Fairness in Machine Learning

Lecture 21

At first sight, we may suppose that AI algorithms are unbiased — especially compared to human decision makers — since the algorithm itself does not have any bias coded into it. However, AI algorithms are trained based on data that may include societal biases, leading to the algorithm learning these biases. In this lecture, we'll explore what it may mean for an AI algorithm to be *fair*, meaning unbiased.

An important figure for developing many of the theoretical foundations of fairness that we are about to see is Cynthia Dwork, a professor of Computer Science at Harvard. In the last 20 years, Professor Dwork has played a pivotal role in the formation of differential privacy and fair AI algorithms.

1 Case Studies

One example of such bias was documented by Amit Datta, Michael Carl Tschantz, and Anupam Datta in a 2015 paper. To test for gender bias in online advertisements, they created many profiles on Google, half with their gender set as man, half set to woman. For each of those accounts, while signed into it, they visited popular job search websites to make Google catch on to the account supposedly looking for a job. Then, they visited the Times of India homepage and recorded the ads shown to the accounts. Using this data, they trained a classifier to predict the gender of an account just from the ads shown to them and found that this classifier has high accuracy, indicating a different treatment of agents based on gender. The table below, taken from the paper, shows the top 5 ads that the classifier found to be the strongest indicator for an account being a man or a woman. Strikingly, the ad that is the strongest indicator for an account being male is an ad for jobs paying '$200k+,' being shown 1816 times to accounts set to male but only 311 times to accounts set to female.

Ad Title	Coefficient	appears in agents		total appearances	
		female	male	female	male
Top ads for identifying the simulated female group					
Jobs (Hiring Now)	0.34	6	3	45	8
4Runner Parts Service	0.281	6	2	36	5
Criminal Justice Program	0.247	5	1	29	1
Goodwill - Hiring	0.22	45	15	121	39
UMUC Cyber Training	0.199	19	17	38	30
Top ads for identifying agents in the simulated male group					
$200k+ Jobs - Execs Only	−0.704	60	402	311	1816
Find Next $200k+ Job	−0.262	2	11	7	36
Become a Youth Counselor	−0.253	0	45	0	310
CDL-A OTR Trucking Jobs	−0.149	0	1	0	8
Free Resume Templates	−0.149	3	1	8	10

Another example was documented in an article by ProPublica. Courts across the US use a software system called COMPAS to predict whether individuals convicted of a crime are likely to commit another crime in the future. Judges then use these predictions in their decision making, for example whether an individual is released on bail pending trial or not. In their report, ProPublica found that people who were similar in all parts of the record except for their race were given different outcomes, with African American defendants generally being predicted to be at a higher risk of committing future crimes than white defendants. Furthermore, as shown in the table below taken from the report, they found that white individuals were more likely to be predicted as unlikely to commit another crime but to end up committing another crime,

while African American individuals were more likely to be predicted as likely to commit another crime but not doing so.

Outcome	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%

These findings suggest a racial bias in the prediction algorithm. However, the picture is more nuanced, as we shall see.

2 Individual Fairness

We first consider what it means for an algorithm to be fair to each individual. Let's start by formalizing the setting at hand.

