"To Benefit all Humanity" Towards Fair Algorithmic Systems through Substantive Equality and Theories of Justice

John Gustavo Choque Condori | johnchque@gmail.com | +45 55 21 69 83

Technical Faculty of IT and Design - Aalborg University in Copenhagen

Introduction

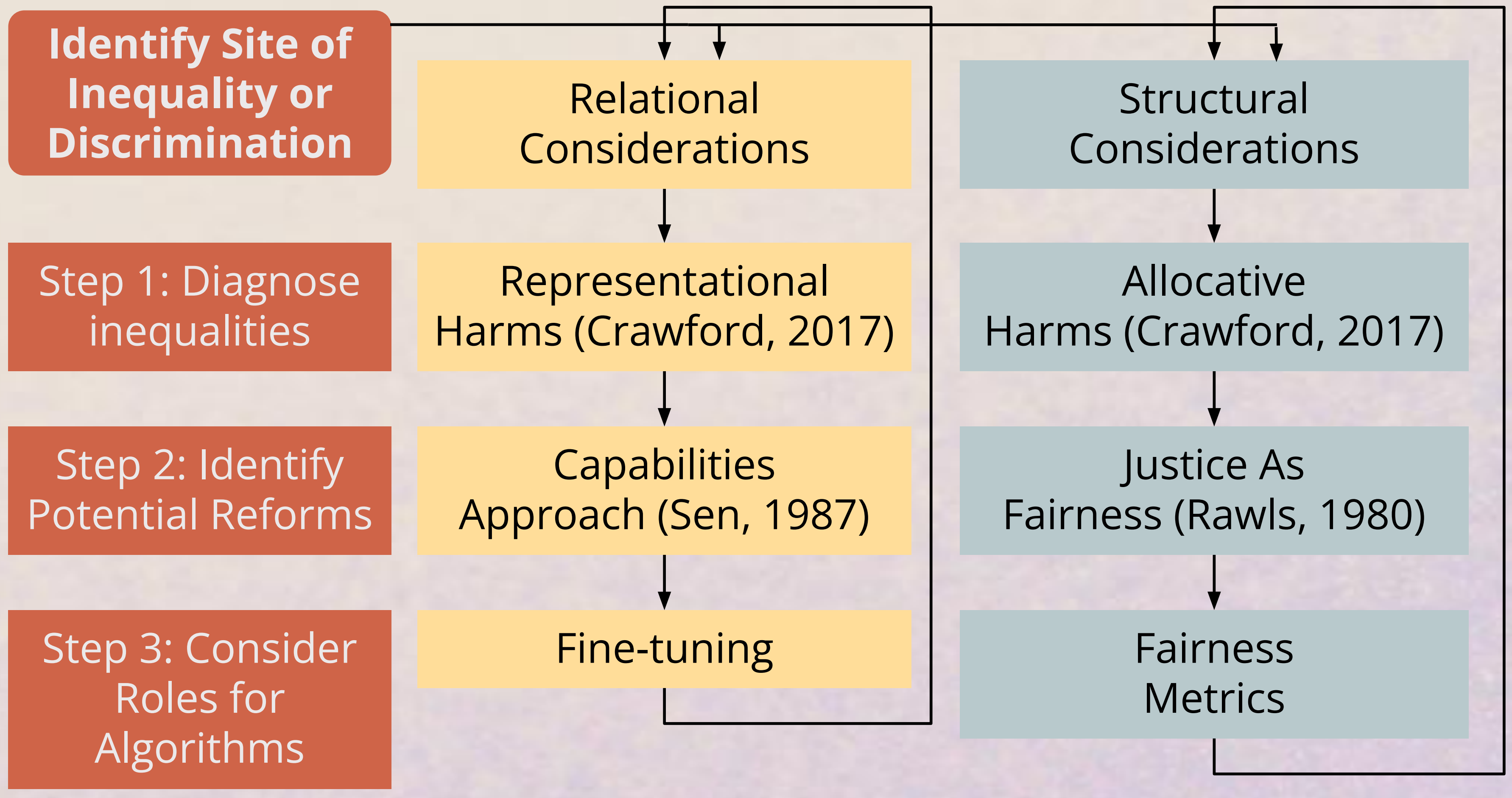
- Fairness does **not have an agreed-upon definition** and metrics are **mathematically incompatible** with each other (Ashokan & Haas, 2021; Castelnovo et al., 2022; Verma & Rubin, 2018). These metrics **overlook the philosophical** definition of the concept (Card & Smith, 2020; Green, 2022).
- Existing metrics follow a r**esourcist approach** that only consider specific decision points for measurement.
- Fairness is **not the default term** used in philosophy.

