

Fair Decisions, Hard and Soft

Kenneth Silver – Trinity College Dublin (Ass. Prof. in Business Ethics)

Gregory Faletto – University of Southern California (PhD student in Dept. of Data Sciences and Operations)

Agenda

- Discuss how certain notions of fairness are apparently mutually unsatisfiable
- Appreciate a critical distinction in mathematical optimization
- Appeal to that distinction to chart a path to mutual satisfiability
- Consider whether we are justified in using the distinction as we do
- Show how the involvement of rights may determine how we use the distinction

An Impossible Problem

The literature has arrived at different notions of fairness in making decisions, and it's not always clear how these notions fit together, or if they even can.

Consider the example of college admissions:

- Suppose there exist two groups of applicants defined by some “sensitive” or “protected” variable (e.g., two races or two genders). (In general, there could be any number of sensitive variables, each containing any number of groups.)
- We can also divide the applicants into “qualified” applicants (in some sense that is not visible at the time of the admissions decision but may be visible later—e.g., whether they go on to graduate in five years or less) and “unqualified” applicants.

Our goal is to create an algorithm to admit students that satisfies some notion of fairness with respect to group membership.

An Impossible Problem

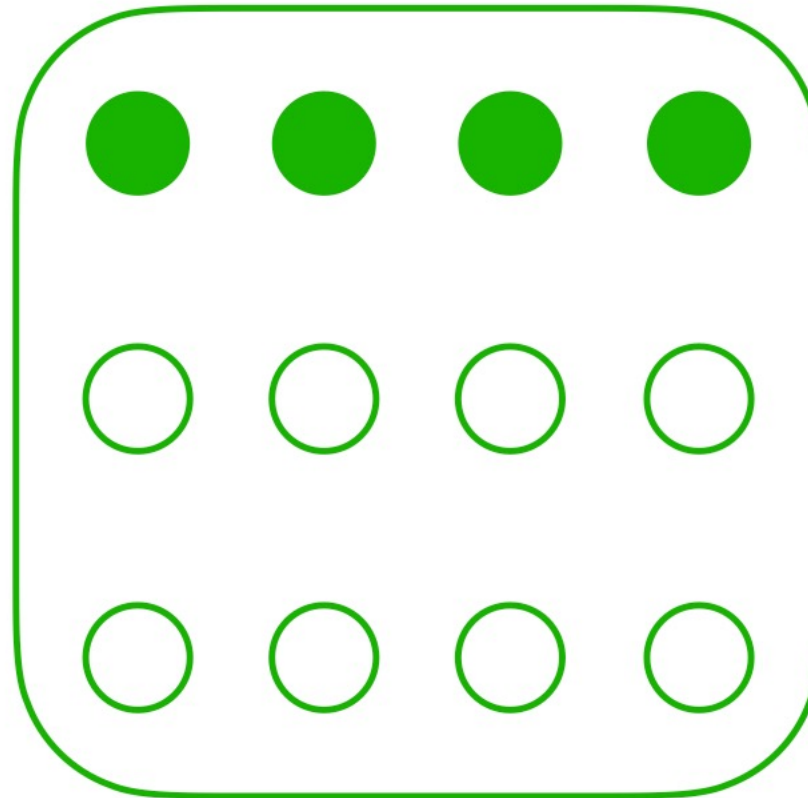
A couple examples of notions of fairness discussed in the statistics and computer science literature (in the college admissions setting):

- **Demographic parity:** the proportion of applicants admitted from each sensitive group should be equal. (That is, the group composition of admitted students should match the group composition of applicants. Similar to affirmative action.)
- **Equalized odds:** the proportion of *qualified* applicants admitted from each sensitive group should be equal, and likewise for unqualified applicants. (That is, conditional on whether you're qualified, whether or not you're in a sensitive group should not affect your admissions chances.)
- **Calibration:** The proportion of Group A admitted applicants who turned out to be qualified should be equal to the proportion of Group B admitted applicants who turned out to be qualified, and likewise for rejected applicants. (That is, the proportion of "accepted Group A members" who are qualified should be equal to the proportion of "accepted Group B members" who are qualified, and likewise for rejections.)

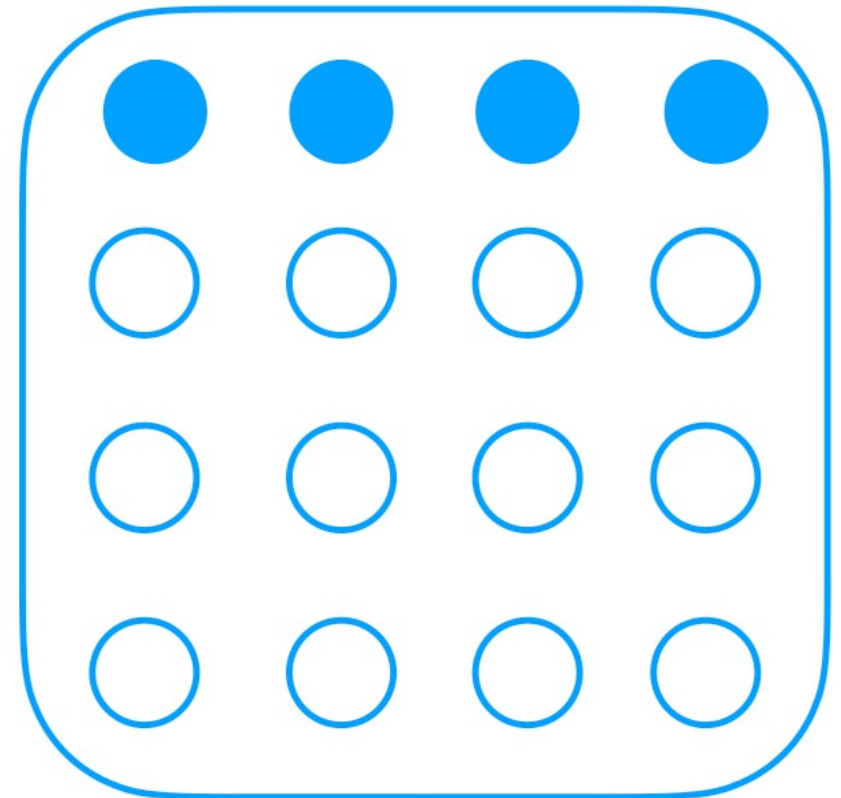
Setting: college admissions. 12 members of Group A apply; 4 are qualified. 16 members of Group B apply, 4 are qualified.

