# Algorithmic Fairness: An Overview

Gregory Faletto

University of Southern California

Math 547, November 20 2019

# Setup

- Input: random sample of $n$ observational units $(X^{(i)}, A^{(i)}, Y^{(i)}) \in \mathbb{R}^p \times \{0, 1\}$, $i \in \{1, \ldots, n\}$.
- Practitioner hopes to fit a model $\hat{f}$ trained on the observed data $(X, A, Y)$ to generate predictions $\hat{Y} = \hat{f}(X, A)$ that satisfy some notion of *algorithmic fairness*, in that individuals are not treated differently on the basis of their protected attributes $A^{(i)}$.
- $Y'$: actual response of interest (may not be observed).

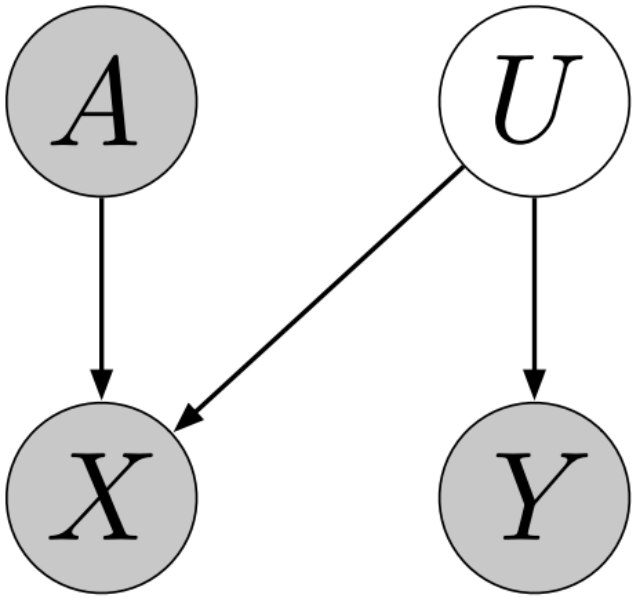
Setup

- Input: random sample of $n$ observational units $(X^{(i)}, A^{(i)}, Y^{(i)}) \in \mathbb{R}^p \times \{0,1\}$, $i \in \{1, \ldots, n\}$.
- Practitioner hopes to fit a model $\hat{f}$ trained on the observed data $(X, A, Y)$ to generate predictions $\hat{Y} = \hat{f}(X, A)$ that satisfy some notion of *algorithmic fairness*, in that individuals are not treated differently on the basis of their protected attributes $A^{(i)}$.
- $Y'$: actual response of interest (may not be observed).

- Goal of algorithmic fairness: create a predictive model that doesn't discriminate with respect to *protected attributes.*
- Assume in this presentation that there is one attribute in $a$ with two categories.
- By assumption we are told which attributes are to be "protected;" determining which attributes should be protected in the first place is outside the scope of algorithmic fairness.
- (e.g., $Y$ is whether someone is arrested. Actual response of interest is $Y'$, committing a crime.

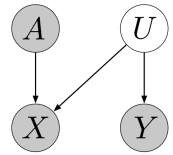
# Fairness Through Unawareness

- *Fairness through unawareness* is the requirement that $\hat{f}$ does not use the information from $A$; that is, $\hat{f} = \hat{f}(X)$.
- Example from Kusner et al. [2017]: regressing on $X$ alone leads to *counterfactual unfairness*, regressing on $A$ and $X$ does not.



Fairness Through Unawareness

- *Fairness through unawareness* is the requirement that $\hat{f}$ does not use the information from $A$; that is, $\hat{f} = \hat{f}(X)$.
- Example from Kusner et al. [2017]: regressing on $X$ alone leads to *counterfactual unfairness*, regressing on $A$ and $X$ does not.




- Most naive criterion: just leave protected attributes out.
- Example of how this goes wrong: suppose car insurance company is trying to set rates. Wants to predict probability driver will get in accident next year. Arrow shows a causal relationship. Assume relationships are linear (plus noise). $U$ is aggression in driving (latent). $X$ is buying a red car. $A$ is religion; Religion 1 is more likely to buy a red car than Religion 2.
- Authors show that under certain assumption regressing $Y$ against $X$ alone leads to unfair outcomes, whereas regressing against both $A$ and $X$ does not.
- Details are too long to explain here, but intuition is that controlling for $A$ allows model to compensate for fact that groups differing in $A$ differ in baseline levels for $X$.
- Fairness through unawareness doesn't work. What's next?

