Six Ways of Fairness

Thore Husfeldt

In Gunnemyr, Mattias & Jönsson, Martin L. (2023) *Post Hoc Interventions: Prospects and Problems.* Lund: Department of Philosophy, Lund University. https://doi.org/10.37852/oblu.184

ISBN: 978-91-89415-60-7 (print)
978-91-89415-61-4 (digital – pdf)
978-91-89415-62-1 (digital – html)

DOI: https://doi.org/10.37852/oblu.184.c509

Published by the Department of Philosophy, Lund University.
Edited by: Mattias Gunnemyr and Martin L. Jönsson
Cover layout by Cecilia von Arnold, Pufendorf Institute for Advanced Studies

9

Six Ways of Fairness

Thore Husfeldt[1]

Abstract. Fairness interventions at any stage of a decision process, including post hoc, necessarily reify a moral intuition about which out- comes are viewed as "fair." Different moral intuitions formally contradict each other, and many suggestions for algorithmic, automated, or trans- parent fairness interventions are necessarily formal. I give a very simple, but complete, overview of such formal fairness notions and observe and solidify some basic contradictions between widely-held intuitions. The presentation aims to be interesting, accessible, minimal, precise, and dispassionate.

1. Introduction

This presentation aims to be an introduction to formalisations of fairness notions that is interesting, accessible, minimal, precise, and dispassionate. In particular:

Interesting. I want to explain some of the core insights, in particular about trade-offs and conflicts between widely-held fairness intuitions.

Accessible. I try to not rely on prior exposure to the technical parts, including machine learning terminology, probability theory, and causality. As best as I can, I either avoid such concepts or define them from first principles.

[1] Professor of Computer Science, Department of Computer Science, Lund University, Sweden, and IT University of Copenhagen, Denmark.

Minimal. I want to introduce as *few* concepts as possible, while still being able to present the phenomena that I find interesting, puzzling, or appealing.

Precise. All concepts are meticulously defined and arguments presented carefully and in their entirety. To the extent that it makes sense, concepts and arguments are supported by diagrams. I've spent some time worrying about appealing notation, favouring an imagined reader that is new to this area and holds no established preferences. There are no incomplete proofs, either of the form 'it can be seen by standard arguments that' or 'the proof follows from chapter 7 in Feller (1950).'

Dispassionate. I aim to be agnostic about political ideals and assume that you and I share no ideological intuitions. Rich examples are important scaffolds for navigating abstraction, but I try to stick to a running *toy* example that aims to avoid triggering our tribal instincts.

This text is not, not does it want to be

Novel. Nothing here is new, except maybe the framework and some work on identifying necessary conditions in our definitions for various relationships to hold. This entire text is an attempt to explain existing notions and findings to myself in a way that I would have liked them explained to me.

Comprehensive. I know many more fairness notions than six, but the whole idea of this text is to be minimal rather than comprehensive. Much more complete presentations can be found in Verma and Rubin (2018) and Barocas et al. (2019) and the references therein.

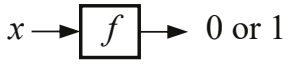
Reflective. Precise definitions, rigorous analysis, and contextual decoupling are *in themselves* epistemologically nontrivial choices, as is my focus on the trade- offs and tensions between various socially adaptive ideas. A lot can be said about this, and I don't. An accessible introduction to this discussion can be found in Friedler et al. (2021).

I also meticulously avoid *resolving* the dilemmas that result from observing the contradictions between various fairness notions. For an example of how such dilemmas may be approached, see Lippert-Rasmussen (2023).

2. Setup

2.1 Selection

Think of S as any decision-making procedure, such as an algorithm, a method, or a law. It takes an individual x and produces an outcome, either 0 or 1:

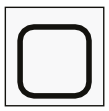


When $S(x) = 1$ we will say that "x has been selected." You can think of selection as "gets their loan application approved," "goes to jail," "is admitted to university," "gets the job," "is shown the ad," "receives the medical treatment," etc. Note that being selected can be beneficial or detrimental for x, depending on the context; mnemonically, I suggest thinking of x being selected for a *scholarship* or a *security check* when $S(x) = 1$.

Formally, S is a mapping

$$S : x \to \{0, 1\}$$

but we will often just write $S = 1$ instead of $S(x) = 1$. We draw the population classified as $S = 1$ using a thick black outline. This may encompass some, or even all of the population.



2.2 Target

The *target* is the quality we try to select for. Think of $T = 1$ as "repays their loan," "commits another crime," "is highly intelligent," "is a pleasant and competent colleague," "will buy the advertised product," "benefits from the treatment," etc. The target value can be a desirable or undesirable quality of x; mnemonically, think of $T = 1$ as talent or *terrorist*. Both are compatible with the corresponding mnemonic for selection: We may want to select talents for the scholarship, and to select terrorists for the security check. In some contexts, you can think of T as *truth*.