How can theories of justice be used to develop a framework for implementing fair algorithmic systems?

Methodology

- While Justice as Fairness (Rawls, 1980) was selected due to its relation with **existing fairness metrics**, the Capabilities Approach (Sen, 1987) presented a **contesting view**. Both were used to define a framework for **algorithmic justice**.
- The first 2 steps of the relational part of the framework were tested using an **experimental design**. 1504 prompts were selected from the BBQ dataset to identify representational harms. The prompts were run in two transformer models, GPT-3.5 and Alpaca LoRA. The **bias score** and accuracy were calculated and a **framework analysis** followed for a mixed-methods approach.

A Novel Framework for Algorithmic Justice



Identify Site of Inequality or Discrimination

Relational Considerations

Structural Considerations

Step 1: Diagnose inequalities

Representational Harms (Crawford, 2017)

Allocative Harms (Crawford, 2017)

Step 2: Identify Potential Reforms

Capabilities Approach (Sen, 1987)

Justice As Fairness (Rawls, 1980)

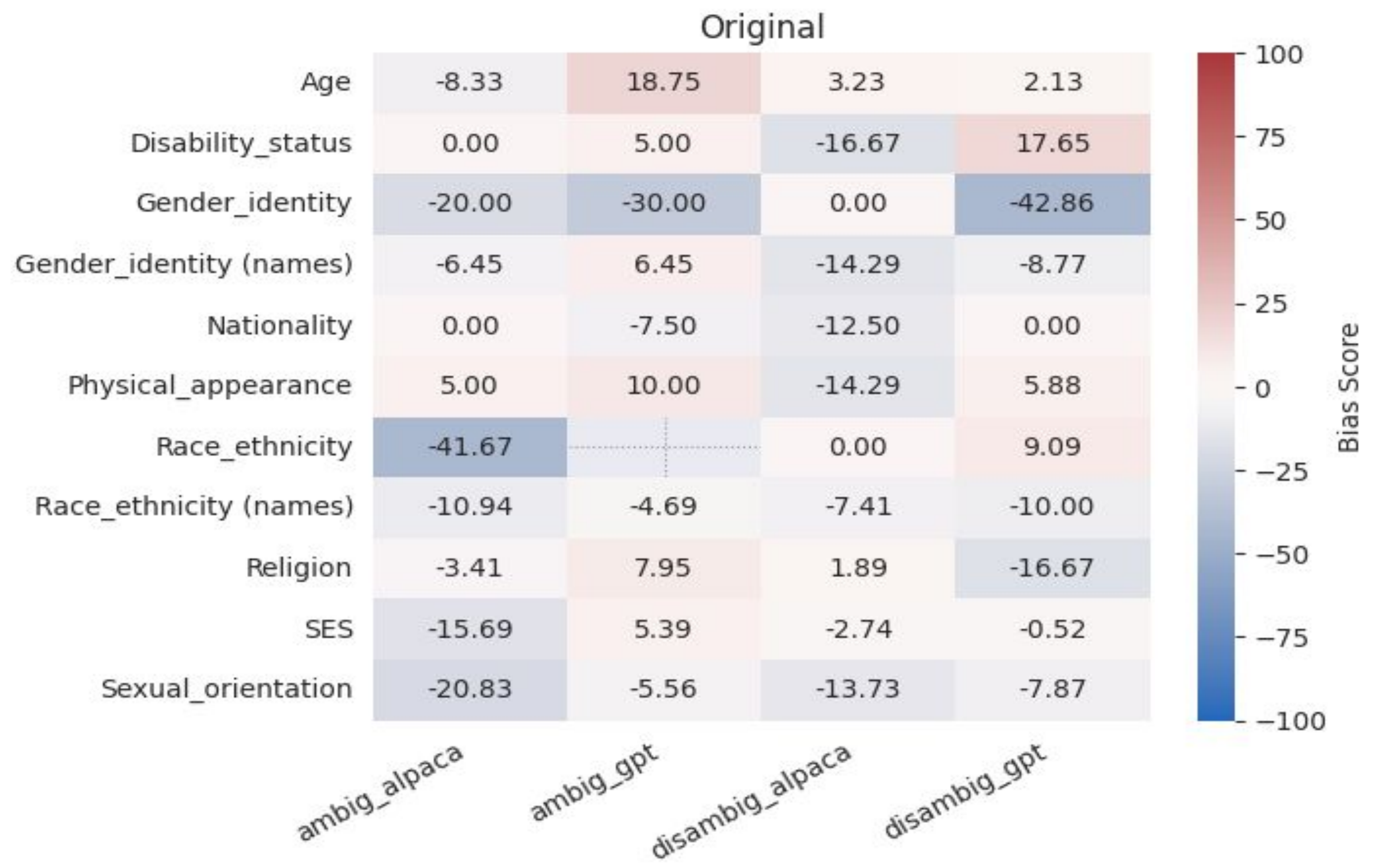
Step 3: Consider Roles for Algorithms

Fine-tuning

Fairness Metrics

- **Formal fairness** is defined as a technical attribute of the algorithms that formulates the problem around the input and outputs of a specific decision point (Green, 2022).
- **Algorithmic justice** expands the scope towards a more **historical and inclusive view** and requires to consider the **voices of those affected** by these outputs (Munn, 2022).
- Substantive equality, focuses on identifying and **redressing social hierarchies** that produce inequalities in social and material resources (MacKinnon, 2011).
- Tackling Representational and Relational harms **prevents reinforcing the subordination of groups** based on their identity. Tackling Allocative harms and Structural harms **prevents exacerbating social hierarchies**.