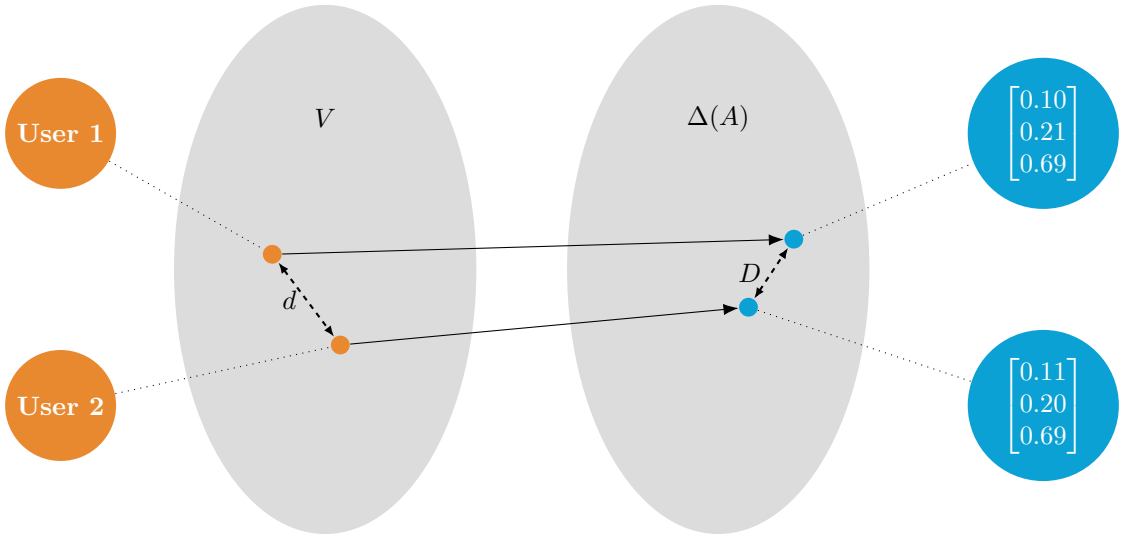
Definition 1 (Similarity-Based Fairness)**.** Let V be a set of *individuals* and let A be a set of possible *outcomes*. Let $M : V \to \Delta(A)$ be a *randomized classifier* that assigns each individual a probability distribution over outcomes.

Let $d : V \times V \to \mathbb{R}_{\geq 0}$ be a *metric*[1] over individuals V. Let D be a metric over possible outcome distributions $\Delta(A)$. The *Lipschitz property* holds for the classifier M if and only if for all $x, y \in V$,

$$D(M(x), M(y)) \leq d(x, y).$$

In other words, for any two similar individuals (as measured by proximity under d), also their outcomes are similar (as measured by proximity under D).

We illustrate this definition below: The two users on the left are embedded in some metric space where metric d captures their similarity (i.e., proximity). They both get matched to a distribution over outcomes, on the right. The Lipschitz condition prescribes that the distance between their outcomes as given by D should not be greater than the distance between the two users as given by d. This is the case in the illustration below.



[1] A metric d over a set V is a function that defines a distance $d(x, y)$ for any pair of elements $x, y \in V$. It is usually assumed that d satisfies

- that the distance of an element to itself is 0, so $d(x, x) = 0$ for any $x \in V$,
- symmetry, so $d(x, y) = d(y, x)$ for any $x, y \in V$, and
- the triangle inequality, so $d(x, z) \leq d(x, y) + d(y, z)$ for any $x, y, z \in V$.

As an example of a metric, consider Euclidean distance or Manhattan distance over vectors in $\mathbb{R}^d$.

There always exist Lipschitz classifiers, for example if we set $M(x) = M(y)$ for all $x, y \in V$, since if all individuals get the same outcome, the left-hand side of the Lipschitz property is always zero. This would likely be a useless classifier in practice since the prediction would be fully independent of the data presented to the algorithm. Thus, we want to also measure the quality of the prediction:

Definition 2 (Loss function)**.** A *loss function* $L : V \times A \to \mathbb{R}_{\geq 0}$ assigns to each individual x and outcome a a loss (i.e., a cost) $L(x, a)$ that is incurred from assigning outcome a to individual x.

This leads to an optimization problem: From the class of classifiers satisfying the Lipschitz property (and thus, arguably, being fair), find the one that minimizes the expected loss over all individuals in V. This optimization problem can be written as

$$\begin{aligned} \min \quad & \sum_{x \in V} \sum_{a \in A} \mu_x(a) \cdot L(x, a) \\ \text{s.t.} \quad & \forall x, y \in V : D(\mu_x, \mu_y) \leq d(x, y) \\ & \forall x \in V : \mu_x \in \Delta(A), \end{aligned}$$

where $\mu_x \in \Delta(A)$ is the distribution over outcomes that we assign to individual $x \in V$. While the above optimization problem may be of exponential size since V may be of exponential size[2], it can be solved efficiently in many practical cases.

