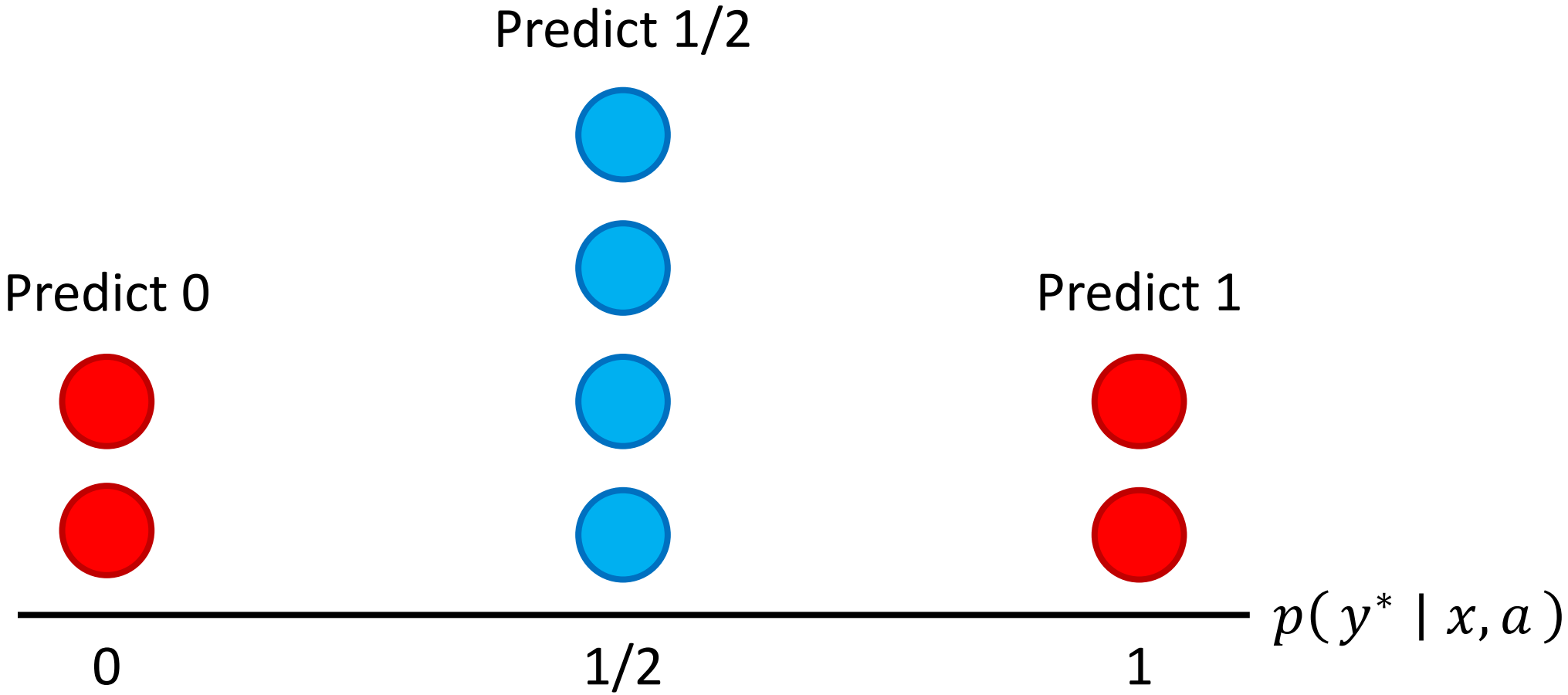
Examples

Satisfies independence / demographic parity
Does not satisfy separation / equal odds
Does not satisfy sufficiency / calibration



