# Fairness and Machine Fairness

**Clinton Castro**
Department of Philosophy
Florida International University
Miami, FL USA
clinton.g.m.castro@gmail.com

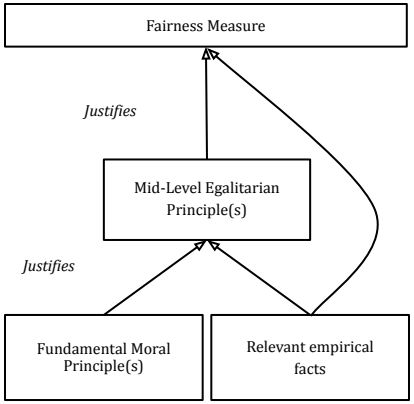
**David O'Brien**
Department of Philosophy
Tulane University
New Orleans, Louisiana USA
d.uibhriain@gmail.com

**Ben Schwan**
Department of Bioethics
Case Western Reserve University
Cleveland, Ohio USA
b.e.schwan@gmail.com

**1. Overview**
We provide a framework for thinking about the connection between fairness measures, their egalitarian roots, and the standards that justify their use in different contexts.

Using the framework, we explore the connections between three fairness measures and three egalitarian ideals.

We show that, although some of these measures align with some of these ideals some of the time, none align with any of these ideals all of the time. Put another way, none of the measures we discuss can be used as an off-the-shelf measure for tracking any of these ideals. Further, we argue, which—if any—of these ideals is correct varies from context to context. So, users of fairness measures must take care to consider which egalitarian ideal is salient in the context of interest and which measure best captures that ideal in that context.




**2. The framework**
Some observations:

Disagreements over a fairness measure can have many sources.

There are a number of normative criteria to evaluate fairness measures.

Defenses and rejections of fairness measures can be extremely limited.

We do not have to agree all the way down to agree on a measure.

**3. Formal Equality of Opportunity (FEO) and Fairness Through Unawareness (FTU)**
Principle FEO requires that, "positions and posts that confer superior advantages [...] be open to all applicants. Applications are assessed on their merits, and the applicant deemed most qualified according to appropriate criteria is offered the position" (Arneson, 2015).

Measure FTU asks that a prediction-based decision system not take as inputs protected attributes (Grgic-Hlaca et al. 2016)

Satisfying FTU is not sufficient for satisfying FEO:

**Graduate School.** The admissions system for a graduate program requires scores of a test that is only administered on a religious holiday for a minority group. Requiring the scores of that test will ensure that most members of the minority group will not be able to take the test and, thus, will be unable to apply.

Nor is it necessary:

**Jobs.** You are hiring. Job applicants take a free aptitude test. You know that members of some minority suffer from a pronounced stereotype threat that reduces their score on this test. (They are just as qualified for the position; the testing environment just has this feature.) So when you assess applications, you take their minority status into account by adjusting their scores.

**4. FEO and Equalised Odds (EO)**
Measure EO asks that the probability that the system correctly predicts that subjects have the property that is being predicted is independent of their protected attributes.

Satisfying EO is not sufficient for satisfying FEO:

Imagine that in Graduate School, a few members of the minority group sit for the exam. We can imagine that the exam itself is perfectly accurate, such that the members of the minority and members of the majority pass iff they are qualified.

Nor is it necessary:

**Jobs 2**. You are hiring. Job applicants take a highly--but not perfectly--accurate, free aptitude test. Members of an oppressed minority group are much more likely to be qualified. This is because other employers discriminate against this group, leaving a high portion of qualified members of the group on the job market. As a result of the discrepancies in the base rates--where members of the minority group are much more likely to be qualified--and the test's being highly--but imperfectly--accurate, true positives are more common among the minority group than in the majority group (i.e., it violates equalised odds).

**5. FEO and being fair**
Note that FEO could not *generally* be the correct egalitarian principle:

That is, satisfying FEO isn't enough to be fair:

**Graduate School 2.** The admissions system for a graduate program requires scores of a difficult test and members of minority groups, on average, do poorly on the test because they cannot afford the test preparation needed to be competitive.