So far, so good, but there are still two underspecified 'ingredients' in the approach above: What distance metrics D and d should we use? Let's first think about D. There are several reasonable options to measure the distance between two probability distributions, for example the following:

Definition 3 (Total Variation Distance)**.** Given two distributions $P, Q \in \Delta(A)$, their *total variation distance* is

$$D_{\text{tv}}(P, Q) = \frac{1}{2} \sum_{a \in A} |P(a) - Q(a)|.$$

A nice property of total variation distance is that if we choose $D = D_{\text{tv}}$, the optimization problem becomes an instance of linear programming[3] for a 'sufficiently linear' choice of d.

Unfortunately, coming up with a distance metric d over individuals is less straight-forward. For example, it is unclear which features of individuals we should even look at to calculate the distance; there are legal restrictions and ethical concerns for certain features. Furthermore, for other features, it is unclear how to meaningfully define 'proximity'. In practice, the choice of d will likely be instance-specific. One idea how to choose d in a general and principled manner is to query random people about how different or similar they perceive two individuals to be, and to computationally learn a distance metric from these samples.

To conclude our discussion of individual fairness, we briefly introduce a completely different approach to similarity-based fairness.

Definition 4 (Envy-Free Individual Fairness)**.** Assume that each individual $x \in V$ has utility $u_x(a)$ for each outcome $a \in A$. A randomized classifier is *envy free* if and only if for all $x, y \in V$,

$$\mathbb{E}_{a \in M(x)}[u_x(a)] \geq \mathbb{E}_{a \in M(y)}[u_x(a)].$$

In other words, no individual would rather have their outcome chosen from another individual's distribution than their own.

Just like for similarity-based fairness, envy-free classifiers always exist, for example, if $M(x) = M(y)$ for all $x, y \in V$. The definition of envy-freeness makes sense if the individuals have different utility functions and for some reason it is difficult for a classifier to give every individual their most preferred outcome. An

[2] For example, if each individual is characterized by m binary features, there may be 2^m possible 'individuals' in V that the algorithm may encounter; if the features aren't binary this number will be even larger or even infinite.

[3] We can introduce a variable $D(\mu_x, \mu_y)(a)$ for every $a \in A$ and $x, y \in V$. We then enforce that $D(\mu_x, \mu_y)(a) \geq \mu_x(a) - \mu_y(a)$ and $D(\mu_x, \mu_y)(a) \geq \mu_y(a) - \mu_x(a)$, as well as $1/2 \cdot \sum_{a \in A} D(\mu_x, \mu_y)(a) \leq d(x, y)$ for all $x, y \in V$. An assignment in the original optimization problem is feasible if and only if the same assignment with $D(\mu_x, \mu_y)(a) = |\mu_x(a) - \mu_y(a)|$ is feasible in the new linear program.

example of this is online ads: Due to limited information about the users, an algorithm may not be able to show a user the *best* ads for them, but we may hope that the algorithm at least shows no other user ads that would be better suited for this user than their current ads.

In contrast, the notion of envy-freeness is not useful in situations where the utility functions of the individuals are very similar and there are generally desired (but limited) and generally undesired outcomes. For example, consider algorithms deciding over bail or loans: All individuals would prefer to get a loan, and envy-freeness implies that everyone should get a loan with the same probability. But the algorithm can only give a loan to individuals that have a sufficiently high chance of paying it back.

3 Group Fairness

We now shift gears to think about algorithms being fair to groups in the population. We'll stick to the simplest setting of there being two groups and a binary outcome, since it turns out to already be quite difficult to get good definitions of fairness in this setting.

Definition 5 (Demographic Parity)**.** Consider a classifier $\hat{Y}$ that assigns each individual an outcome in $\{0, 1\}$. Each individual is either *qualified*, meaning they should be assigned value 1, or *unqualified*, meaning they should be assigned value 0; we denote this by Y. In some cases, we will think of 1 as a desirable outcome and 0 as an undesirable outcome.