- **Demographic parity:** admit an equal proportion of members of Group A and Group B.
- **Equalized odds:** admit an equal proportion of members of each group when you stratify by whether or not they are qualified.
- **Calibration:** the proportion of applicants who are qualified is equal in each group when you stratify by whether or not they are admitted.

Group A



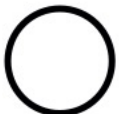
Group B



Key



Qualified




Unqualified



Admitted

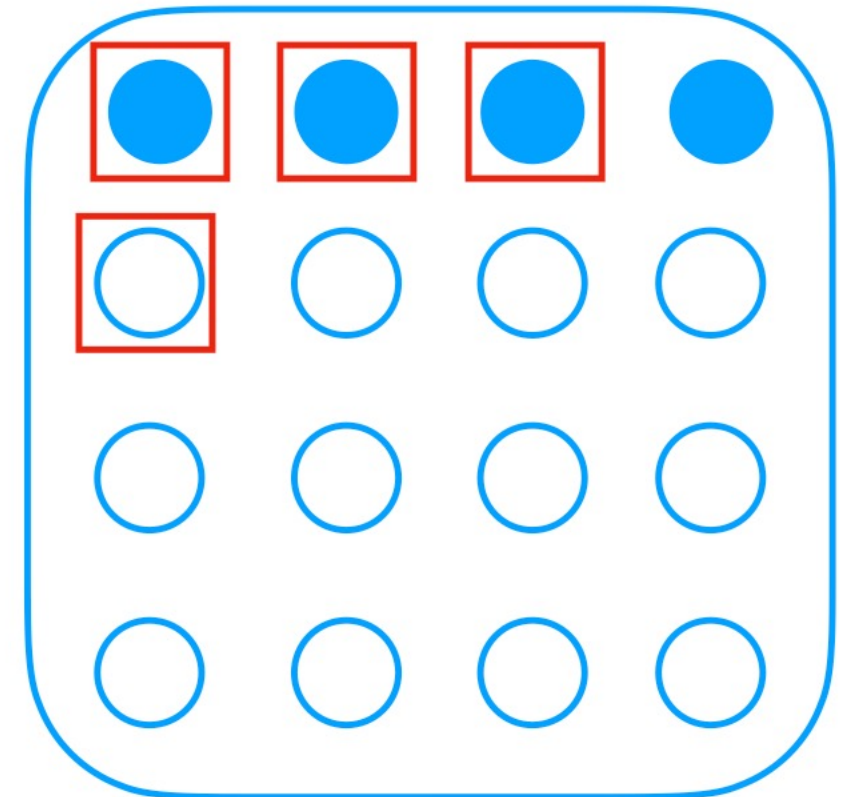
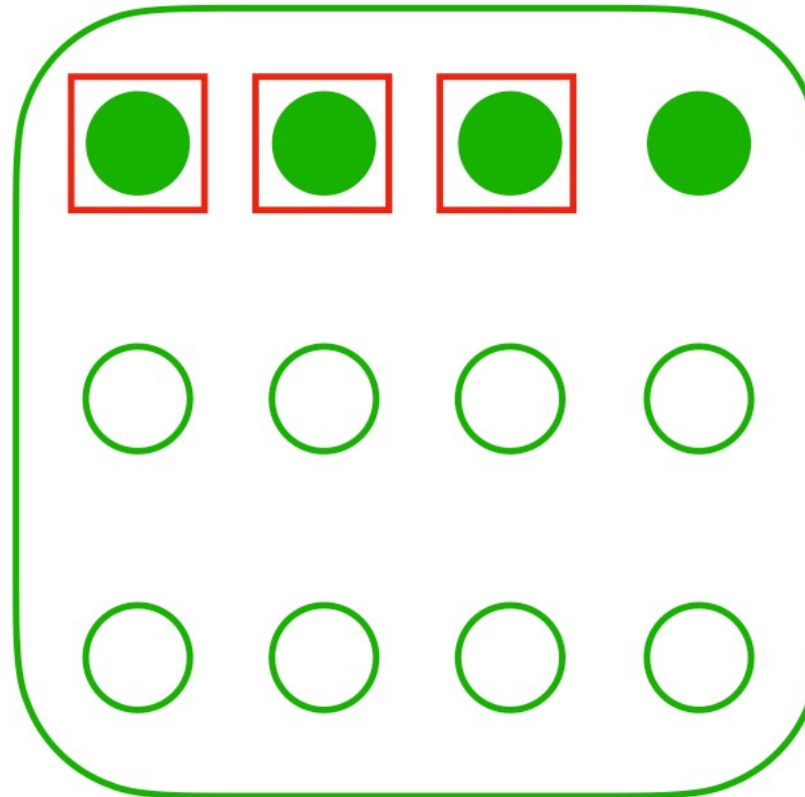
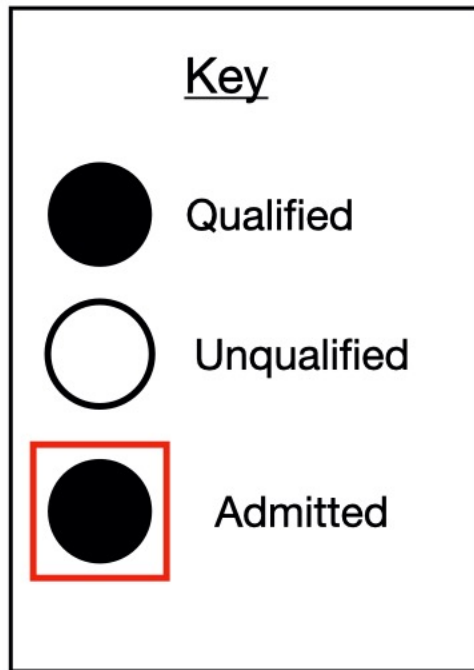
Demographic parity?  because for each group, 1/4 of the applicants were admitted.