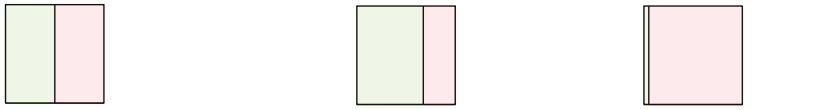
In pictures, the population where $T = 1$ (the *target population*) is drawn in a more opaque colour. The mnemonic is *tinted*. We can draw examples where the target population makes up various fractions of the total population:

very few almost half many

2.3 Groups

The population is partitioned into two groups 0 and 1. If x belongs to group 1 then $G = 1$, otherwise $G = 0$. This grouping may be by sex, gender, religion, ethnicity, caste, age, etc.[2] Think of $G = 1$ as *Greeks* in a (fictional) ancient population consisting entirely of Greeks and Romans. If the two subpopulations for which $G = 1$ or $G = 0$ have roughly equal size, you may want to think of G as gender.

We will draw group 1 in green, and the other group using not-green (in fact, red).[3] If you like the Graeco–Roman example, Greeks are green and Romans are red.



equal sized groups unequal sized tiny minority

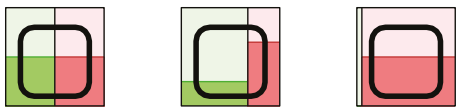
Depending on context, the group membership of x may be called "sensitive" or "protected," leading to implicit or explicit legal, social, or cultural ambitions for the interplay between S, T and G.

[2] Class is another plausible grouping. We avoid the word "class" here so as to avoid confusion with the classification provided S, which is often called a classifier.

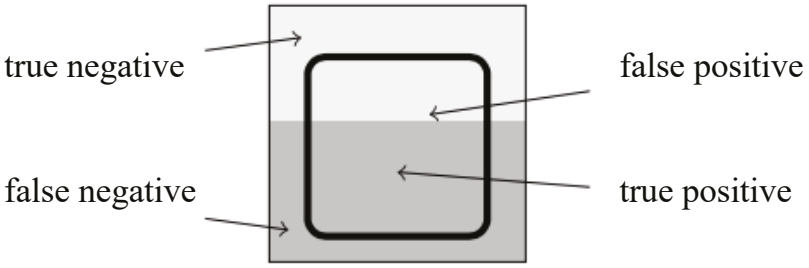
[3] If you cannot distinguish the two colours, green will be on the left (*gauche* in French), and red on the right.

2.4 What we want to study

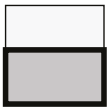
The three values S, T, and G provide a framework for studying the result of f, and we can draw them using schematic representations like this:



Example 1 (*S and T: accuracy, correctness, utility*). The interplay between the selection S and target T models notions like accuracy or correctness. We can express some standard terminology: The true positives are targeted individuals ($T = 1$) that are (correctly) selected ($S = 1$). Similarly, true negatives have $T = 0$ and are (correctly) de-selected ($S = 0$). False positives have $T = 0$ yet are selected ($S = 1$), and the false negatives have $T = 1$ and are de-selected ($S = 0$). Graphically:



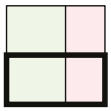
The closer S and T are, the more accurate or correct is the classifier. When $S = T$ for all inputs then the classifier is perfect, in the sense of making no mistakes:



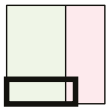
perfect classification, S = T.

We say that a classifier has high utility if $S = T$. Note that "perfection," "utility," "accuracy," and "correctness" are value-laden words. Note also that it is not clear for whom a perfect classifier has high utility; the outcomes for the individual, the group, or society are often at variance with each other. For instance, failing to select terrorist x for security screening is the desired outcome for x, but catastrophic for others.

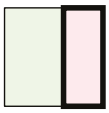
Example 2 *(S and G: representation and bias*). The classification given by *S* can relate to the grouping given by *G* in various ways. If one group is selected with more than their fraction of the population, that group is called overrepresented among the selected individuals, which is sometimes called bias.



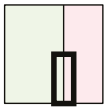
half of population, no bias, more G



small selection, biased towards G



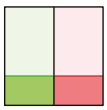
perfectly partisan selection, S = 1 − G



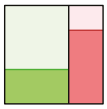
equal numbers from each group, G underrepresented

Our usual caveat applies: bias is a value-laden word that has negative connotations.

Example 3 (*G and T: diversity*). Finally, group membership *G* may relate with the target value *T*. The target quality may be rare or ubiquitous, and it may be equally or unequally represented in the two groups. Whether targeted individuals belong to either group depends on the relative group sizes.



target equally rare in both groups



target rarer among G

the majority of targeted individuals belong to G.

In the last example, note that all individuals in the red group are in the target population. Be aware that in many contexts the mere idea that target values are not equally distributed among groups is outrageous.

