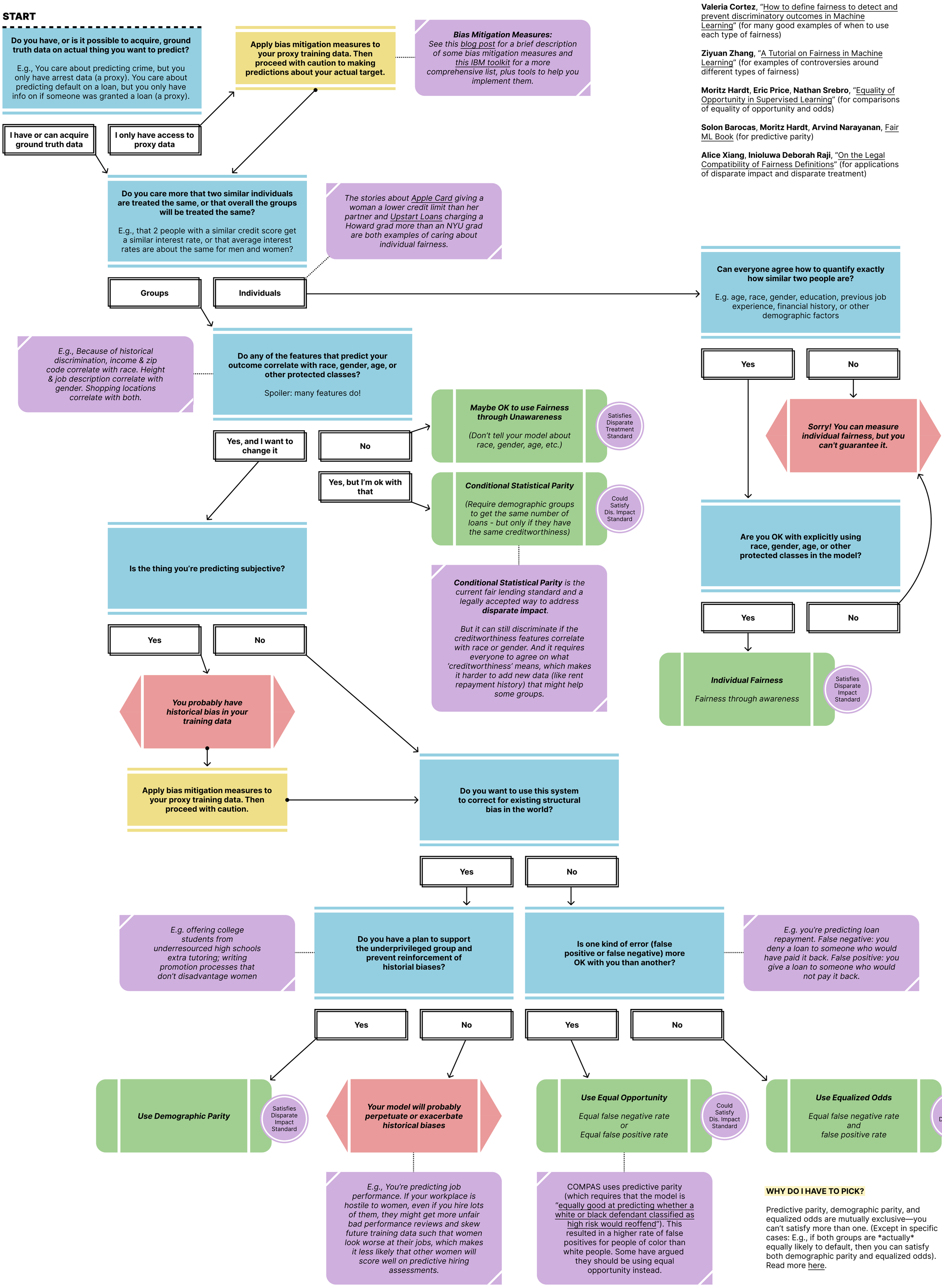
Which type of statistical fairness should you strive for?

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, "A Tutorial on Fairness in Machine Learning" (for examples of controversies around different types of fairness)

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

Solon Barocas, **Moritz Hardt**, **Arvind Narayanan**, Fair ML Book (for predictive parity)

Alice Xiang, **Inioluwa Deborah Raji**, "On the Legal Compatibility of Fairness Definitions" (for applications of disparate impact and disparate treatment)

WHY DO I HAVE TO PICK?

Predictive parity, demographic parity, and equalized odds are mutually exclusive—you can't satisfy more than one. (Except in specific cases: E.g., if both groups are *actually* equally likely to default, then you can satisfy both demographic parity and equalized odds). Read more here.