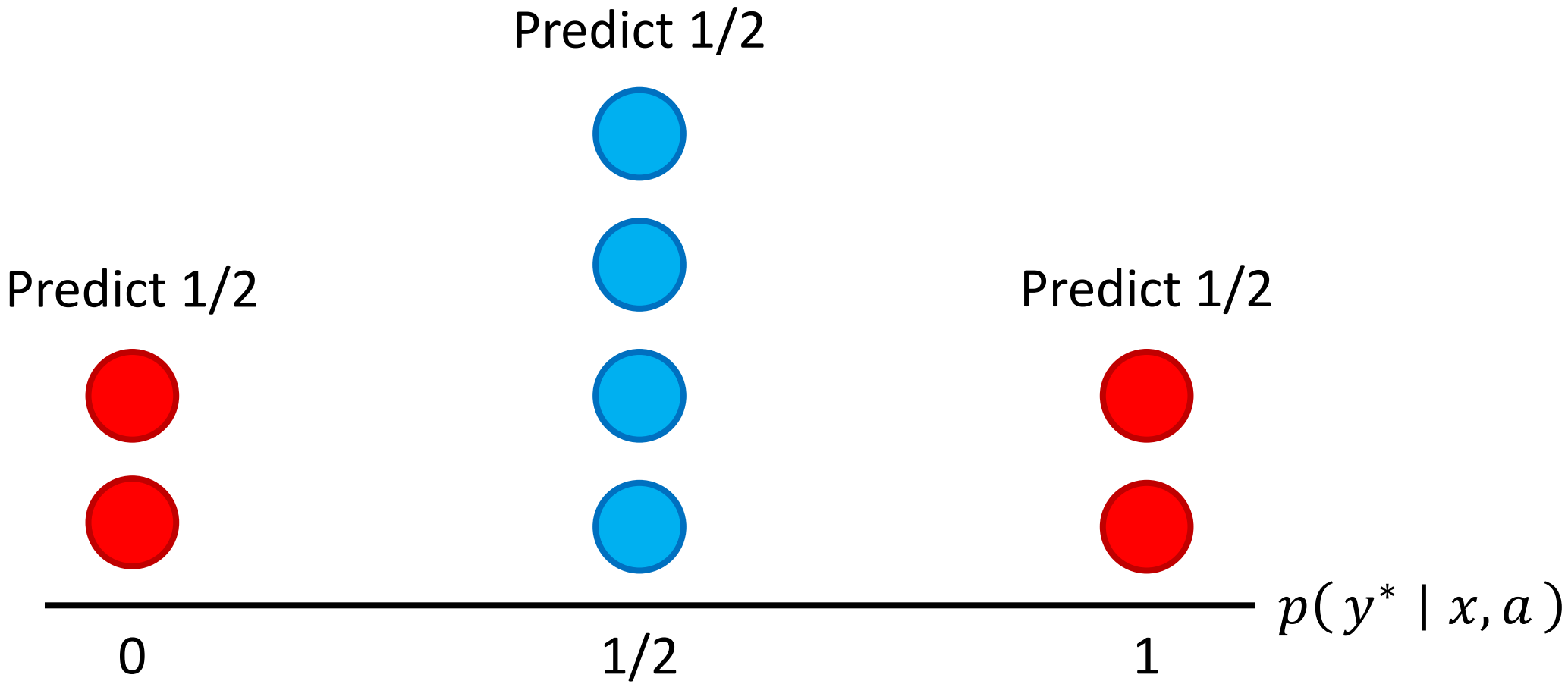
Examples

Satisfies independence / demographic parity
Does not satisfy separation / equal odds
Satisfies sufficiency / calibration



Examples

Satisfies independence / demographic parity
Satisfies separation / equal odds
Satisfies sufficiency / calibration