Example 4 *(Homerian Poetry School).* I will stick with the Graeco–Roman setting as a sufficiently silly and culturally remote toy example. The selection is a scholarship to the Athenian School for Homerian Poetry and the targeted value is talent (for Homerian poetry). The Greek population is tiny compared to the vast Roman empire; yet talent for Homerian poetry is much more widespread among the Greeks. (This may have entirely cultural reasons; Homerian poetry is written in Greek, not Latin, and highly valued in the Greek elite.)

If you're a formalist and happily navigate *S, T,* and *G* as mathematical abstractions, you can ignore my attempts at building intuition.

3. Fairness as Independence

We were able to express a few things using equality, such as $S = T$, but for the fairness notions we need *independence* from probability theory.

This will allow us to write expressions like "$S \perp G$" for "S is independent of G". The intended meaning is that S, which determines whether x is selected, is "independent of" (in the sense of "is not affected by" or "is indifferent to" or "contains no information about") G, the group that x belongs to.

If you want, you can largely ignore the fact that $\perp$ is a shorthand for a very rigorous and simple definition of "is independent of" and skip the next subsection. You can do the same if you do not need or want to be reminded of basic probability theory.

3.1 Event, Condition, Independence, Random Variable

For our purposes, an *event* E is a subset of the set Ω, with an associated *probability* $\Pr(E)$ satisfying $0 \leq \Pr(E) \leq 1$, $\Pr(\Omega) = 1$, and $\Pr(E \cup F) = \Pr(E) + \Pr(F)$ when E and F are disjoint.

Example 5. If you want, you can view Ω as 'the population,' so that events are subsets of the population, such as 'incompetent Romans falsely given a scholarship.' Then the population is finite and you can understand the probability function as $\Pr(E) = |E|/|\Omega|$.

The *conditional probability of E given F* written $E \perp F$, is the probability that E occurred given that F has occurred, and defined as

$$\Pr(E \mid F) = \frac{\Pr(E \cap F)}{\Pr(F)} \qquad \text{if } \Pr(F) > 0$$

(If $\Pr(F) = 0$ then $\Pr(E \mid F)$ is not defined.) Note that

$$\Pr(E \cap F) = \Pr(E \mid F) \Pr(F)$$

always holds, even if $\Pr(F) = 0$ (in which case both sides are 0).

Intuitively, an event E is independent of another event F, written $E \perp F$, if the fact that E happened includes no information about whether F happened. Formally, two events are *independent* if $\Pr(E \cap F) = \Pr(E) \Pr(F)$.

Proposition 1. The following are equivalent to $E \perp F$:

1. $F \perp E$.
2. $E \perp \overline{F}$.
3. $\Pr(E \mid F) = \Pr(E)$ if $0 < \Pr(F)$.
4. $\Pr(E \mid F) = \Pr(E \mid \overline{F})$ if $0 < \Pr(F) < 1$.

Proof. Set intersection and multiplication are both symmetric. For 3, we have

$$\Pr(E \mid F) = \frac{\Pr(E \cap F)}{\Pr(F)} = \frac{\Pr(E) \Pr(F)}{\Pr(F)} = \Pr(E)$$

For 2, observe

$$\Pr(E) \Pr(\overline{F}) = \Pr(E) \Pr(1 - \Pr(F)) = \Pr(E) - \Pr(E) \Pr(F) = \Pr(E) - \Pr(E \cap F) = \Pr(E \cap \overline{F}).$$ □

Whereas equality and independence are symmetric concepts, it is not true in general that $\Pr(E \mid F)$ is the same as $\Pr(F \mid E)$ (even if $E \perp F$).

Example 6. Alice and Bob each have their own (biased) coin. A is the event that Alice's coin comes up "heads," with probability $\Pr(A) = \frac{1}{10}$, Bob's with $\Pr(B) = \frac{1}{4}$. By tedious enumeration of the 400 different outcomes, we see that $\Pr(A \cap B) = \frac{10}{400} = \frac{1}{40}$. Then $A \perp B$.

Example 7. Claire has two coins. Coin 1 comes up "heads" with probability $\frac{1}{10}$, coin 2 with probability $\frac{1}{4}$. Claire picks one of her coins uniformly at random and lets both Alice and Bob toss it. The probability that Alice comes up "heads", is

$$\Pr(A) = \frac{1}{10}\ \frac{1}{2} + \frac{1}{4}\ \frac{1}{2} = \frac{7}{40}$$

Bob tosses the same coin, so $\Pr(B) = \frac{7}{40}$. However,

$$\Pr(A \cap B) = (\tfrac{1}{10})^2\ \tfrac{1}{2} + (\tfrac{1}{4})^2\ \tfrac{1}{2} = \tfrac{29}{800}$$

Thus, A and B are not independent.