# Demographic Parity

*Demographic parity* is the requirement that

$$\mathbb{P}(\hat{Y}^{(1)} = \hat{y} \mid A^{(1)} = 0) = \mathbb{P}(\hat{Y}^{(1)} = \hat{y} \mid A^{(1)} = 1), \qquad \forall \hat{y} \in \{0, 1\}, \quad (1)$$

where the probability is taken over the distribution of the random sample of the observed data and any randomness used to create the model.

*Demographic parity* is the requirement that

$$\mathbb{P}(\hat{Y}^{(1)} = \hat{y} \mid A^{(1)} = 0) = \mathbb{P}(\hat{Y}^{(1)} = \hat{y} \mid A^{(1)} = 1), \qquad \forall \hat{y} \in \{0, 1\}, \quad (1)$$

where the probability is taken over the distribution of the random sample of the observed data and any randomness used to create the model.

- Kind of like affirmative action.
- Makes sense if $Y' \perp\!\!\!\perp A$, so $\hat{Y}$ ought to be independent of $A$ as well in a fair model.
- Advantages: intuitive, simple, good "gut check."
- Disadvantages: can result in significantly different treatment of two people with similar attributes who differ only on protected attribute, if distributions of $X$ differ significantly across groups.
- Also makes sense even if $Y'$ is not independent of $A$, if you believe differences in distributions of $X$ across different protected groups only exist because of discrimination you'd like to compensate for.
- Might be a little too aggressive—doesn't take into account that distribution of $Y'$ might be different in groups with different protected statuses. Can we take that into account?

# Equalized Odds

*Equalized odds* is the requirement that

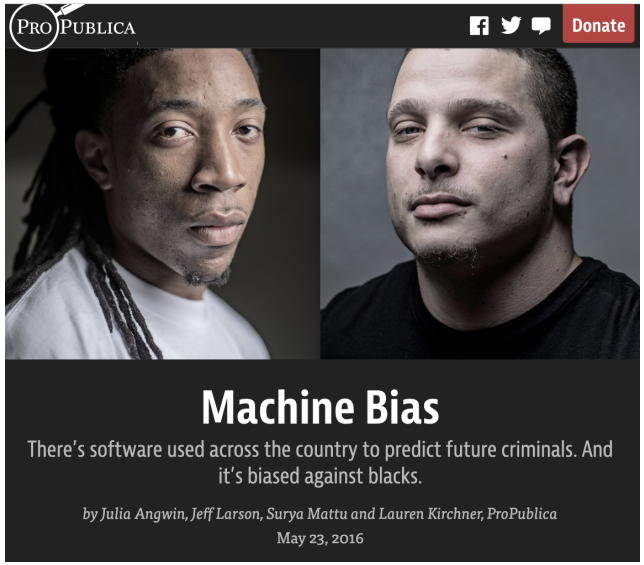
$$\mathbb{P}(\hat{Y}^{(1)} = \hat{y} \mid Y^{(1)} = y, A^{(1)} = 0) = \mathbb{P}(\hat{Y}^{(1)} = \hat{y} \mid Y^{(1)} = y, A^{(1)} = 1), \quad \forall \hat{y}, y \in \{0, 1\}. \quad (2)$$

*Equalized odds* is the requirement that

$$\mathbb{P}(\hat{Y}^{(1)} = \hat{y} \mid Y^{(1)} = y, A^{(1)} = 0) = \mathbb{P}(\hat{Y}^{(1)} = \hat{y} \mid Y^{(1)} = y, A^{(1)} = 1), \quad \forall \hat{y}, y \in \{0, 1\}. \quad (2)$$

- Intuition: given that a person really has state $y$, probability that algorithm will predict that state is equal across protected groups.
- Another way of saying: $\hat{Y}$ is independent of $A$ conditional on $Y$.
- Makes sense if $Y'$ is independent of $A$ conditional on $Y$ (that is, any measurement error of $Y$ is the same across all protected groups) (special case: $Y' = Y$)
- Advantage: unlike demographic parity, allows us to differentiate between groups based on their observed response $Y$ (takes $Y$ into account).
- Disadvantages: requires observing $Y$. Also, relies a lot on measurement of $Y$—may perpetuate bias if observed values of $Y$ are themselves biased.
- Example of $Y$ depending on $A$: $Y$ is whether someone is arrested. Actual response of interest is $Y'$, committing a crime. Because different races have different levels of contact with law enforcement all else equal, the distribution of $Y$ differs from $Y'$ across race.

