On the Moral Justification of Statistical Parity

Corinna Hertweck[1,2], Christoph Heitz[1] and Michele Loi[2]
[1] Zurich University of Applied Sciences, [2] University of Zurich

Imagine you are in charge of admissions to a prestigious university: When and why should low- and high-income students be accepted at the same rate?

1. Introduction

Artificial intelligence is increasingly influencing every aspect of our lives. Automated Decision Making (ADM) systems use machine learning to decide who should receive a loan, who gets admitted to university or who is invited for a job interview. With such systems influencing our lives at such a scale, it's natural to ask whether these systems are fair and how we can check whether they are fair. The topic has been discussed in philosophy and law for thousands of years, so the field of algorithmic fairness aims to apply these discussions to ADM systems. Our work discusses the philosophical reasoning behind one of the most popular mathematical definitions of fairness ("statistical parity"). In essence, we attempt to lay the foundation for answering the question "What philosophical justification is there to enforce the metric statistical parity in a given application?"

2. Statistical Parity

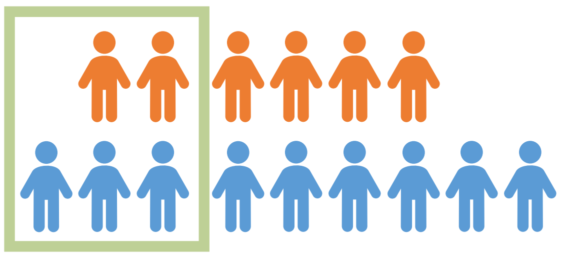
Statistical parity demands that the probability of getting a certain decision (e.g., being admitted to a prestigious university) is the same across different socio-demographic groups (i.e., the metric is „fulfilled" if the probability of being admitted to a prestigious university is the same for low- and high-income students).

In the illustration below, 1/3 of the high-income students (in orange) are admitted to university. The same share has to be admitted for low-income students (in blue).

Acceptance rate

High-income: 1/3

Low-income: 1/3



Admitted

4. Rule

When justifying statistical parity *with repsect to the people affected by the decisions* (i.e., university applicants), it appears as if the rule for when to enforce statistical parity should be:

IF AND ONLY IF
- the groups are equal in the potential space **AND**
- there is measurement bias **OR** unjust life's bias,

THEN
- statistical parity should be enforced

5. Counterexamples

Next, we check whether this rule works as a universal rule. The idea of such a universal rule would be that one can check whether the conditions hold in any given context. If so, statistical parity should be enforced, otherwise not. When looking for counterexamples, we found counterexamples in both directions:

- Preconditions are fulfilled, but statistical parity should not be fulfilled
- Preconditions are not fulfilled, but statistical parity should be enforced