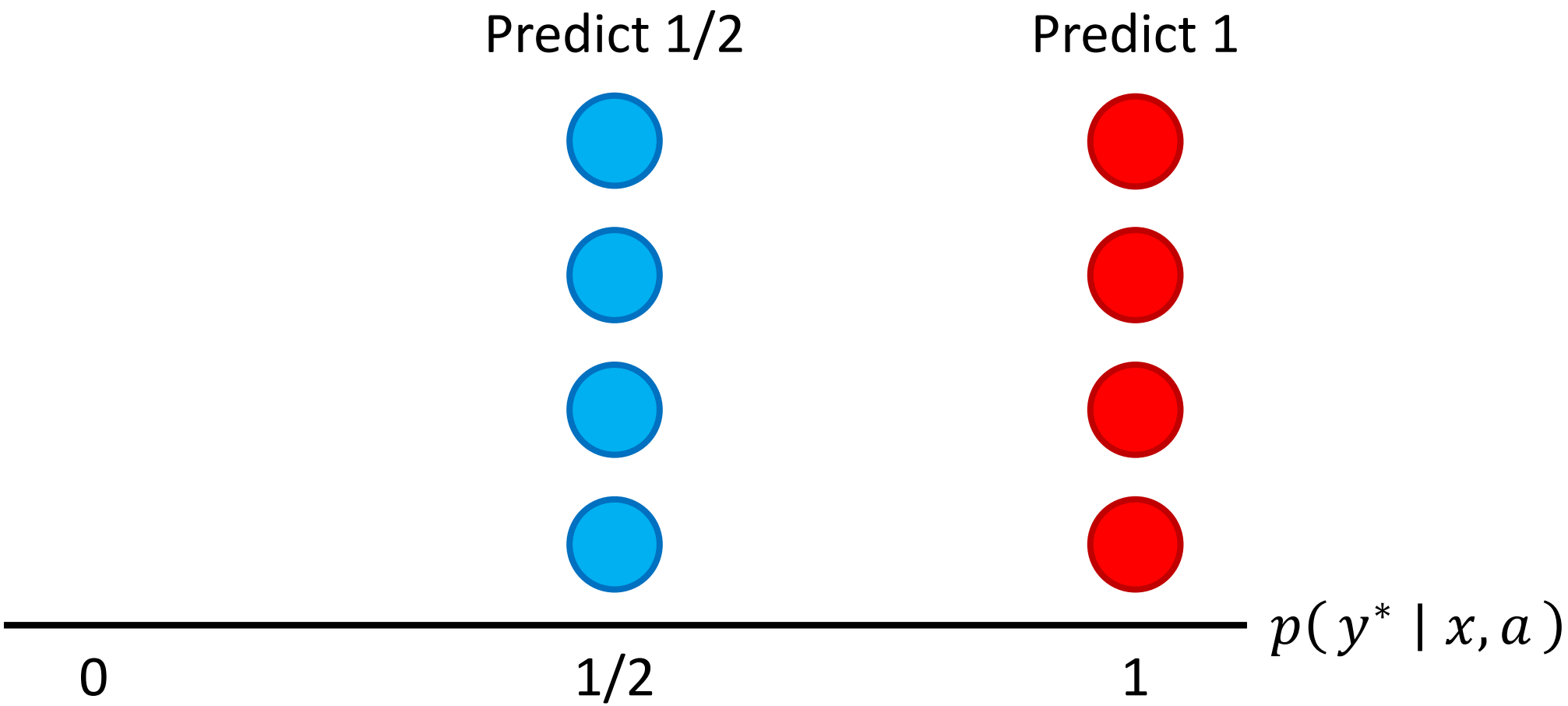


Examples

Does not satisfy independence / demographic parity

Does not satisfy separation / equal odds

Satisfies sufficiency / calibration

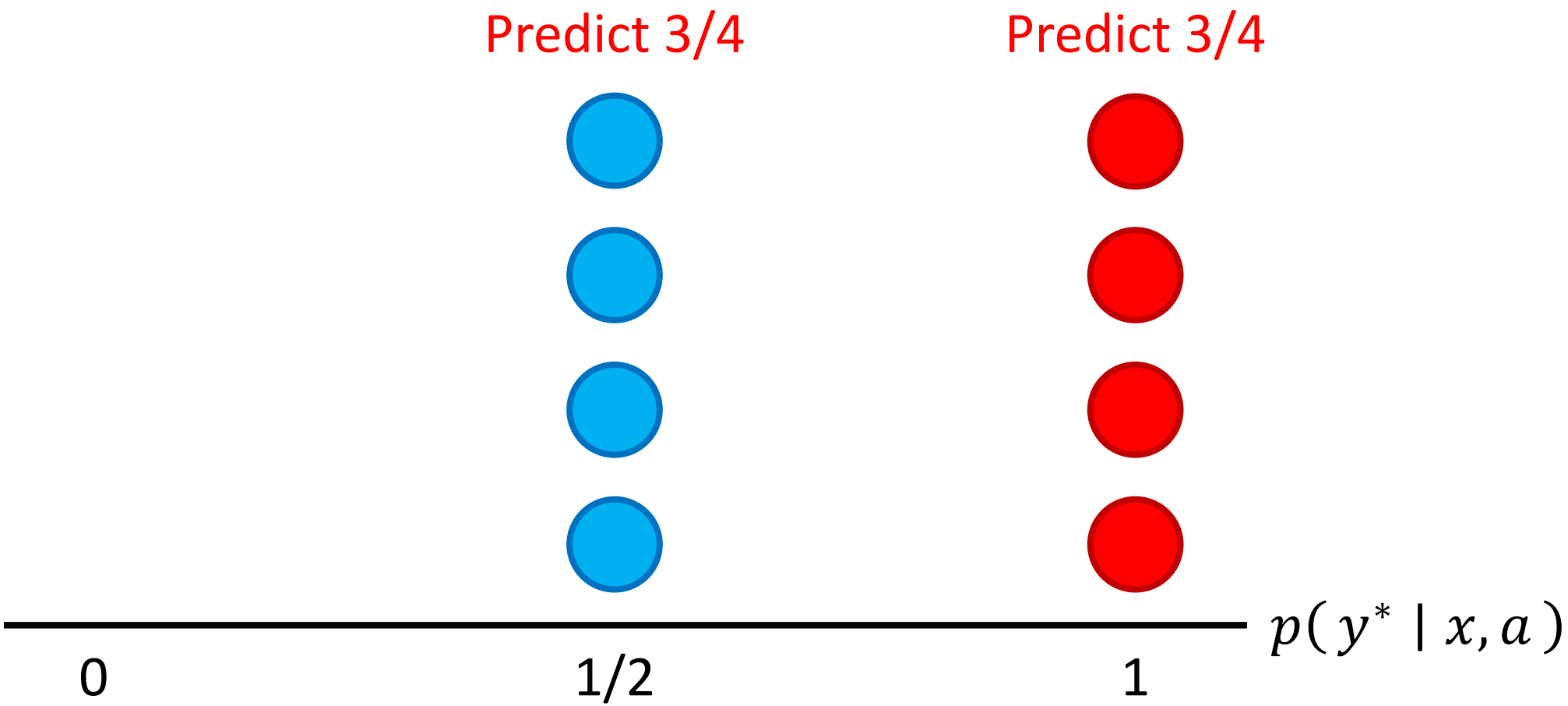


Examples

Satisfies independence / demographic parity

Does not satisfy separation / equal odds

Does not satisfy sufficiency / calibration



Algorithms for Ensuring Fairness

- Given a notion of fairness, there are a few ways of achieving it

- **Example:** Independence
  - **Pre-processing:** Adjust features to be uncorrelated with sensitive attribute
  - **Training constraints:** Impose the constraint during training
  - **Post-processing:** Adjust the learned classifier so its predictions are uncorrelated with the sensitive attribute

- **Goodhart's law:** "When a measure becomes a target, it ceases to be a good measure" – Marilyn Strathern
  - Do not blindly impose fairness, need to carefully examine predictions

Human-in-the-Loop Fairness

- **Potential solution:** Have domain experts weigh in on what performance metrics result in fair model selection/training

- **Challenges**
  - Experts may not understand limitations of ML models (e.g., does a judge using a system understand that it only has 60% accuracy?)
  - Potential for selective enforcement based on human biases