Equalized odds?  because 0% of the unqualified group A members were admitted, 8.3% for Group B (unfair to Group A)

Calibration?  because 75% of admitted Group B members were qualified but 100% of admitted Group A members were (unfair to Group A)

Group A

Group B



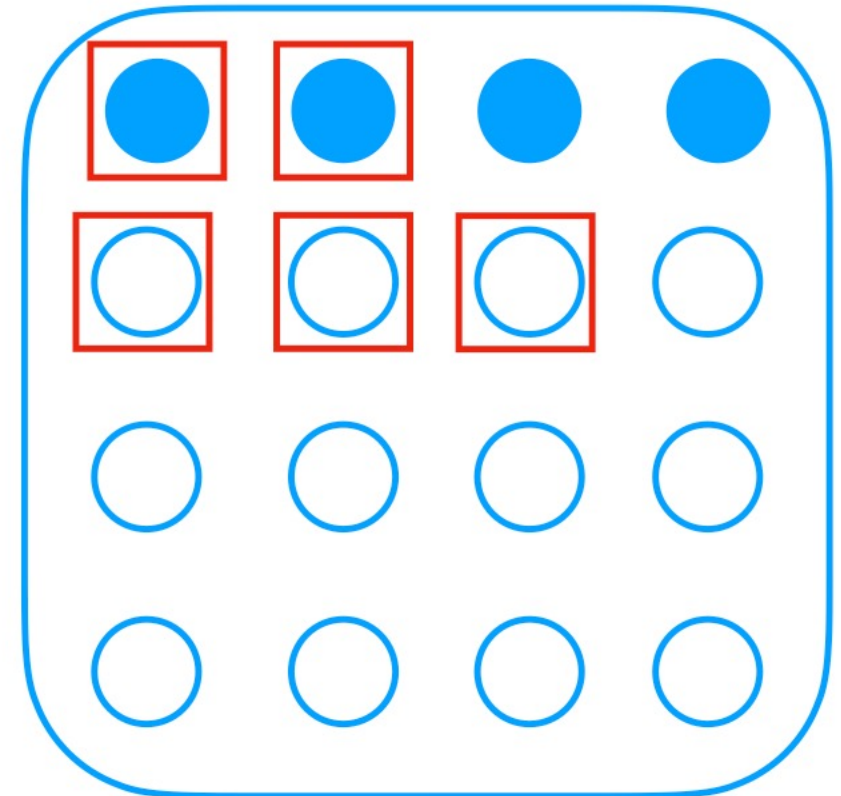
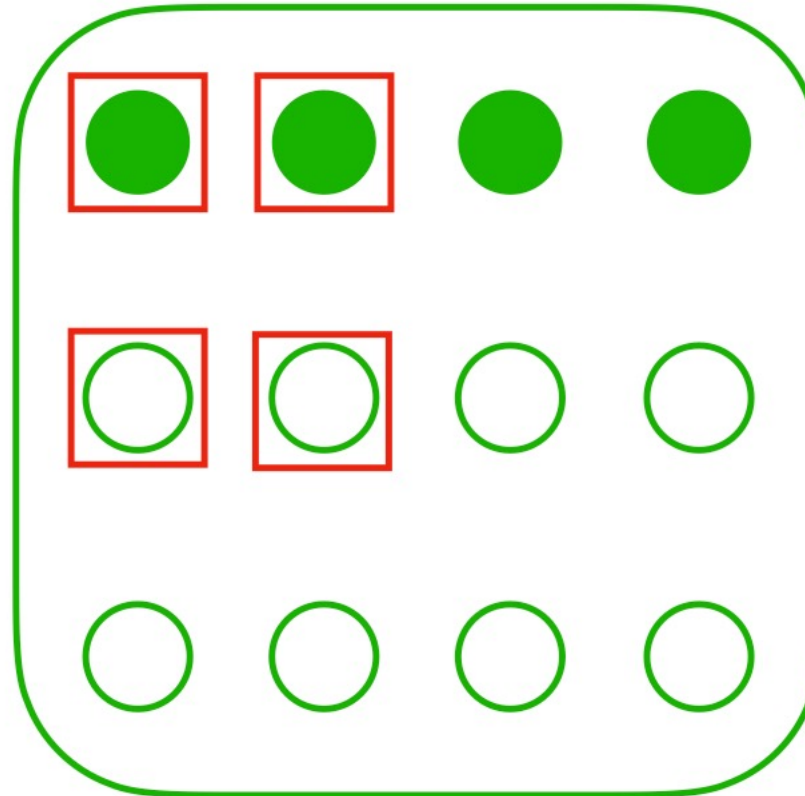
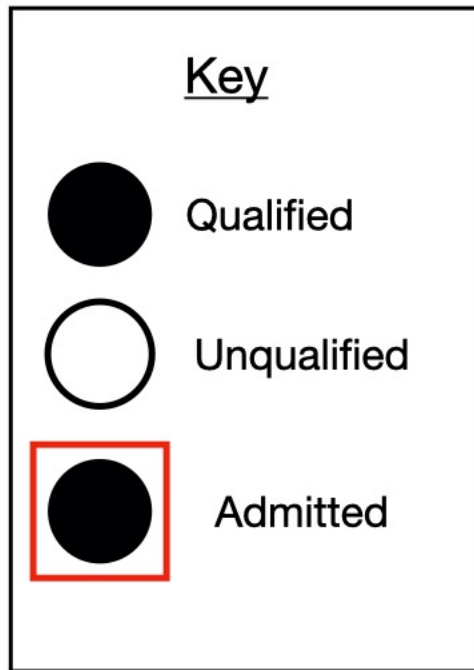
Demographic parity? ✗ because 33.3% of Group A members were admitted and 31.3% of Group B members were (unfair to B)

Equalized odds? ✔ because in both groups, half of the qualified members and 1/4 of the unqualified members were admitted

Calibration? ✗ because 40% of admitted Group B members were qualified but 50% of admitted Group A members were (unfair to Group A)