To build some intuition, $\Pr(B \mid A) = \frac{29}{140}$, so having observed Alice's heads coin toss, Bob has a roughly 20% chance (much better than $\Pr(B) = \frac{7}{40}$) of a heads outcome. Intuitively, this is because it is quite likely that Alice received the 2nd coin from Claire.

Proposition 2 (Total probability)**.** Let $E_1, \ldots, E_n$ be a disjoint partition of Ω. Then for any event F, we have

$$\Pr(F) = \Pr(F \mid E_1) \Pr(E_1) + \cdots + \Pr(F \mid E_n) \Pr(E_n). \qquad (1)$$

Proof. Since the E_i for a partition of Ω and $F \subseteq \Omega$, we can write F as a disjoint union $F = (F \cap E_1) \cup \cdots \cup (F \cap E_n)$, which implies $\Pr(F) = \Pr(F \cap E_1) + \cdots + \Pr(F \cap E_n)$. By definition, $\Pr(F \cap E_i) = \Pr(F \mid E_i) \Pr(E_n)$. □

In particular, for $n = 2$, we have

$$\Pr(F) = \Pr(F \mid E) \Pr(E) + \Pr(F \mid \overline{E}) \Pr(\overline{E}),$$

which is the only version we need.

A *random indicator variable* A is a function $A : \Omega \to \{0, 1\}$. For a value a we write the (formally meaningless and abusive, but intuitively useful) expression "$A = a$" to denote the event $\{ x \in \Omega \mid A(x) = a \}$. Two random variables A and B are *independent* if for all a, b

$$\Pr(A = a \cap B = b) = \Pr(A = a) \cdot \Pr(B = b)$$

We extend the notation from events to random variables and write $A \perp B$. Note that if $A \perp B$ and $\Pr(B = b) \neq 0$ then

$$\Pr(A = a \mid B = b) = \frac{\Pr(A = a \cap B = b)}{\Pr(B = b)} = \Pr(A = a)$$

Two random variables are equal, $A = B$, if $A(x) = B(x)$ for all $x \in \Omega$. We write $A \not\perp B$ and $A \neq B$ do indicate that $A \perp B$ and $A = B$ fail to hold, respectively.

> **Proposition 3.** Assume $0 < \Pr(A = a) < 1$ or $0 < \Pr(B = b) < 1$ *for some a or b*. If $A = B$ then $A \not\perp B$. If $A \perp B$ then $A \neq B$.
>
> *Proof.* Assume $\Pr(B) < 1$; the other case is symmetric. If $A = B$ then $\Pr(A = a \cap B = a) = \Pr(A = a \cap A = a) = \Pr(A = a) \neq \Pr(A = a) \Pr(B = b)$, so A and B are not independent. Now assume $A \perp B$. If also $A = B$ then in particular $(\Pr(B = b))^2 = \Pr(A = b \cap B = b) = \Pr(B = b \cap B = b) = \Pr(B)$. But this can only hold if $\Pr(B = b) \in \{0, 1\}$, violating the assumption. We conclude $A \neq B$.

To avoid a misunderstanding: $A \neq B$ does not imply $A \perp B$, and $A \not\perp B$ does not imply $A \neq B$. Independence and equality are both very restrictive notions, and the relationship between A and B can fail to satisfy either.

> **Example 8.** Alice flips a fair coin, and Bob flips the same coin if it comes up 1 (else he accepts Alice 's outcome as his own). Then $\Pr(A = 1) = \frac{1}{2}$ and $\Pr(B = 1) = \Pr(A = 1) + \Pr(A = 0)\frac{1}{2} = \frac{3}{4}$. Clearly, $A \neq B$. (In fact, $\Pr(A = 1) \neq \Pr(B = 1)$.) Also, A and B are clearly not independent, and we can verify $\Pr(A = 1 \cap B = 0) = 0 \neq \frac{1}{2} \cdot \frac{1}{4} = \Pr(A = 1) \cdot \Pr(B = 0)$.

To avoid other misunderstandings: $A \perp B$ does not imply $\Pr(A = 1) \neq \Pr(B = 1)$. (Consider two independent coin flips.) $A \neq B$ does not imply $\Pr(A = 1) \neq \Pr(B = 1)$. (Let A be a random coin flip and define $B = 1 - A$.)

3.2 Demographic Parity

The relationship

$$S \perp G \tag{2}$$

means that selection is independent of group membership. In particular, group membership does not affect the classifier's selection outcome.

This is a very well-studied fairness notion and goes by many names: demographic parity, group fairness, statistical parity, equal outcomes, absence of disparate impact, Darlington's 4th criterion, equity, or just independence.