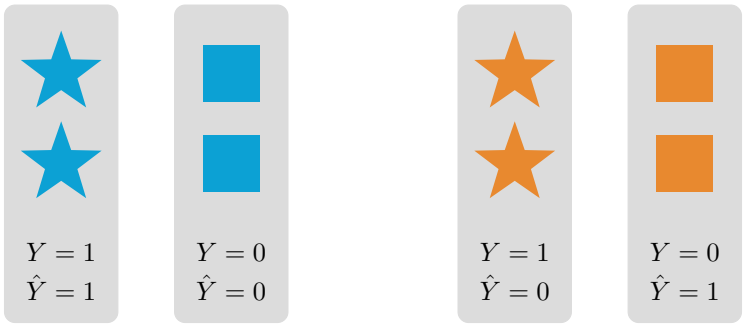
Assume that there are two groups, numbered 0 and 1, and that each individual is in either of the two groups. For an outcome $\hat{y} \in \{0, 1\}$ and group $g \in \{0, 1\}$, we write $\Pr[\hat{Y} = \hat{y} \mid G = g]$ to denote the probability that a (random) individual gets outcome $\hat{y}$, conditioned on the individual being from group g. A classifier satisfies *demographic parity* if

$$\Pr[\hat{Y} = 1 \mid G = 0] = \Pr[\hat{Y} = 1 \mid G = 1].$$

That is, individuals in both groups are equally likely to get outcome 1.[4]

Unfortunately, there exist examples where a classifier performs in intuitively unfair or biased ways but still satisfies demographic parity.

Example 1 (Demographic Parity of Bad Classifier)**.** Assume there are four individuals in each group, two qualified and two unqualified ones. In group 0, all individuals receive the correct outcome: $\hat{Y} = 1$ if $Y = 1$ and $\hat{Y} = 0$ if $Y = 0$. In contrast, in group 1, all individuals receive the wrong outcome: $\hat{Y} = 0$ if $Y = 1$ and $\hat{Y} = 1$ if $Y = 0$. This is illustrated in the figure below where groups are indicated by color; qualified individuals are stars while unqualified individuals are squares.



One can check that this classifier satisfies demographic parity since in each group, half the individuals receive outcome 1, so

$$\Pr[\hat{Y} = 1 \mid G = 0] = \frac{1}{2} = \Pr[\hat{Y} = 1 \mid G = 1].$$

However, intuitively this outcome feels unfair: All individuals in group 0 are classified correctly while all individuals in group 1 are classified incorrectly, so the accuracy of the classifier is very unevenly distributed.

[4] Since probabilities add to 1, this implies that they are also equally likely to get outcome 0.

Let's try to modify the definition of demographic parity to avoid cases as seen in Example 1.

Definition 6 (Equalized Odds)**.** For an outcome $\hat{y} \in \{0,1\}$, a true qualification $y \in \{0,1\}$, and group $g \in \{0,1\}$, let $\Pr[\hat{Y} = \hat{y} \mid G = g, Y = y]$ denote the probability that a (random) individual gets outcome $\hat{y}$ conditioned on the individual being from group g and having qualification y.

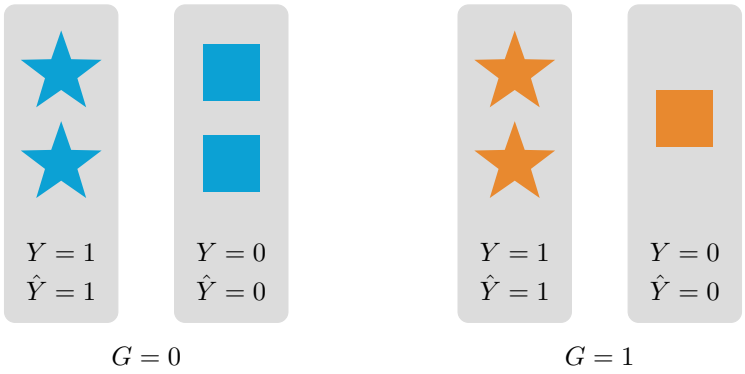
A classifier satisfies *equalized odds* if for all $y, \hat{y} \in \{0,1\}$, it holds that

$$\Pr[\hat{Y} = \hat{y} \mid G = 0, Y = y] = \Pr[\hat{Y} = \hat{y} \mid G = 1, Y = y].$$

That is, individuals in both groups have the same probability of getting 1 conditioned on being qualified and getting 0 conditioned on being unqualified. Furthermore, the false positive and false negative rates are the same in both groups.