# Violation of Equalized Odds




## 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% |


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% |

- Famous case: company called Northpointe created algorithm COMPAS. Purpose: when someone is arrested, run COMPAS, generate predicted probability arrestee will commit another crime in next two years. Used to set bail amounts. Propublica wrote article showing their algorithm violated equalized odds.
- (Table is confusingly labeled/mislabeled: if you examine details, the label "Labeled higher risk, but didn't re-offend" is really "percentage of people who didn't end up committing a crime that were predicted to commit a crime;" that is, (empirical) probability algorithm predicted person would commit a crime conditional on race and actual $Y$; so this is in fact equalized odds.
- See that algorithm is much harsher on African-American arrestees than white arrestees.

# Calibration

*Calibration* is the requirement that

$$\mathbb{P}(Y^{(1)} = y \mid \hat{Y}^{(1)} = \hat{y}, A^{(1)} = 0) = \mathbb{P}(Y^{(1)} = y \mid \hat{Y}^{(1)} = \hat{y}, A^{(1)} = 1), \quad \forall \hat{y}, y \in \{0, 1\}. \quad (3)$$

Calibration

*Calibration* is the requirement that

$$\mathbb{P}(Y^{(1)} = y \mid \hat{Y}^{(1)} = \hat{y}, A^{(1)} = 0) = \mathbb{P}(Y^{(1)} = y \mid \hat{Y}^{(1)} = \hat{y}, A^{(1)} = 1), \quad \forall \hat{y}, y \in \{0, 1\}. \quad (3)$$

- Reverses equalized odds; given that algorithm predicted a person has state $\hat{y}$, probability that the person really has that state must be equal across protected groups. (Within the sets of people who received the same prediction, $Y$ should be independent of $A$. $Y$ is independent of $A$ conditional on $\hat{Y}$.)
- Borrowed from educational and psychological testing and assessment, where it is widely used
- Disadvantage: like equalized odds, may perpetuate bias if observed values of $Y$ are themselves biased.
- Prima facie, calibration and equalized odds may both seem reasonable and desirable.

# Northpointe's Response: Calibration is Satisfied

- Angwin et al. used the incorrect classification statistics to frame the COMPAS risk scales as biased against blacks. They compared the complements of *Sensitivity* and *Specificity* for blacks and whites. These are operating characteristics calculated separately on recidivists only and non-recidivists only. They should have used the complements of the predictive values that take into account the base rate of recidivism. In their main table, the PP authors misrepresented the *percentage of non-recidivists with a positive test* result ("Not Low" risk level) as the *percentage of persons with a positive test result that did not recidivate* ("Labeled Higher Risk, But Didn't Re-Offend").

| | White | African American |
|---|---|---|
| Labeled Higher Risk, But Didn't Re-Offend | 41% | 37% |
| Labeled Lower Risk, Yet Did Re-Offend | 29% | 35% |

- Angwin et al. used the incorrect classification statistics to frame the COMPAS risk scales as biased against blacks. They compared the complements of *Sensitivity* and *Specificity* for blacks and whites. These are operating characteristics calculated separately on recidivists only and non-recidivists only. They should have used the complements of the predictive values that take into account the base rate of recidivism. In their main table, the PP authors misrepresented the *percentage of non-recidivists with a positive test* result ("Not Low" risk level) as the *percentage of persons with a positive test result that did not recidivate* ("Labeled Higher Risk, But Didn't Re-Offend").

| | White | African American |
|---|---|---|
| Labeled Higher Risk, But Didn't Re-Offend | 41% | 37% |
| Labeled Lower Risk, Yet Did Re-Offend | 29% | 35% |

- Northpoint countered Propublica by showing their algorithm satisfied calibration.
- Used same phrasing as Propublica article because they argued their metric was "correct" and Propublica's was wrong.
- Not actually clear that calibration is "more correct"—subject of research
- Interesting: both criteria seem reasonable and desirable, yet one is strongly violated by the exact same algorithm that satisfies the other. What's going on?