Nor is it necessary:

**Jobs 3.** You are hiring. All are welcome to apply but minorities don't know this and are unlikely to be qualified, due to educational inequities. You devise a predictive system for predicting which applicants will either arrive qualified or able learn on the job. You devise a lottery system to locate to allocate positions among them (so as not to perpetuate inequality).

**6. Substantive Equality of Opportunity (SEO) and Counterfactual Fairness (CF)**
Principle SEO prevails with respect to some desirable position or ranked order of positions just in case all members of society are eligible to apply for the position, applications are fairly judged on their merits and the most meritorious are selected, and sufficient opportunity to develop the qualifications needed for successful application is available to all (Arneson, 2015; emphasis ours).

Measure CF says that a prediction-based decision is fair *iff* it is the same in the actual world and any counterfactual world where the individual belongs to a different demographic group. NB: if any variable is sensitive to a protected attribute, then the prediction is counterfactually unfair (Kusner et al. 2017).

Satisfying CF is not sufficient for satisfying SEO:

**Law School Success 2.** Seeing that race affects studiousness and valuing counterfactual fairness, the admissions team seeks to base entry on a different variable. As it turns out, applicants come from two different, perfectly integrated high schools to which students were randomly assigned. One highschool happens to have high quality pre-law courses, whereas the other has no pre-law courses whatsoever. The admissions team uses the pre-law course as a determining factor in their admissions decisions, reasoning that anyone who passed their pre-law courses would excel in law school.

Nor is it necessary:

**Internship.** You are hiring for an internship that requires english/spanish fluency. All students had to take a language class, but only some took spanish (even though all had the opportunity). Among the cohort, there are many hispanic students who speak spanish at home, and thus, are fluent whether or not they took the class. You have applicants take a spanish exam as part of their application.

**5. Luck Egalitarianism (LE) and CF**
Luck egalitarianism is the view that it is unfair for some to be worse off than others through no choice of their own (Cohen 2008).

Satisfying CF is not sufficient for satisfying LE:

**Insurance.** A bank offers no-questions-asked loan insurance to businesses. To decide whom to approve, they use an algorithm that is based on a sophisticated, proprietary prediction of the quality of of applicants' future business decisions. Frank and Rita both apply for insurance. Frank has a history of foolhardy business decisions, but as it happens (and as the algorithm correctly predicts) will not soon make another one. Rita has a history of responsible business decisions, but as it happens (and as the algorithm correctly predicts) will soon make an uncharacteristically foolhardy one. The bank thus denies Rita's application and approves Frank's.

Nor is it necessary:

See Internship (above).

**6. Some Lessons**
Choosing proper measures requires nuance and great sensitivity to the egalitarian principles that undergird our choice of measure.

Despite our confidence in a kind of pluralism about fairness measures, some general patterns exist; being aware of these patterns can help in making decisions about which measures to use. Here are four patterns from our discussion:

As demonstrated by Graduate School, fairness through unawareness and equalised odds are liable to miss cases of unfairness where qualifications are caused by protected attributes.

As evidenced by Jobs and Jobs 2, fairness through unawareness and equalised odds misdiagnose cases where protected attributes are good evidence of qualifications.

While counterfactual fairness can at least sometimes correct for these shortcomings, it has its own flaws. As Law School Success 2 shows, counterfactual fairness can miss unfairness where "qualifications" are unfair because, for instance, they are not under subjects' control.

As Internship demonstrates, counterfactual fairness can--much like fairness through unawareness and equalised odds--misdiagnose cases where protected attributes are good evidence of qualifications.

**7. References**
Arneson, Richard, "Equality of Opportunity", The Stanford Encyclopedia of Philosophy (Summer 2015 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2015/entries/equal-opportunity/>.

Cohen, G. A. (2008). Rescuing Justice and Equality. Cambridge, MA: Harvard University Press.

Grgic-Hlaca, Nina, Zafar, Muhammad Bilal, Gummadi, Krishna P, and Weller, Adrian. The case for process fairness in learning: Feature selection for fair decision making. NIPS Symposium on Machine Learning and the Law, 2016.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In Advances In Neural Information Processing Systems, pages 3315–3323.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. In Advances in Neural Information Processing Systems, pages 4066–4076.