Group A

Group B



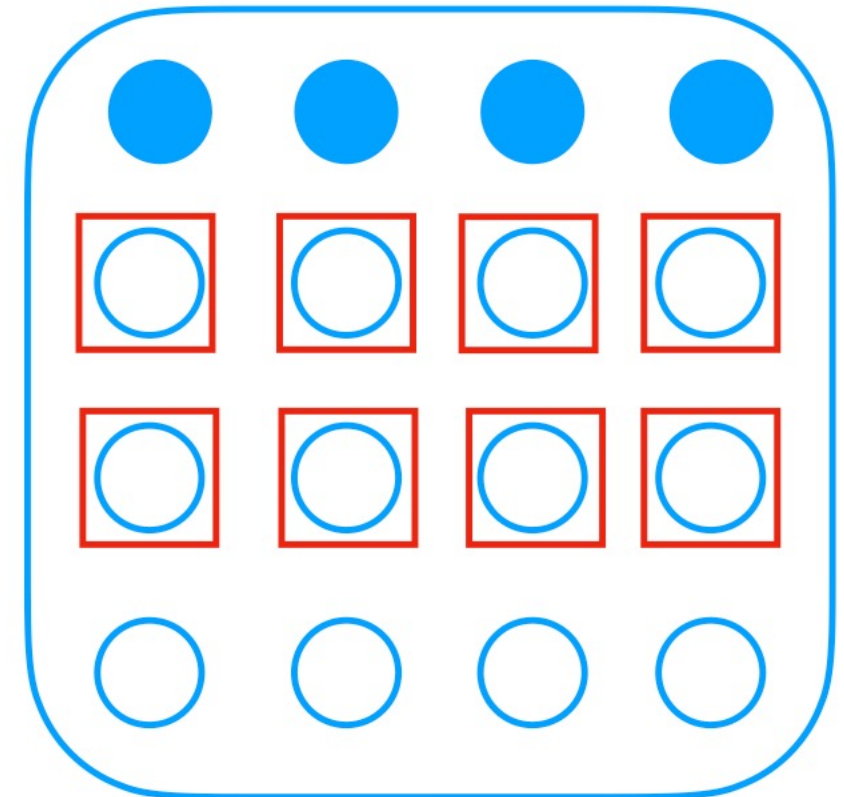
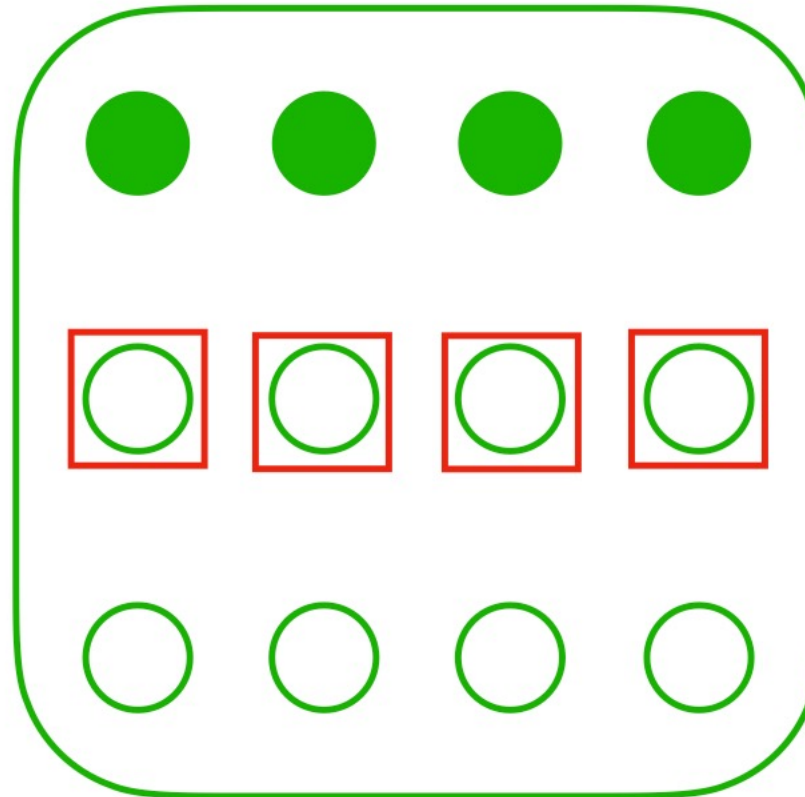
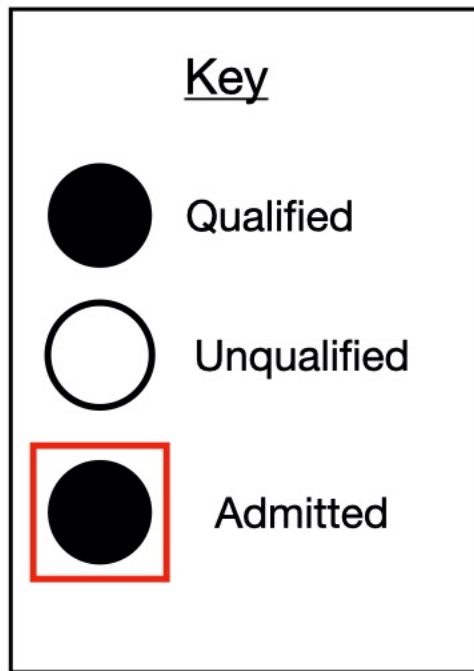
Demographic parity? ✗ because 33.3% of Group A members were admitted and 50% of Group B members were (unfair to A)

Equalized odds? ✗ because only 50% of the unqualified members in Group A were admitted but 66.7% of unqualified Group B members were admitted (unfair to Group A)

Calibration? ✓ because in both groups 100% of admitted members were unqualified and 50% of rejected members were qualified.