# Calibration and Equalized Odds are Mutually Incompatible

**Theorem (Theorem 1.1, Kleinberg et al. [2017])**

*Suppose an algorithm satisfies both calibration and equalized odds. Then either $Y \perp\!\!\!\perp A$ or the relationship between $Y$ and $X$ within each subgroup of $A$ is deterministic ($\mathbb{P}(Y = y \mid X, A = a) \in \{0, 1\}$ for all $a$ and $y$).*

Similar result in Chouldechova [2017].

**Theorem (Theorem 1.1, Kleinberg et al. [2017])**

*Suppose an algorithm satisfies both calibration and equalized odds. Then either $Y \perp A$ or the relationship between $Y$ and $X$ within each subgroup of $A$ is deterministic ($\mathbb{P}(Y = y \mid X, A = a) \in \{0, 1\}$ for all $a$ and $y$).*

Similar result in Chouldechova [2017].

- Probably the single most famous and important result in algorithmic fairness.
- Raises question: how to pick which criterion is better in a given situation? Seems like it might depend on situation—discussed how demographic parity intuitively makes sense if $Y' \perp\!\!\!\perp A$, equalized odds makes sense if $Y'$ is independent of $A$ conditional on $Y$. Would be good to have some results validating intuition.

# Output Disparity Control

*Output disparity control* [Yeom and Tschantz, 2018] is the requirement that

$$d_{TV}(\hat{Y}^{(1)}|A^{(1)}=0, \hat{Y}^{(1)}|A^{(1)}=1) \leq d_{TV}(Y'^{(1)}|A^{(1)}=0, Y'^{(1)}|A^{(1)}=1). \tag{4}$$

where $d_{TV}(X, Y)$ is the total variational distance between random variables $X$ and $Y$:

$$d_{TV}(X, Y) = \frac{1}{2} \sum_{y \in \{0,1\}} |\mathbb{P}(X = y) - \mathbb{P}(Y = y)| .$$

Output Disparity Control

*Output disparity control* [Yeom and Tschantz, 2018] is the requirement that

$$d_{TV}(\hat{Y}^{(1)}|A^{(1)}=0, \hat{Y}^{(1)}|A^{(1)}=1) \leq d_{TV}(Y'^{(1)}|A^{(1)}=0, Y'^{(1)}|A^{(1)}=1). \quad (4)$$

where $d_{TV}(X, Y)$ is the total variational distance between random variables $X$ and $Y$:

$$d_{TV}(X, Y) = \frac{1}{2} \sum_{y \in \{0,1\}} |\mathbb{P}(X = y) - \mathbb{P}(Y = y)|.$$

- Right side of inequality: total variational distance between distributions of $Y'$ for different values of $A$.
- Basic idea: the distributions of our predictions conditional on $A$ should be no further apart than the distributions in the population.
- Weak notion of fairness; makes sense if increasing disparity with respect to protected attributes can't be good, which seems reasonable. (Seems like violating it is definitely always bad, and we can rule out models that violate it.)
- Agnostic to whether existing disparity is benign or bad, and it seems like there could be ways to reduce disparity that might arguably be unfair (per discussion on demographic parity). So satisfying this condition doesn't mean algorithm is good.

# Summary of Some Results From Yeom and Tschantz [2018]

**Theorem (From Theorems 1, 5, and 6 in Yeom and Tschantz [2018])**

*Any model that satisfies demographic parity also controls output disparity. Further, suppose $Y' \perp\!\!\!\perp A$. Then every model satisfying $\hat{Y} = Y'$ satisfies demographic parity, and models satisfying equalized odds and calibration may allow output disparity amplification.*

**Theorem (From Theorems 2, 4, and 6 in Yeom and Tschantz [2018])**

*Suppose $Y = Y'$. Then every model satisfying $\hat{Y} = Y'$ satisfies equalized odds, and models satisfying demographic parity and calibration may allow output disparity amplification.*