> **Example 9** (*Proportional representation*). Under (2), we have $\Pr(G = g \mid S = s) = \Pr(G = g)$. For instance, for $g = s = 1$, this means that $\Pr(G = 1 \mid S = 1)$ (the proportion of Greeks among those selected for a scholarship) equals $\Pr(G = 1)$ (the proportion of Greeks in the entire population). In words, Greeks (as well as non-Greeks) are represented among the selected (as well as among the de-selected) in proportion to their population size.

Perhaps misleadingly, the notion is often understood as the selected group *representing* the whole population. (This is misleading because selected in individuals may have very little else in common with their group.)

3.3 Target indifference

The relationship

$$S \perp T$$

means that the classifier selects individuals independently of their target value. Thus, $S \perp T$ means that selection is *indifferent* to the target value. In contrast, $S = T$ means that exactly the target value is selected for. The latter is sometimes called maximal *utility* and is often a desirable property of selection.

The choice between $S \perp T$ and $S = T$ reflects the importance of accurate selection.

> **Example 10** (*Sortition*)**.** This corresponds to flipping a coin (or some other random process) for each individual, weighted by the desired size of the selected set.

Many real-world societies have used or still use a random process for civic obligations such as jury duty in the United States. The very fact that the process achieves demographic parity (across all thinkable groups, not only G) is felt to outweigh the potential absence of any legal expertise or ethical schooling among jury members. In our notation, the benefits of achieving demographic parity or representativeness (in particular, $S \perp G$) are felt to outweigh the negative consequences of not selecting for competence ($S \perp T$). See Sec. 4.4.

In a raffle or lottery, individuals are selected based on a random process. For instance, if we want to select $k = |S|$ many individuals from the population $\mathscr{P}$, we can hand out numbered tickets numbered 1, . . ., $|\mathscr{P}|$ randomly and select those individuals receiving a number at most k. Lotteries can be used for entertainment (and the perceived fairness of the process is important for attracting customers that want to buy a lottery ticket), for selecting school children for an exciting activity such as a school trip, or for unpleasant activities such as latrine duty. See Stone (2009) for an introduction to random selection.

3.4 We're all equal

Consider

$$G \perp T\,.$$

In words, the target variable is independent of group membership. Sometimes called "equal base rates" or "the world is just." It represents a model of reality underlying many social theories, religions, and scholarly disciplines. When $G \perp T$ holds, a perfect classifier with $S = T$ satisfies demographic equality $G \perp S$. Most of the phenomena that make fairness definitions interesting simply vanish under this assumption.

3.5 Relationships

We have arrived at three different fairness notions,

$$G \perp S, \qquad G \perp T, \qquad S \perp T\,,$$

making up half of our "sixpack" of fairness. To set the stage for next two sections, we want to understand the interaction of these notions.

Luckily, there isn't much to understand. For instance, all three notions can hold simultaneously. Imagine for instance a situation in which we're all equal (so $G \perp T$ holds by assumption about the target distribution, such as letting T be the last digit of an individual's number of nose hairs in binary) and let S be the outcome of a fair coin flip (so $S \perp T$ and $S \perp G$). Then all three fairness notions are simultaneously satisfied.

We can also imagine $G = T$ (so the target is "membership in G") and still use a fair coin flip for S, so that $S \perp T$ and $S \perp G$. Now two notions are simultaneously satisfied and the third is maximally unsatisfied. The most attractive of these settings is where $S = T$ (perfect prediction), yet $S \perp G$ (equal outcomes) and $T \perp G$ (we're all equal.)

It is also thinkable that only one of the fairness conditions is satisfied. However, the two others cannot both be maximally unsatisfied: if both $S = T$ and $G = T$ then we cannot have $S \not\perp G$ (in fact, we do have $S = G$.) Finally, all three conditions can of course fail to hold. (In fact, in reality they presumably *do* fail to hold. The entire framework is a simplification that tries to conceptualise desirable properties.)

Even though I try to be almost comically agnostic and symmetric about the different notions, you may want to view the conditions Target Indifference $S \perp T$ and We're All Equal $G \perp T$ as *trivialising* conditions, at least on first reading, because of the following two examples.

> **Example 11** (*Sortition*). Let S be the result of a random process, such as a lottery. Then $S \perp T$ and $S \perp G$ hold. Thus, target indifference and demographic parity are very easy to achieve.

> **Example 12** (*Perfect world*)**.** Assume We're All Equal $T \perp G$. Now assume that the selection mechanism achieves perfect utility $S = T$. Then Demographic Parity $S \perp G$ holds. In other words, Demographic Parity is achieved by merely maximising the utility of the selection mechanism, so that the concepts of utility and fairness are identical. No conflicting goals arise, and the meritocratic intuition is well-aligned with the equity intuition, and the fairness perspective has added nothing new.

In other words, even though $S \perp T$ and $G \perp T$ may be very attractive fairness notions, keep in mind that our explorations become interesting mainly in settings where they fail. They are in some sense perfect, trivial, irrelevant, unrealistic, degenerate, boring, utopian, or even dystopian.