We have seen in Example 1 that demographic parity does not imply equalized odds. Maybe surprisingly, equalized odds does also not imply demographic parity:

Example 2. Let's consider a similar setup as in Example 1 with two changes: Every individual gets labeled correctly, and there now only exists one unqualified individual in group 1. This is illustrated below:



One can check that this classifier satisfies equalized odds since

$$\Pr[\hat{Y} = \hat{y} \mid G = 0, Y = y] = 1 = \Pr[\hat{Y} = \hat{y} \mid G = 1, Y = y]$$

when $y = \hat{y}$ and

$$\Pr[\hat{Y} = \hat{y} \mid G = 0, Y = y] = 0 = \Pr[\hat{Y} = \hat{y} \mid G = 1, Y = y]$$

when $y \neq \hat{y}$. However, this classifier does not satisfy demographic parity since only half of the individuals in group 0 receive outcome 1 while $^2/_3$ of the individuals in group 1 do, so

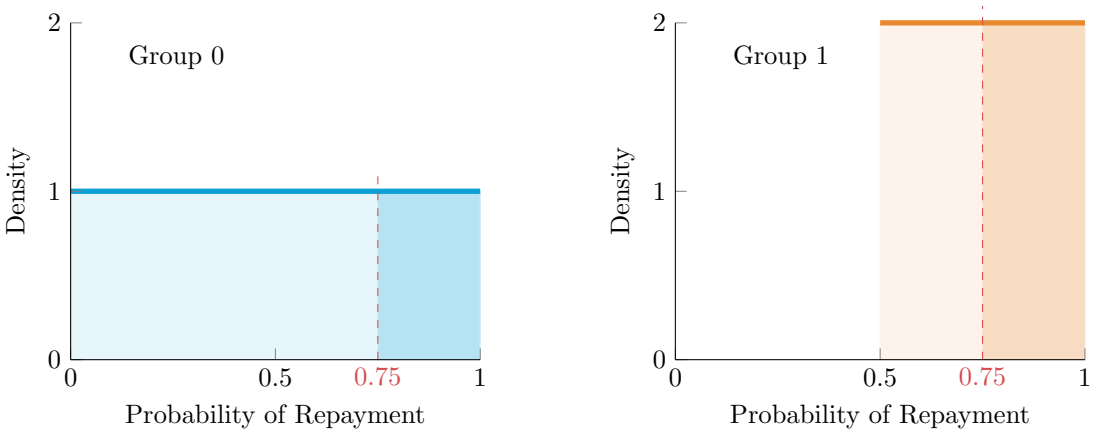
$$\Pr[\hat{Y} = 1 \mid G = 0] = \frac{1}{2} \neq \frac{2}{3} = \Pr[\hat{Y} = 1 \mid G = 1].$$

Intuitively, this outcome feels fair, since every individual gets classified correctly, but there happens to be more qualified individuals in one group. This may be seen as another argument against demographic parity.

To better understand equalized odds, let's consider another, intuitively fair, classifier and see if it satisfies equalized odds.

Example 3 (Risk Scores)**.** In the US, FICO scores are widely used to predict creditworthiness. The scores range from 300 to 850, with a low score indicating a low probability of paying back a loan, while a high score indicates a high probability of paying back a loan. Commonly, a score of 620, corresponding to a default rate of 18%, is used as the cutoff for prime-rate loans. Let's assume for simplicity that the FICO scores are perfectly *calibrated*, meaning that the default probability predicted by the score is the true probability with which an individual will default (independently of their group). Then, this is an intuitively fair classifier: Every individual that meets a fixed probability of repaying the loan gets a loan, independently of which group they are a part of. Does it also satisfy equalized odds?

To simplify the math, let's say that a bank gives a loan ($\hat{Y} = 1$) if and only if the estimated probability of repayment is at least 75%. Assume that there are two groups. In group 0, there exist individuals with any probability of repayment; in particular, the probability of a random person in group 0 repaying the loan is uniformly distributed on $[0, 1]$. In group 1, all individuals have a probability of repayment of at least $1/2$; in particular, the probability of a random person in group 1 repaying the loan is uniformly distributed on $[1/2, 1]$. These probability densities are shown below.