Group A

Group B



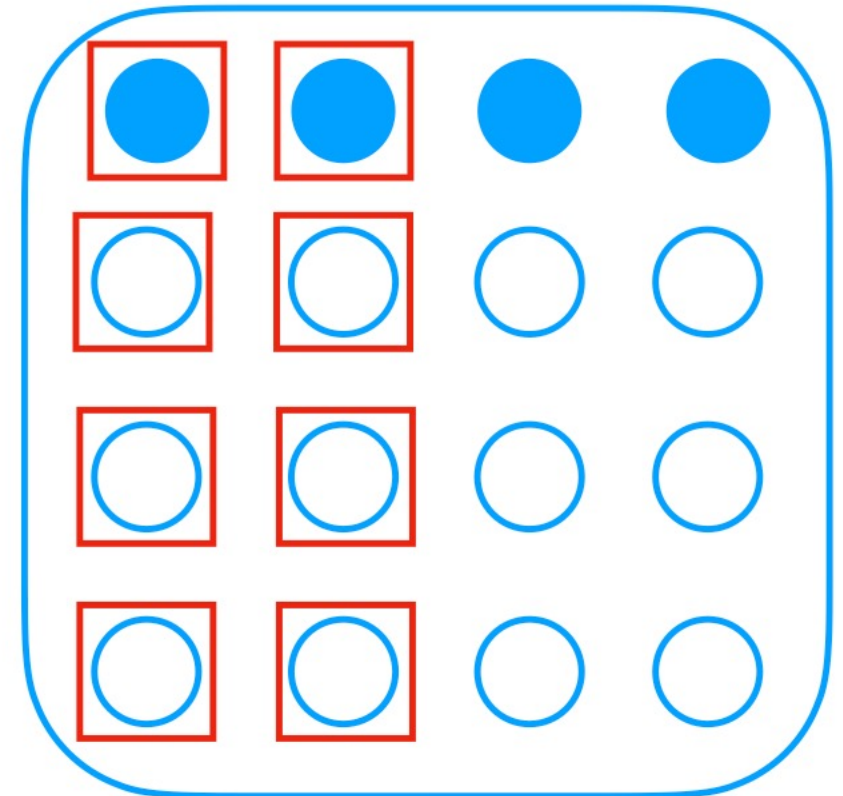
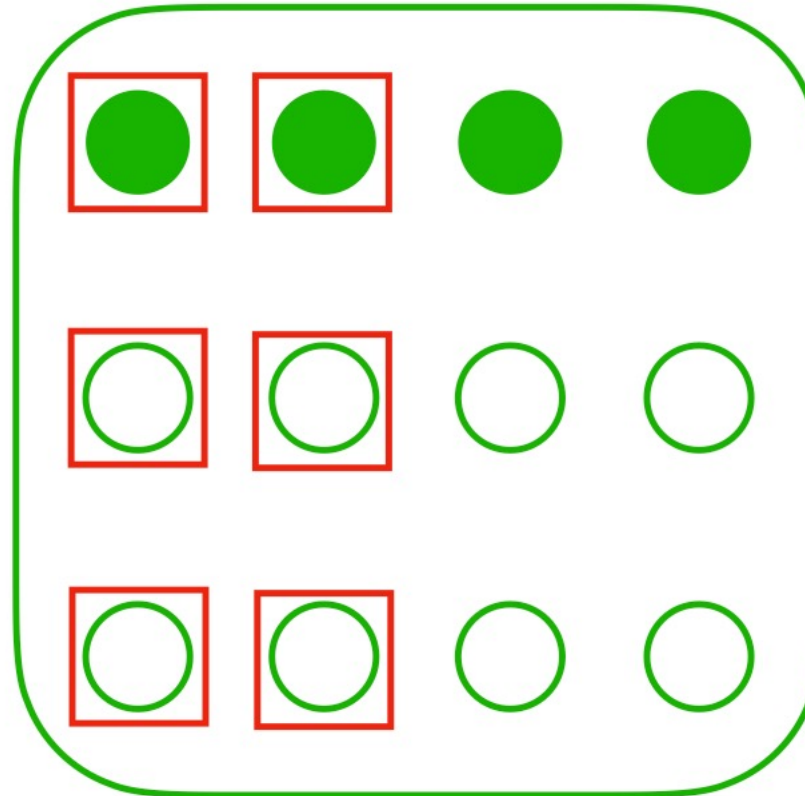
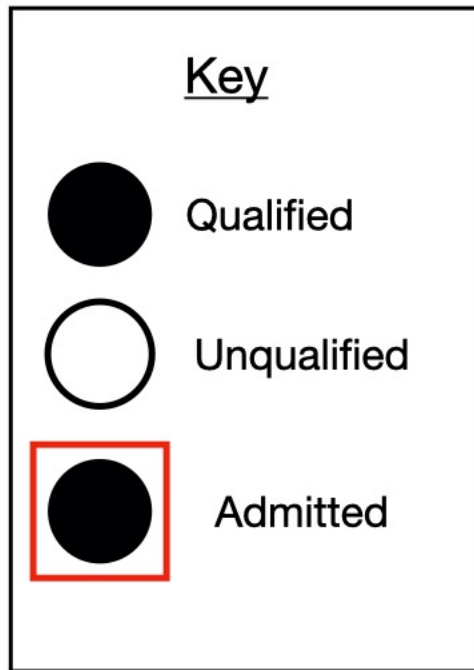
Demographic parity? ✓ because half of Group A members were admitted and half of Group B members were admitted.

Equalized odds? ✓ because in both groups, half of the qualified members and half of the unqualified members were admitted

Calibration? ✗ because 33.3% of admitted Group A members were qualified but 25% of admitted Group B members were (unfair to Group A)