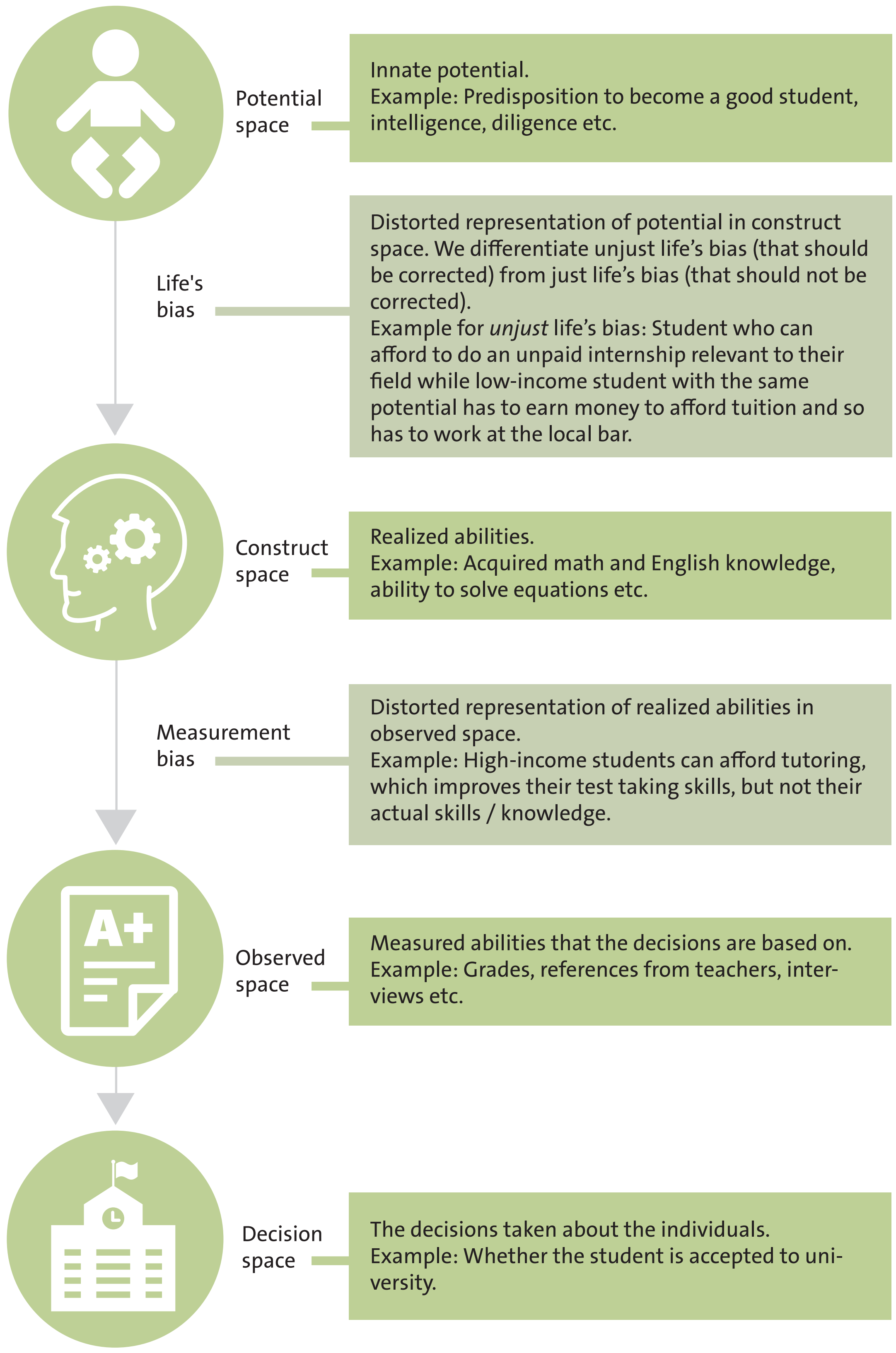
Our counterexamples are based on the fact that the rule does not consider the consequences of the decisions that have been taken – it only considers past and current injustices in the form of measurement and life's bias. The rule will thus have to be extended in the future to cover all cases.

References

[1] Friedler, Sorelle A., Carlos Scheidegger, and Suresh Venkatasubramanian. „On the (im)possibility of fairness." arXiv preprint arXiv:1609.07236 (2016).

3. Framework

We build on the framework proposed by [1]. The authors propose a way to think about how differences between socio-demographic groups come to be. Their framework divides the process through which we arrive at decisions, e.g., university admission decisions, into different "spaces" and "biases". Our contribution is the addition of the *potential space* and *life's bias*.



6. Discussion

- At least two types of biases justify statistical parity from the view of decision subjects: measurement bias and unjust life's bias
- However, considering these biases alone is still insufficient
- Measurement and life's bias do not consider the consequences of the decisions
- Consequences of decisions should be considered in future frameworks

Contact:
Corinna Hertweck
corinna.hertweck@zhaw.ch

Learn more about our work

Zürcher Hochschule für Angewandte Wissenschaften
zhaw School of Engineering
IDP Institut für Datenanalyse und Prozessdesign

Universität Zürich UZH

FNSNF Swiss National Science Foundation