4. Fairness as Conditional Independence

Our central tool for modelling causality and fairness is the notion of conditional independence (Dawid, 1979; Pearl, 2009).

4.1 Conditional Independence

Let A, B, C be random variables. We say that A and B are *conditionally independent given* C, written

$$A \underset{C}{\perp} B \qquad \text{or} \qquad A \perp B \mid C \qquad \text{or} \qquad (A \perp B) \mid C,$$

if for all $a, b, c \in \{0, 1\}$,

$$\Pr(A = a \cap B = b \mid C = c) = \Pr(A = a \mid C = c) \cdot \Pr(B = b \mid C = c).$$

By $E \cap F \mid G$ we mean $(E \cap F) \mid G$.

Three conditional independence notions can be expressed:

$$S \underset{G}{\perp} T, \qquad G \underset{S}{\perp} T, \qquad \text{and} \qquad G \underset{T}{\perp} S.$$

Because of symmetry, these are *all* the ways in which our three variables can be conditionally independent.

4.2 Equal Odds

The relation

$$G \underset{T}{\perp} S$$

is called equal odds, equal treatment, conditional procedure accuracy equality, or separation.

It is easily understood in terms of *errors*: If you insist that no group is 'treated worse' (or better) than the other, then you are for equal odds. In particular, by 'treated equally' you mean the false positive rate should be the same for both groups, and the false negative rate should be the same for both groups. (By implication, the true positive rates and true negative rates are also the same.)

For instance, you achieve equal odds if you admit half of the targeted population in each group (say, half the talented Romans and half the talented Greeks), and one quarter of the untargeted population (say, a quarter of the untalented Romans and a quarter of the untalented Greeks.) From the perspective of a talented Roman, her odds of being selected are the same as a talented Greek (namely, $\frac{1}{2}$).

In other words, if we restrict our attention to only the individuals with the same T, we achieve demographic parity $G \perp S$. Yet another way of saying this is that any dependence between group membership G and selection S (i.e., deviation from demographic parity) is 'explained away' by the target distribution T.

4.3 Equally Good Prediction

The relationship

$$G \underset{S}{\perp} T$$

is also known as the Cleary model (absence of differential prediction), *sufficiency*, predictive rate parity, conditional use accuracy equality, or well-calibration within groups.

This notion is easier to understand by first looking at the relaxed version, for $S = 1$. The idea is that $G \perp T$ ("we're all equal"), when we restrict the population to the selected individuals. The selection may be heavily skewed towards one group or the other, and talent may be very unequally distributed in the population. The corresponding requirement for $S = 0$ is that the classifier *deselects* individuals *outside* of the target group with equal probability. If $G \perp T$ holds conditioned on both $S = 1$ and $S = 0$ then we have $G \perp T \mid S$.

> **Example 13.** The Homerian Poetry School of Athens admits students on a very harsh entrance exam. Greeks are much better at poetry (in particular in Greek!), so the cohort of freshmen is dominated by them. However, the (pitifully few) Romans who make it into the Homerian are every bit as talented as their classmates from across the Aegis. Students on campus can detect no group differences in performance. In fact, long-time teachers at the school, who seldom wander off-campus and only ever interact with the selected group, have the (false) impression that Greeks and Romans in general are equally good at Homerian poetry, and will lecture their worldlier friends at length about this.

Graduates from the Homerian are in high demand in the booming poetry economy, no matter their group membership.

This is an example of $G \perp T \mid S = 1$, sometimes called positive prediction parity or just predictive parity. The story says nothing about $S = 0$. For instance, in the story it is still possible that rejected Roman applicants on average are much better poets than rejected Greek applicants.

4.4 Stratified indifference

The cleanest example is sex-based draft lottery, used in many countries that select a random subset of the male population for military service, and none from the female. Also, the ancient Greeks used stratified sortition by implementing aleatoric democracy yet restricting it to males.

5. Relationships, Implications, and Trade-Offs

Three of the six fairness notions, stand out as being popular, intuitively appealing, politically viable, consistent with correctness, and achievable by manipulating the classifier:

$$\begin{array}{ccc} G \perp S & G \perp T & S \perp T \\ G \underset{T}{\perp} S & G \underset{S}{\perp} T & S \underset{G}{\perp} T \end{array} \tag{3}$$

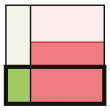
Mimicking our easy observations from 3.5, we will investigate the formal relationship between these notions. We saw that from the top row, it was possible to satisfy 1, 2, or 3 of the notions. The gist of this section is that this is not true of the second row.

This insight turns out to be a relatively pedestrian observation about the properties of conditionally independent random variables – it has nothing to do with which three fairness notions we picked. Thus, we will give a general and very simple treatment in terms of A, B and C, and spell out the implications to the popular case after each result.