Human-in-the-Loop Fairness

- **Example:** In bail decision-making, judges selectively follow model
  - Less lenient against younger defendants, especially minorities
  - Younger defendants are actually more risky, but judges may have been lenient due to societal norms (e.g., “second chance”)
  - Judges followed algorithm less and less over time

https://www.washingtonpost.com/business/2019/11/19/algorithms-were-supposed-make-virginia-judges-more-fair-what-actually-happened-was-far-more-complicated/

Agenda

- Sources of bias
- Naïve fairness definitions
- Group fairness
- Other fairness definitions

Causal Fairness

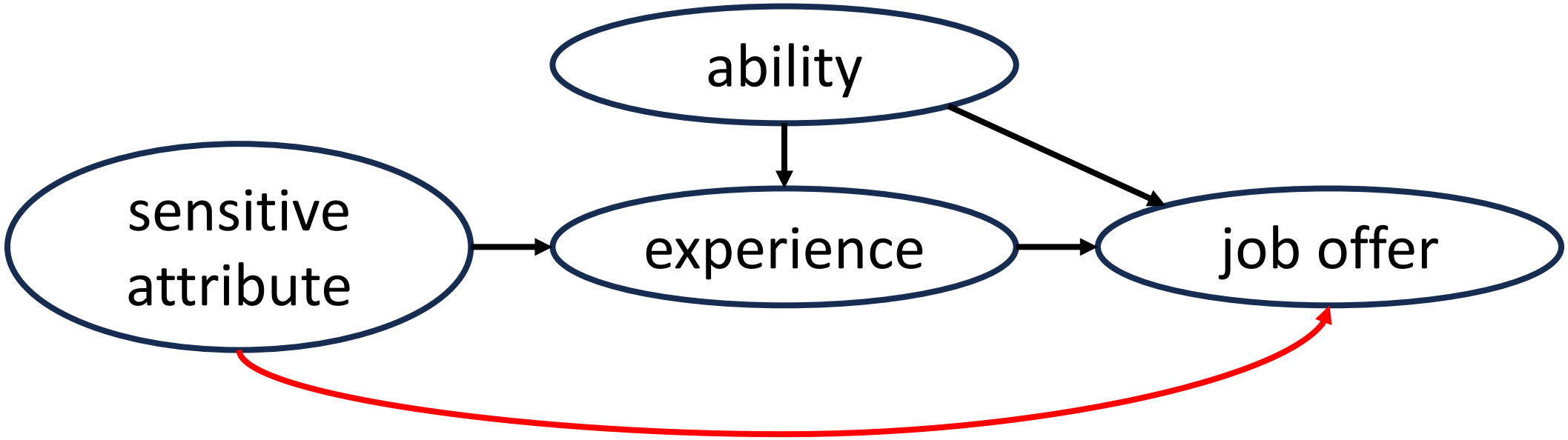
- Fairness definitions discussed so far ignore the difference between causation and correlation
- **Intuitive definition:** The individual should be treated the same as if you changed their sensitive attribute
- **Example**
  - If this applicant's sensitive attribute is changed, would you give them the job?
  - If so, then you should give them the job!

Causal Fairness

- **Challenge:** How to formalize “changed their sensitive attribute”?
  - Maybe if their sensitive attribute were different, they would have had better job experience by avoiding past discrimination
  - How should this be counted?

- Formalized via **counterfactual reasoning**

Causal Fairness



Takes into account effect of sensitive attribute on experience when changing sensitive attribute

Multicalibration

- In group fairness, if the subgroups are disjoint, then post-hoc enforcing fairness is typically easy
  - In calibration, we simply calibrate each subgroup separately

- What if the subgroups overlap?
  - E.g., Gender, ethnicity, age
  - We could calibrate for each intersection of subgroups, but there are exponentially many subgroups in the number of sensitive attributes

- **Multicalibration** uses **boosting** to ensure calibration for all subgroups in a computationally and statistically efficient way

Agenda

- Sources of bias
- Naïve fairness definitions
- Group fairness
- Other fairness definitions