Group A

Group B



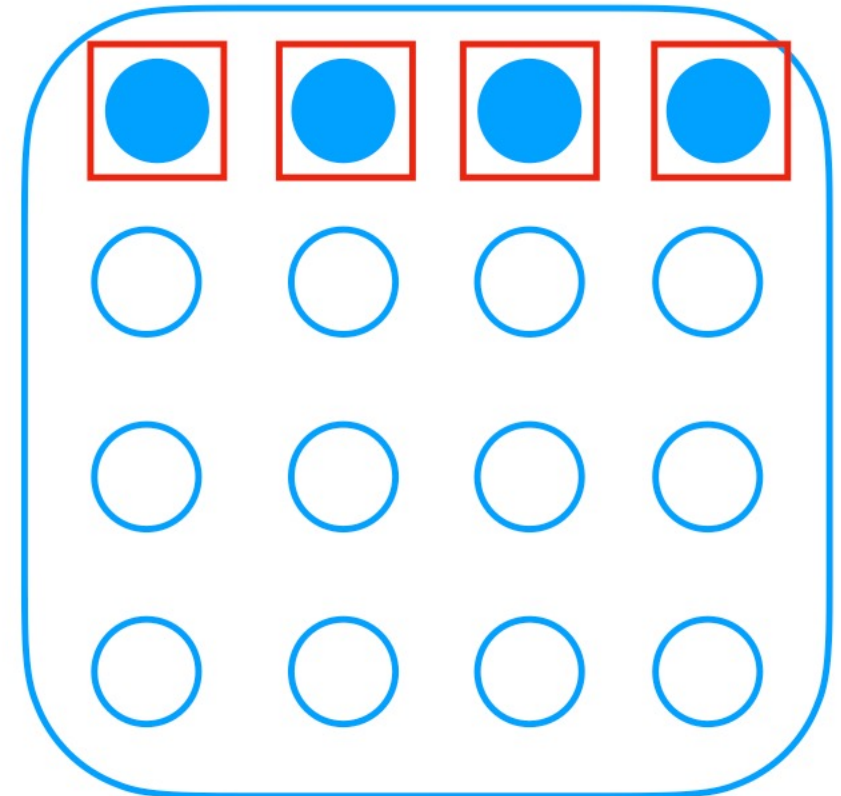
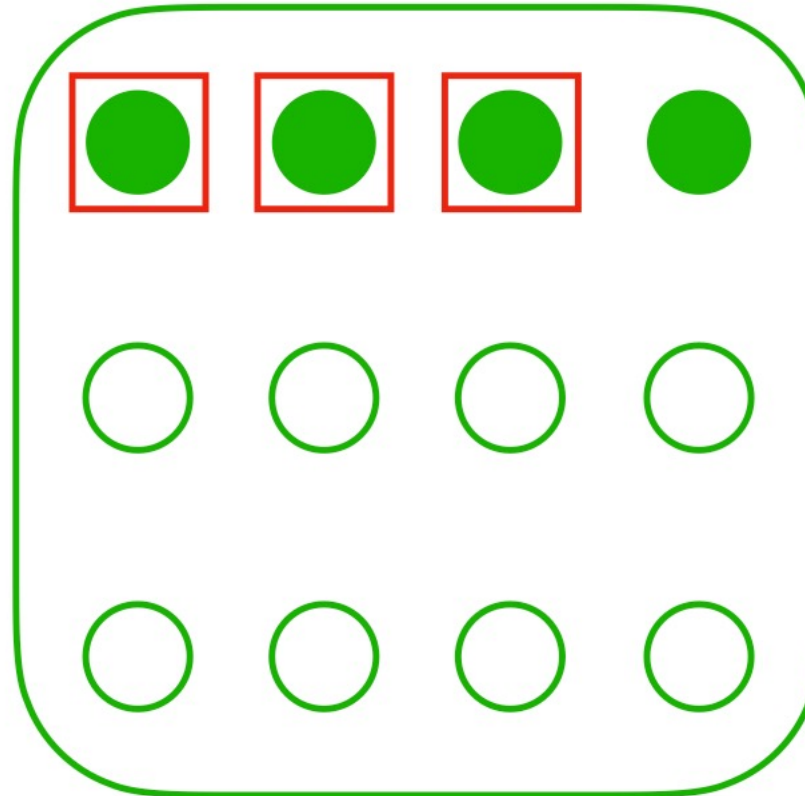
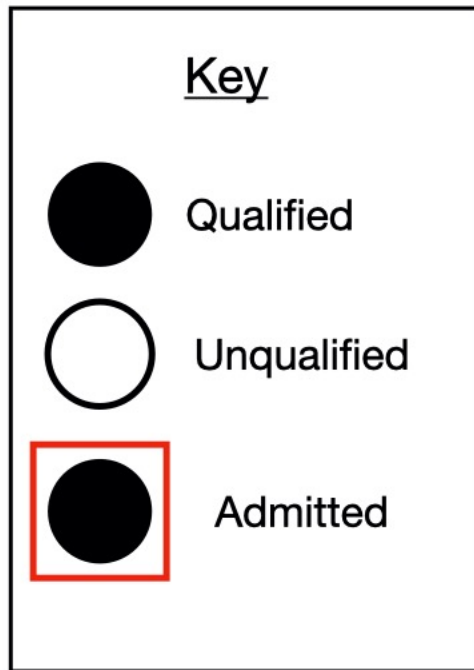
Demographic parity? ☒ because 25% of Group A members were admitted and 25% of Group B members were admitted.

Equalized odds? ☒ because only 75% of the qualified members of Group A were admitted but 100% of the qualified Group B members were (unfair to Group A)

Calibration? ☒ because in Group A 11.1% of the rejected candidates were qualified but 0% of the rejected candidates from Group B were qualified (unfair to Group A)

Group A

Group B



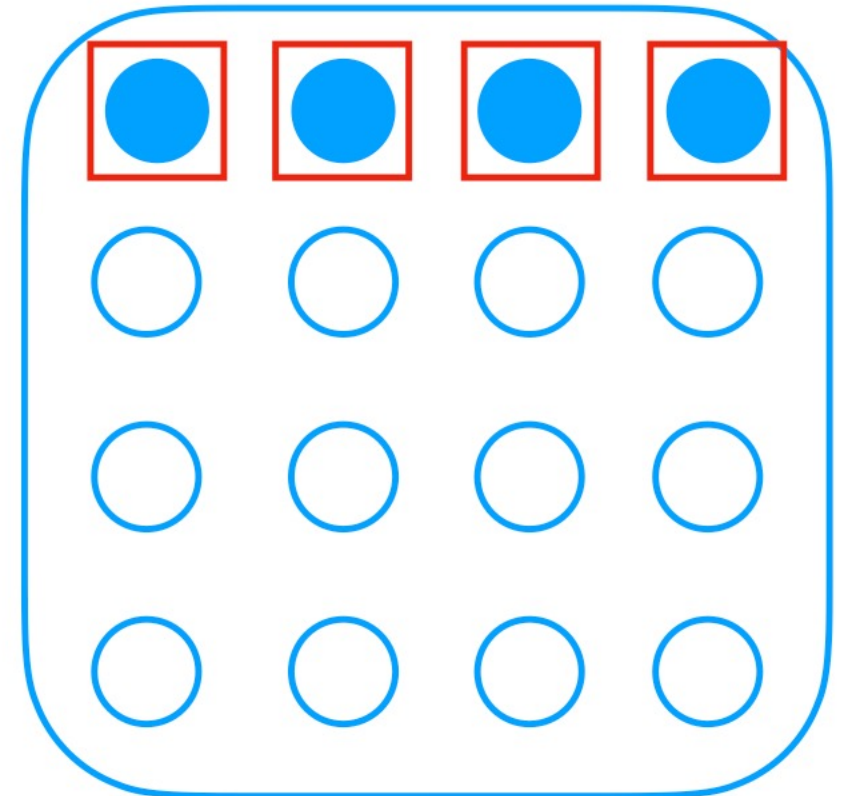
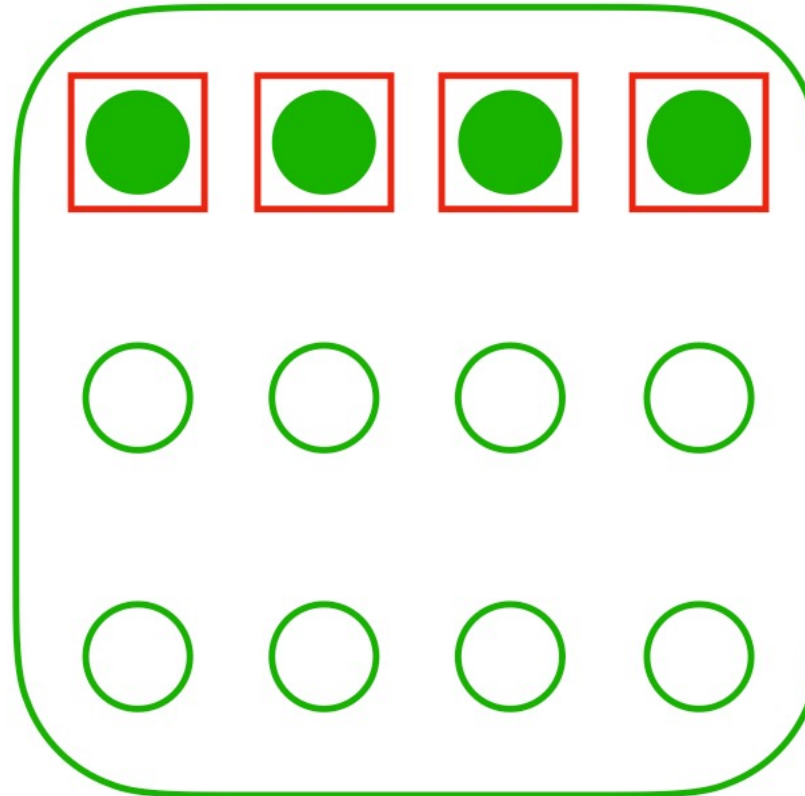
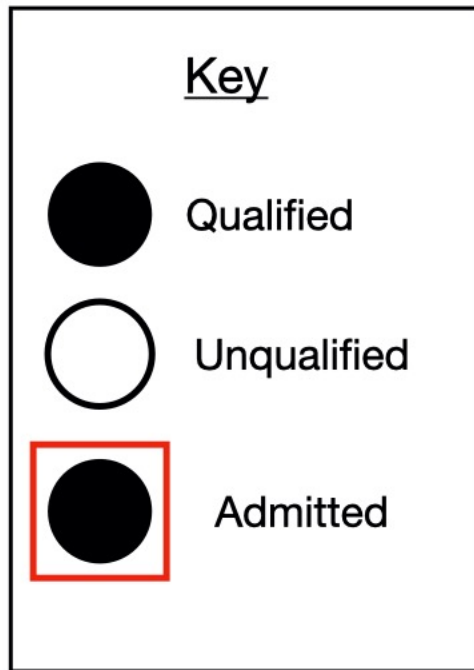
Demographic parity? ✗ because 33.3% Group A members were admitted but only 25% of Group B members were (unfair to B)