5.1 Independence versus Conditional Independence

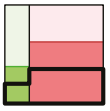
First we convince ourselves that the second row is indeed different from the first, in that the independence notion $A \perp B$ does not imply, nor is implied by, its conditional counterpart $A \perp B \mid C$. This is entirely pedestrian but seems to be psychologically counter-intuitive, so here are some counterexamples for $A = S$, $B = G$, $C = T$:

1. Demographic parity holds: Both groups are proportionally represented in the selection, roughly with 1/3 of their populations. Equal odds does not hold: targeted Greeks are certain to be selected, Romans aren't.



$S \underset{T}{\perp} G$ yet $S \not\perp G$

2. Equal odds holds: In each group, roughly half of the targeted individuals are selected, and none of the untargeted. But demographic parity fails: Romans are overrepresented in the selection – almost 1/3 of the Roman population is selected, but less than 1/6 of the Greek.



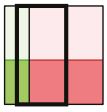
$S \not\underset{T}{\perp} G$ yet $S \perp G$;

3. If $T \perp G$ (we're all equal) then both can hold.

both $S \underset{T}{\perp} G$ and $S \perp G$;

In fact, both would hold if we set $S = T$, simultaneously achieving perfect prediction, equal odds and demographic parity.

4. If $S \perp T$ (target indifference), then both can hold as well:



both $S \underset{T}{\perp} G$ and $S \perp G$;

Since the two notions do not imply the other, we need assume both, requiring both demographic parity and equal odds, i.e., $G \perp S$ and $G \perp S$ mod T . This combination of two fairness notions is very close to the moral intuition of many

people and will be felt as a desirable requirement in the selection process. We already saw two simple examples above (3 and 4) for how this could be achieved (namely, assuming we're all equal $S \perp G$ or target indifference $S \perp T$.)

We now show that those two are the *only* possibilities.

Proposition 4. Let A, B, C denote random variables with $C \in \{0, 1\}$. If $A \perp B$ and $A \perp B \mid C$ then $A \perp C$ or $B \perp C$.

Proof. Write a for the event $A = a$, and similarly for b and c. By $\Pr(\bar{c})$ we mean $\Pr(\overline{C = c}) = \Pr(C = 1 - c) = 1 - \Pr(c)$.

By total probability we have

$$\begin{aligned}
&\Pr(a) = \Pr(a \mid c)\ \Pr(c) + \Pr(a \mid \bar{c})\ \Pr(\bar{c}) = \\
&\Pr(a \mid c)\ \Pr(c) + \Pr(a \mid \bar{c})\ [1 - \Pr(c)] = \\
&[\Pr(a \mid c) - \Pr(a \mid \bar{c})]\ \Pr(c) + \Pr(a \mid \bar{c}) = \\
&q \cdot \Pr(c) + \Pr(a \mid \bar{c}),
\end{aligned}$$

where

$$q = \Pr(a \mid c) - \Pr(a \mid \bar{c}).$$

From our assumptions, we can also write

$$\begin{aligned}
&\Pr(a) = \Pr(a \mid b) = \\
&\Pr(a \mid b,c)\ \Pr(c \mid b) + \Pr(a \mid b,\bar{c})\ \Pr(\bar{c} \mid b) = \\
&\Pr(a \mid c)\ \Pr(c \mid b) + \Pr(a \mid \bar{c})\ \Pr(\bar{c} \mid b)] = \\
&\Pr(a \mid c)\ \Pr(c \mid b) + \Pr(a \mid \bar{c})[1 - \Pr(c \mid b)] = \\
&q \cdot \Pr(c \mid b) + \Pr(a \mid \bar{c}).
\end{aligned}$$

(Note that $\Pr(a,b \mid \bar{c}) = \Pr(a \mid \bar{c})\ \Pr(b \mid \bar{c})$ holds because $\overline{\{C = c\}} = \{C = 1 - c\}$ is an event.) Combining these two expressions, we arrive at

$$q \cdot \Pr(c) = q \cdot \Pr(c \mid b).$$

For this to be true, either $q = 0$, i.e.,

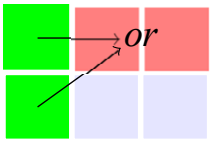
$$\Pr(a \mid c) = \Pr(a \mid \bar{c}),$$

which means $A \perp C$, or

$$\Pr(c) = \Pr(c \mid b),$$

which means $C \perp B$. □

In particular, assuming any column in (3) implies one of the other unconditional independence notions. Graphically,



and this is true for every column in (3), not only the first. Still, our most important conclusion is, in prose

> **Trade-off 1:** if equal odds and equal outcomes both holds, then selection is target indifferent or we're all equal.