**Theorem (From Theorems 1, 5, and 6 in Yeom and Tschantz [2018])**

*Any model that satisfies demographic parity also controls output disparity. Further, suppose $Y' \perp A$. Then every model satisfying $\hat{Y} = Y'$ satisfies demographic parity, and models satisfying equalized odds and calibration may allow output disparity amplification.*

**Theorem (From Theorems 2, 4, and 6 in Yeom and Tschantz [2018])**

*Suppose $Y = Y'$. Then every model satisfying $\hat{Y} = Y'$ satisfies equalized odds, and models satisfying demographic parity and calibration may allow output disparity amplification.*

- Demographic parity seems best if $Y' \perp\!\!\!\perp A$. (makes sense: we said earlier demographic parity is justified if this is true, because it enforces this property in the model. This result validates informal intuition.)
- Equalized odds seems best if $Y = Y'$ (makes sense: equalized odds justified in this case).
- Calibration seems like the worst out of these three criteria—it turns out it always allows disparity amplification (Theorem 6 in Yeom and Tschantz [2018]).
- Of course it is not always true that $Y' = Y$. And remember that $Y' \perp\!\!\!\perp A$ doesn't always make sense; for example, in the red car example $Y'$ is not independent of $A$ because of the correlation induced by latent variables. So this is helpful in selecting which criterion makes sense under which assumptions, but still doesn't provide a one-size-fits-all solution.

# Counterfactual Fairness

*Counterfactual fairness* [Kusner et al., 2017, Kilbertus et al., 2017] is satisfied for a model $\hat{f}(A^{(i)}, U^{(i)})$ if

$$\mathbb{P}(\hat{f}(a, U) = \hat{y} \mid X = x, A = a) = \mathbb{P}(\hat{f}(a', U) = \hat{y} \mid X = x, A = a),$$
$$\forall \hat{y} \in \{0,1\}^n, x \in \text{supp}(X), a, a' \in \{0,1\} \quad (5)$$

where $U$ is the set of all unobserved latent variables that may cause $X$, $A$, and $Y$.

*Counterfactual fairness* [Kusner et al., 2017, Kilbertus et al., 2017] is satisfied for a model $\hat{f}(A^{(i)}, U^{(i)})$ if

$$\mathbb{P}(\hat{f}(a, U) = \hat{y} \mid X = x, A = a) = \mathbb{P}(\hat{f}(a', U) = \hat{y} \mid X = x, A = a), \quad \forall \hat{y} \in \{0,1\}^n, x \in \text{supp}(X), a, a' \in \{0,1\} \quad (5)$$

where $U$ is the set of all unobserved latent variables that may cause $X$, $A$, and $Y$.

- Notice that conditional values of variables don't change. This is because we are interested in the probability conditional on the same variables after intervening on the variable of interest (distribution is different than the observational distribution when that variable is changed).
- Follows principle that it is unfair to treat individuals differently based on factors outside of their control (as protected attributes often are). Thinking about fairness based on causation puts individual control at center of question of fairness.
- Advantage: which criteria is best seems to depend on relationship between $A$, $Y$, and $Y'$. This approach relies on modeling these relationships explicitly, which should work better.
- Disadvantages: in order to verify that it is met (empirically or theoretically), need either (a) knowledge of or assumptions on the causal graph as well as the functions relating the variables, or (b) the ability to intervene on the variables of interest.

A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153—-163, 2017. URL `https://arxiv.org/pdf/1610.07524.pdf`.

N. Kilbertus, M. Rojas-Carulla, G. Parascandolo, M. Hardt, D. Janzing, B. Schölkopf, and B. M. De. Avoiding Discrimination through Causal Reasoning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. URL `https://arxiv.org/pdf/1706.02744.pdf`.

J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *The 8th Innovations in Theoretical Computer Science Conference*, 2017. URL `https://arxiv.org/pdf/1609.05807.pdf`.

M. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual Fairness. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, pages 4069—-4079. Curran Associates, Inc., 2017. URL `https://obamawhitehouse.archives.gov/blog/2016/05/04/big-risks-big-opportunities-intersection-big-data-`.

S. Yeom and M. C. Tschantz. Discriminative but Not Discriminatory: A Comparison of Fairness Definitions under Different Worldviews. Technical report, 2018. URL `http://arxiv.org/abs/1808.08619`.