Equalized odds? ✓ because in both groups 100% of the qualified members were admitted and 0% of the unqualified members were

Calibration? ✓ because in both groups 100% of the admitted members were qualified and 0% of the rejected members were qualified

Group A

Group B



Note: per the impossibility theorem, this is the only scenario in which both equalized odds and calibration are satisfied (since the proportions of each group that are qualified are not equal—33.3% of Group A is qualified, only 25% of Group B is). This means there is no scenario where all three criteria are satisfied.

An Impossible Problem

THEOREM 1.1 (in the college admissions example context; Kleinberg et al. [2017]). *Suppose an algorithm satisfies both calibration and equalized odds. Then either (a) there is no correlation between sensitive group membership and being qualified, or (b) the algorithm can perfectly predict who's qualified and who isn't.*

Since there is almost always some correlation between sensitive group membership and being qualified (e.g. through unequal access to opportunities or present-day discrimination) and real-world data doesn't allow for perfect algorithms, we interpret this as an impossibility result: **both notions of fairness can't be simultaneously satisfied.**

Possible Solution?

Give up at least one of the conceptions of fairness...

Perhaps this is yet another time when we face an inconsistent group of plausible claims and have to bite the bullet in denying one.

Perhaps there are compelling reasons to sacrifice one of these notions.

From Sune's talk, we saw how we might be able to sacrifice accuracy for Equalized Odds (perhaps similar considerations allow us to sacrifice Calibration).

Or perhaps different criteria matter in different contexts.

From Clinton, David, and Ben's talk, we saw how different conceptions of fairness might cut across each other in ways that suggest that different conceptions may be better more appropriate for different contexts.

Reexamining the Source of the Problem

The problem concerns the constraints on algorithms. We are trying to balance distinct considerations:

- We want the algorithm to predict who's qualified as successfully as possible.
- We want one or more notions of fairness to be satisfied.

The impossibility theorem shows we can't insist on perfectly satisfying two notions of fairness. And the problem is that we need at least one algorithm here to actually use.

What if there were a way to take these fairness criteria into account in *some* way, but where an algorithm's failure to perfectly satisfy the constraint does not automatically mean that we will not use it?

Hard and Soft Constraints in Mathematical Optimization

In mathematical optimization, frequently problems are encountered where we would like to satisfy some constraint exactly, but doing so is infeasible (perhaps for computational reasons, perhaps because the real-life data won't allow it).

In these cases, a common approach is to *relax* the constraint. We solve the optimization problem so that violations of the constraint are *discouraged* (in proportion to how bad the violations are) rather than *ruled out*.

We follow a popular mathematical optimization textbook (Boyd & Vandenberghe 2004:Sec.5.1.4) in referring to these exact constraints as **hard constraints** and the relaxed versions as **soft constraints**.

Hard and Soft Constraints in Mathematical Optimization

Example: An institutional investor needs to allocate an endowment across a collection of stocks. The investor would like there to be no volatility in returns at all. She looks at stock returns over the last five years and looks for a portfolio allocation that would provide maximum returns with a **hard constraint** of 0 volatility; that is, she seeks a solution to the following (stylized) optimization problem:

Maximize	returns
Subject to	volatility = 0.

Hard and Soft Constraints in Mathematical Optimization

However, no allocation makes this possible. She relaxes her constraint of 0 volatility and solves the following optimization problem with a **soft constraint** on volatility:

$$\text{Maximize} \quad \text{returns} - \lambda \cdot \text{volatility}$$

where λ is a constant needed to reconcile different units of measurement and assign relative weight to the goals of maximizing returns versus minimizing volatility.

λ gives us a knob to twist: when it's small, we care mostly about maximizing returns; when it's large, we care mostly about minimizing volatility; and we can choose any point in between to balance our concerns.

Applying the Distinction

It's easy to see that we have been treating the distinct fairness criteria as ***hard constraints*** in our decision about which algorithm to use.

Although we might be interested in more than one notion of fairness, we often can't insist on satisfying all of them with exact equality due to the impossibility theorem.

This makes it clear how the impossible result might be averted: It might be possible to relax our understanding of these constraints, even if they remain in place, and instead treat them as ***soft constraints***.