For group 0, we can determine that the probability of a false negative (not giving the loan to a person that would have repaid it) is

$$\Pr[\hat{Y} = 0 \mid G = 0, Y = 1] = \frac{\Pr[\hat{Y} = 0 \wedge Y = 1 \mid G = 0]}{\Pr[Y = 1 \mid G = 0]} = \frac{\int_0^{0.75} x \; dx}{\int_0^1 x \; dx} = \frac{0.75 \cdot 0.375}{1 \cdot 0.5} \approx 0.56.$$

We compute the probability in the numerator by averaging over all individuals that didn't get a loan ($\hat{Y} = 0$) since their probability x of repaying is less than 0.75. For those individuals, we average over the probability that they would have actually repaid the loan ($Y = 1$) which happens with probability x.

Similarly, for group 1, we can determine that the probability of a false negative is

$$\Pr[\hat{Y} = 0 \mid G = 1, Y = 1] = \frac{\Pr[\hat{Y} = 0 \wedge Y = 1 \mid G = 1]}{\Pr[Y = 1 \mid G = 1]} = \frac{\int_{0.5}^{0.75} 2x \; dx}{\int_{0.5}^1 2x \; dx} = \frac{2 \cdot 0.25 \cdot 0.625}{2 \cdot 0.5 \cdot 0.75} \approx 0.41.$$

This classifier does not satisfy equalized odds! Individuals in group 0 are more likely to be denied a loan even though they would have paid it back.

It turns out that such an intuitively fair loan policy of picking a fixed percent threshold satisfies equalized odds only in very particular settings.

Definition 7 (Equal Base Rates)**.** We say that an instance satisfies *equal base rates* if $\Pr[Y = 1 \mid G = 0] = \Pr[Y = 1 \mid G = 1]$. That is, the mean probability of being qualified is equal in both groups.

Example 4 (Equal Base Rates)**.** Consider the distributions over individuals' probabilities of repayment shown below.

The mean probability of an individual being qualified is the same in both groups, since

$$\Pr[Y = 1 \mid G = 0] = \int_0^1 x \; dx = \frac{1}{2} = \int_{0.25}^{0.75} 2x \; dx = \Pr[Y = 1 \mid G = 1].$$

Thus, this instance satisfies equal base rates.

Definition 8 (Perfect Prediction)**.** We say that an instance satisfies *perfect prediction* if for each individual, the probability of being qualified is 0 or 1, i.e., each individual is either certainly qualified or certainly unqualified.

Example 5 (Perfect Prediction)**.** Consider the distributions over individuals' probabilities of repayment shown below, where the marks represent point masses with probability as shown on the y-axis.

In group 0, every individual is qualified with probability 1. In group 1, half the individuals are qualified with probability 0 while the other half is qualified with probability 1. Thus, this instance satisfies perfect prediction.

Theorem 1 (Informal)**.** *If a risk assignment*[5] *satisfies calibration*[6] *and equalized odds, the underlying instance either satisfies equal base rates or perfect prediction.*

To summarize, even if the predictor is perfectly unbiased, a classifier based on risk threshold would fail equalized odds in any realistic setting. Unfortunately, this is an obstacle to our hope of using equalized odds as a general measure of group fairness.

Over the last decade, dozens of fairness axioms have been proposed, yet the community has not converged on a single notion that can be regarded as the "right" one. This stands in contrast to concepts such as differential privacy or envy-freeness in fair division, where there is broad agreement that the definitions capture a compelling and operationally meaningful ideal. The absence of such consensus in algorithmic fairness has practical consequences: it creates ambiguity for practitioners and slows the deployment of fairness-aware systems. Moving toward impactful applications will likely require narrowing this gap.

[5] A *risk assignment* is any classifier that gives the (desired) outcome $\hat{Y} = 1$ exactly to those individuals for which the probability that they are qualified ($Y = 1$) exceeds a fixed threshold t.

[6] A classifier is *calibrated* if the predicted probability with which an individual is qualified is correct for all individuals.