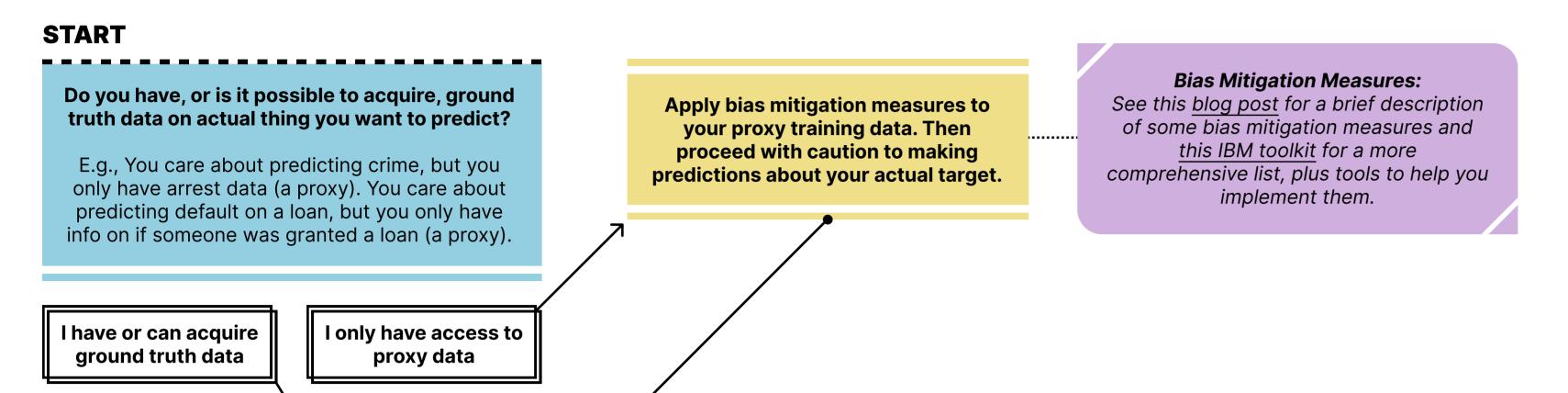
Which type of statistical fairness should you strive for?

You've got a machine learning system. You want it to be fair. But there are so many ways to be fair! Which should you choose?

By Samara Trilling and Madison Jacobs



RELEVANT DEFINITIONS

DISPARATE TREATMENT

Disparate treatment is a legal term defined as negative treatment of a loan candidate or group of loan candidates due solely to that candidate's protected status (race, ethnicity, gender, etc).

DISPARATE IMPACT

Disparate impact is a legal term defined as unintentional but systemic negative treatment of a protected group of loan candidates... but because ML models lack a human decision maker to ask about their intent or reasoning, it's not always clear how disparate treatment and impact should apply to algorithms. Regulators should clarify this.

Credit to:

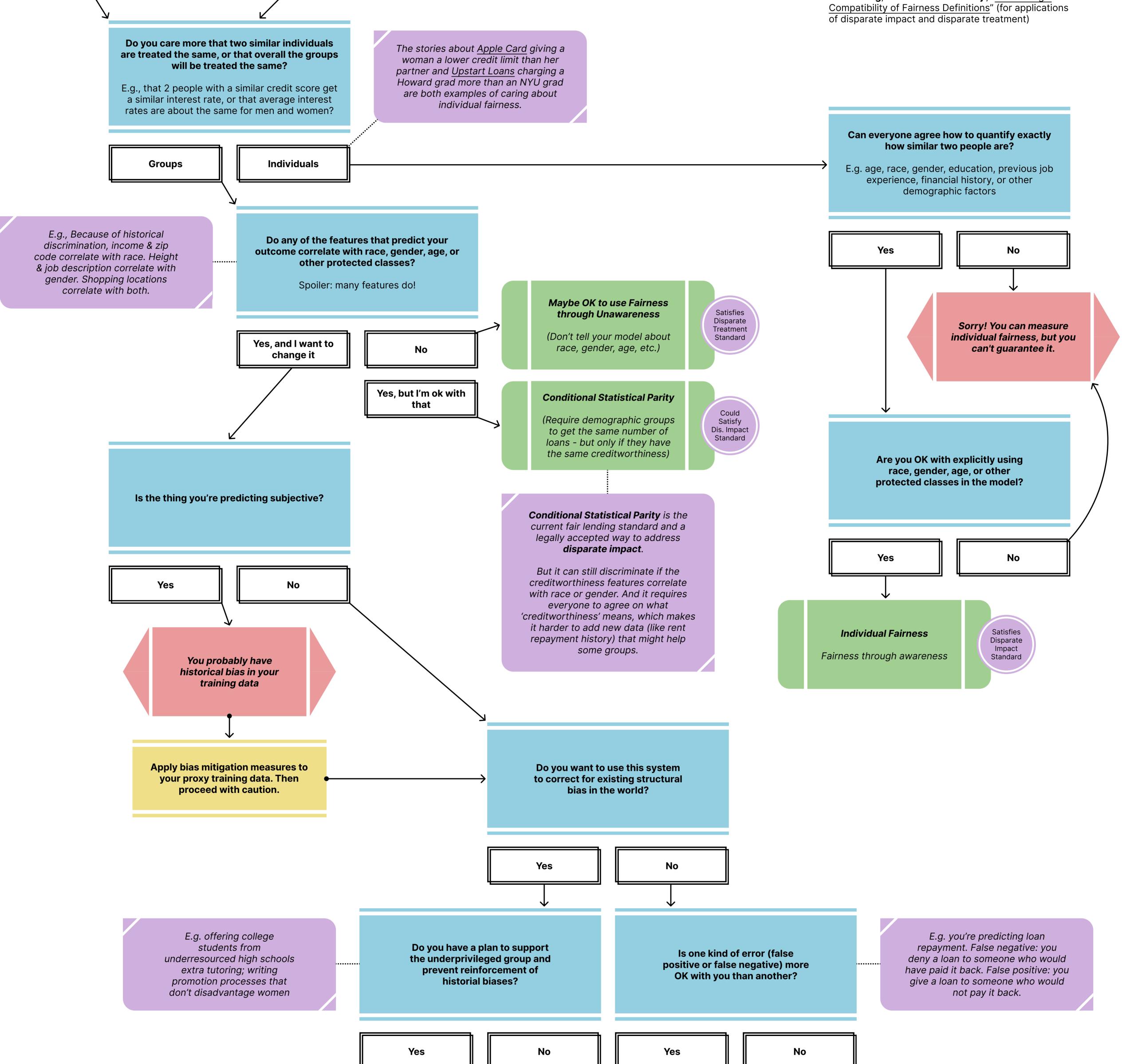
Valeria Cortez, "How to define fairness to detect and prevent discriminatory outcomes in Machine Learning" (for many good examples of when to use each type of fairness)

Ziyuan Zhang, "<u>A Tutorial on Fairness in Machine</u> <u>Learning</u>" (for examples of controversies around different types of fairness)

Moritz Hardt, **Eric Price**, **Nathan Srebro**, "<u>Equality of</u> <u>Opportunity in Supervised Learning</u>" (for comparisons of equality of opportunity and odds)

Solon Barocas, Moritz Hardt, Arvind Narayanan, Fair <u>ML Book</u> (for predictive parity)

Alice Xiang, Inioluwa Deborah Raji, "On the Legal



Use Equal Opportunity **Use Equalized Odds** Could Satisfy Dis. Impact Could Satisfy Your model will probably Satisfies **Use Demographic Parity** Equal false negative rate perpetuate or exacerbate Equal false negative rate Disparate Dis. Impact Impact historical biases and Standard Standard Standard Equal false positive rate false positive rate E.g., You're predicting job COMPAS uses predictive parity WHY DO I HAVE TO PICK? performance. If your workplace is (which requires that the model is hostile to women, even if you hire lots "equally good at predicting whether a Predictive parity, demographic parity, and white or black defendant classified as of them, they might get more unfair equalized odds are mutually exclusive—you bad performance reviews and skew high risk would reoffend"). This can't satisfy more than one. (Except in specific resulted in a higher rate of false future training data such that women cases: E.g., if both groups are *actually* look worse at their jobs, which makes positives for people of color than equally likely to default, then you can satisfy it less likely that other women will white people. Some have argued both demographic parity and equalized odds). score well on predictive hiring they should be using equal Read more here. opportunity instead. assessments.