Equivalently, unless groups have equal target base rates, or selection is indifferent, the ideals of equal odds and equal outcomes are incompatible.

5.2 Triangulating

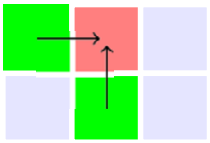
Proposition 5. If $B \perp C$ and $A \perp B \mid C$ then $A \perp B$.

Proof. Using again the shorthand a for the event $\{A = a\}$, etc., we have

$$\begin{aligned}
&\Pr(a,b) = \Pr(a,b \mid c)\Pr(c) + \Pr(a,b \mid \bar{c})\Pr(\bar{c}) = \\
&\Pr(a \mid c)\Pr(b \mid c)\Pr(c) + \Pr(a \mid \bar{c})\Pr(b \mid \bar{c})\Pr(\bar{c}) = \\
&\Pr(a)\Pr(b \mid c)\Pr(c) + \Pr(a)\Pr(b \mid \bar{c})\Pr(\bar{c}) = \\
&\Pr(a)\left[\Pr(b \mid c)\Pr(c) + \Pr(b \mid \bar{c})\Pr(\bar{c})\right] = \\
&\Pr(a)\Pr(b).
\end{aligned}$$

□

In particular, assuming an entry in each row in (3) from different columns implies the unconditional notion at the top row. Graphically,



The result is true no matter which two boxes we pick, as long as they are in different columns. But the most important conclusion, in prose, is that

Trade-off 2: if demographic parity and predictive parity both hold, then we're all equal.

Equivalently, unless we're all equal, a classifier cannot achieve both demographic and predictive parity.

Proposition 6. Assume $A \perp B \mid C$ and $A \perp C \mid B$. Then $A \perp B$ and $B \perp C$, unless the involved probabilities are zero.

Proof. To be precise, we will show that under the assumptions,

1. $A \perp B$ if for all b and c, we have $\Pr(B = b, C = c) > 0$, and

2. $A \perp C$ if for all a and c, we have $\Pr(A = a, C = c) > 0$.

We show the first statement; the other is similar. Write again a for $\{A = a\}$, etc. Using conditional probability and the assumptions, we have

$$\Pr(a,b,c) = \Pr(a\, b \mid c)\, \Pr(c) =$$
$$\Pr(a \mid c)\, \Pr(b \mid c)\, \Pr(c) = \Pr(a \mid c)\, \Pr(b,c).$$

and

$$\Pr(a,b,c) = \Pr(a,c \mid b)\, \Pr(b) =$$
$$\Pr(a \mid b)\, \Pr(c \mid b)\, \Pr(b) = \Pr(a \mid b)\, \Pr(c,b)$$

Since $\Pr(b,c) = \Pr(c,b) \neq 0$, we deduce $\Pr(a \mid c) = \Pr(a \mid b)$ for all a, b, c. We can use this (for c and $\bar{c}$) in the following derivation,

$$\Pr(a)\, \Pr(b) = [\Pr(a \mid c)\, \Pr(c) + \Pr(a \mid \bar{c})\, \Pr(\bar{c})]\, \Pr(b) =$$
$$[\Pr(a \mid b)\, \Pr(c)\, \Pr(b) + \Pr(a \mid b)\, \Pr(\bar{c})]\, \Pr(b) =$$
$$\Pr(a \mid b)\, \Pr(b)\, \Pr(c) + \Pr(a \mid b)\, \Pr(b)\, \Pr(\bar{c}) =$$
$$\Pr(a,b)\, [\Pr(c) + \Pr(\bar{c})] = \Pr(a,b),$$

which establishes $A \perp B$. □

In particular, assuming two of the fairness notions from the bottom row implies their unconditional counterparts. Graphically,

Most importantly:

> **Trade-off 3:** if equal odds and predictive parity both hold then we have demographic parity and we're all equal.

Equivalently, unless we're all equal and the classifier achieves demographic parity, then the classifier cannot both guarantee equal odds and predictive parity.

References

Barocas S., Hardt M., and Narayanan, A. (2019) *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org. http://www.fairmlbook.org.

Dawid, A.P. (1979) Conditional independence in statistical theory. *Journal of the Royal Statistical Society, Series B.*, 41(1):1–31.

Friedler S.A., Scheidegger C., and Venkatasubramanian S. (2021) The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making. *Communications of the ACM*, 64(4):136–143.

Lippert-Rasmussen, K (2023). Post hoc interventions and machine bias. *This volume*.

Pearl, J. (2009) *Causality*. Cambridge University Press.

Stone P. (2009) Logic of random selection. *Political theory*, 37(3), 2009.

Verma S. and Rubin J. (2018) Fairness definitions explained. In *Proc. of 2018 ACM/IEEE International Workshop on Software Fairness, FairWare'18, May 29, 2018, Gothenburg, Sweden*.