In Practice

Consider again the college admissions problem. The college has past data on students they admitted and did not admit, as well as information on whether each student was in hindsight “qualified.” Using the past data, the college wants to fit a decision algorithm by solving the following optimization problem with hard constraints:

Maximize	accuracy of predictions about whether each student is qualified
Subject to	violations of equalized odds = 0
	violations of calibration = 0.

Motivation

Why would we think that the constraints at hand are better thought of as soft constraints?

Historically, Broome suggested that fairness mattered *but could be outweighed*. He says,

“It may be that fairness requires a lottery, so that it would be unfair not to hold one, but that in this case fairness is outweighed by expediency, so that on balance it is right to send the talented candidate without a lottery. This depends on the circumstances. If it is vital that there should be no slip in the execution of the mission, the unfairness will be tolerable. But if a less than perfect performance is acceptable, more importance can be given to fairness. In some circumstances, fairness will win, and a lottery should be held.” (1990:90)

And so, perhaps this is really what we are talking about when we are talking about soft constraints. So, that an algorithm is unfair in some way may provide a significant reason not to use it, but this is a reason to be weighed against the positive outcome of its use and compared with calculations for similar algorithms.

A Natural Further Challenge

If we assume that the fairness constraints are all legitimate soft constraints, how do we decide the right way to weigh them? (That is, how do we choose those λ parameters?)

One idea: If all fairness constraints seem relevant, with no clear reason to weight one more, then weight them equally (after accounting for any differences in units of measure). (Perhaps there is some meta principle of fairness that calls for equal weighting absence a reason to prioritize any of them.)

Another idea: There are times where several fairness constraints seem relevant, but one may seem more important than others. (We saw this in a few of the talks yesterday.) We will need some way for judging the weights, but it might be *that* bad.

Must They Be Soft Constraints?

So, there is some reason to treat fairness considerations as soft constraints, and doing so makes the math tractable, but *must* they be soft constraints?

From the impossibility theorem, we know that they can't ***all*** be hard constraints. But perhaps some of them are and some of them are not.

Looking at them again, for example, we might think that Calibration and Demographic Parity offer genuine conceptions of fairness that we should aim to achieve. But we may think that Equalized Odds offers a more significant constraint. Perhaps it is not simply something bad to fail to satisfy, but something it is *wrong* to fail to satisfy.

If an algorithm does not satisfy Equalized Odds, then a prospective qualified student may have a smaller chance of being admitted *purely because of their race or gender*. If they are not admitted, then we may take them to have been discriminated against, and we may understand ourselves as having a **right** against this form of discrimination.

(In Moreau's [2010] terms, the prospective student will have had a certain right interfered with, the right to make decisions about how to live their life free from the effects of normatively extraneous features of them.)

So, crucially, there is a plausible story to tell that ties together this conception of fairness with the having of a certain right.

Why would that matter?

Rights, Exclusionary Reasons, and Hard Constraints

It's controversial, but some scholars understand rights as generating reasons that cannot simply be weighed as in a normal cost-benefit calculation.

If you want to do X, then you have a reason to do X. But if I have a claim right that gives you a duty to not do X, then this doesn't simply give you a reason of the same kind (even a strong one) to not do X. Instead, the fact that you have a duty to not X seems to exclude X from among your available options of things to do.

On this understanding of rights, we say that having a right generates ***exclusionary reasons***, and exclusionary reasons do not get weighed directly against considerations for or against some option. Instead, they come first to constrain the options themselves, which are then weighed as normal. (See Adams [2019] – “In Defense of Exclusionary Reasons”)

Rights, Exclusionary Reasons, and Hard Constraints

Example...

If I own my house, then I have a right to its private use, and you have a duty not to enter my house without my consent. Now, suppose you're trying to decide what to do, and it might be really fun to hang out at my house without my consent. Still, your duty excludes the possible fun as counting as a reason to be weighed in the decision of what to do, because going to my house is off the table before we even start thinking about it.

Rights, Exclusionary Reasons, and Hard Constraints

Consider these reasons concerning your going to my house:

- (1) It is usually fun to hang out at my house (especially when I'm not there).
- (2) My house is across town from you.
- (3) Chad might also be at my house, and he's not fun.
- (4) I have a right to who comes to my house and don't consent to your coming over.

Now...

- (1) is a reason in favor of going to my house.
- (2) is a reason against going to my house that *rebutts* (1) – a rebutting defeater.
- (3) is a reason against going to my house that *undercuts* (1) – an undercutting defeater.
- (4) is a reason against going to my house that arguably *excludes* (1) from counting as a reason – exclusionary defeat

Mixed Approach

Insofar as rights generate reasons for agents that do not simply weigh for or against certain options, but instead constrain the field of options, it seems that rights function to generate hard constraints. So, the presence of rights justifies developing mixed approaches that respect all considerations of fairness, where some are treated as soft constraints and others as hard.

Going back to our example, where Equalized Odds is taken to provide a hard constraint grounded in a right against discrimination, and where Demographic Parity has been argued to be a good end (justified by its capacity for facilitating distributive justice, say), we can frame the mixed problem like this:

Maximize	accuracy of prediction - λ \cdot violations of demographic parity
Subject to	violations of equalized odds = 0

Assumptions & Further Challenges

Rights might not work this way in reasoning. We are assuming that rights genuinely do generate exclusionary reasons. And some people think that they do not. (See Essert [2012], Whiting [2017], Gur [2018], among others.)

There may be no rights and duties in the offing. We are assuming that there is a genuine right that some duty to a kind of fairness protects. But we might think that there is not. We may accept harm-based explanations of the badness of these sorts of unfairness, or think that we have no genuine rights in this context.

There may be *more* rights and duties in the offing. We are assuming that at most one of the fairness constraints are grounded in a right. But what if more than one is, re-threatening the impossibility problem? (Likely solution, treat both as soft constraints, but assume the further duties of precaution and moral repair for failing to meet the obligation.)

We might be able to outweigh reasons from rights *too*. There's been a string of works lately that tries to show how genuinely deontological duties can exist yet can *sometimes* be outweighed. (See Lee-Stronach [2018], Lazar [2019], Lazar & Lee-Stronach [2019]) So we'd need a view about when and how that happens and how it applies here, how rights can generally provide reasons that exclude but perhaps with exceptions.

Conclusion

So, for all we know, perhaps fairness constraints are never hard. Or, perhaps they may be in certain contexts. We are not committed either way.

Instead, our broader points are that:

- Apparently mutually unsatisfiable constraints can be accommodated as long as they are not all treated as hard constraints. And
- Considerations involving rights may be important to determining when fairness constraints must be treated as hard.

Thank you!