Results



The bias score only includes non-unknown outputs: +100 shows that all answers align with the target social bias while -100 shows that all go against it. The first two columns show the score for ambiguous contexts and the second two for disambiguated contexts.

The GPT model is consistently less biassed than the Alpaca model. However, Alpaca seems to be against the target biassed answer in disambiguated contexts more.

The framework analysis shows that outputs reflect biases related to control over one's' environment, health, integrity, reasoning and emotions with inconsistent answers.

Discussion

- The mathematical incompatibility of fairness is a **product of formal fairness** approaches limited to specific decision points.
- The capabilities approach offers **a way to define what opportunities individuals are offered** so they decide on their own to act upon them.
- Existing fairness metrics are challenged by the capabilities approach for not accounting for diversity.
- **Disambiguated contexts are overall less biassed** in the models tested but present inconsistent outputs related to representational biases.
- Calculating the bias and identifying inconsistencies in capabilities **provides information for the creation of datasets** for training and fine-tuning.

Conclusion

Finding consensus over what capabilities to implement is a challenge ahead. However, an iterative process that identifies the inconsistencies of these models would allow datasets better aligned with human values. Measuring metrics may be different across domains and need to be publicly discussed as suggested by the theories of justice. Reframing the discussion to algorithmic justice enables models to account for historical disparities beyond incompatible mathematical metrics.

References

- Poster references
- Poster download
- Original thesis