Perception as a Fairness Parameter

Jose M. Alvarez
University of Pisa &
Scuola Normale Superiore
Pisa, Italy
jose.alvarez@sns.it

Mayra Russo
L3S Research Centre &
Leibniz Hannover Universität
Hannover, Germany
mrusso@l3s.de

Perception refers to the process by which two or more agents make sense of the same information differently [5]. Largely unexplored within the Fair AI literature, in this work, we consider perception as a parameter of interest for fairness problems and present the *fair causal perception* (FCP) framework. FCP allows for an algorithmic decision-maker h to elicit group-specific representations, or *perceptions*, around a sensitive attribute A to enhance the information set $\mathbf{X}$ used for calculating the decision outcome $h(\mathbf{X}) = \widehat{Y}$. It combines ontologies [2] and structural causal models [4], by using the former to semantically enrich the latter [1]. Under FCP, we can enrich A by operationalizing two distinct processes around it: *categorization* and *signification* [3]. Categorization entails sorting instances into categories, while signification entails an interpretive act in which we represent the social meanings relevant to the decision context. This act of interpretation becomes available to h by introducing a set of perceptions around A that both signify and prescribe additional information on how belonging to A can affect $\mathbf{X}$ beyond what is recorded. We envision an h that can choose to reinterpret $\mathbf{X}$ using A-specific perceptions depending on its fairness goals, meaning that it is possible for the same individual instance to be classified differently by h depending on the evoked representations. We showcase our framework using a college admissions problem. In the case of a tie between two observably alike candidates with different socioeconomic backgrounds A, h can non-randomly break the tie in favor of the under-privileged candidate using FCP. With FCP we can describe what it means to be a candidate from the under-privileged group and, in turn, how it causally affects a candidate's SAT scores $\mathbf{X}$ by describing local penalties to be introduced by h when comparing these candidates. Although, under FCP, h deviates from the individual fairness notion of treating similar individuals alike, it can also lead to fairer results as under-privileged candidates can be prioritized when all else is equal. We will test this hypothesis and evaluate our proposed framework by assessing our use case against individual fairness benchmarks.

Acknowledgments and Disclosure of Funding

We thank our supervisors Salvatore Ruggieri and Maria-Esther Vidal for the advice and support. This work received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie Actions (GA number 860630) for the project "NoBIAS".

References

[1] E. Blomqvist, M. Alirezaie, and M. Santini. Towards Causal Knowledge Graphs. In *KDH@ECAI*, volume 2675 of *CEUR Workshop Proceedings*, pages 58–62. CEUR-WS.org, 2020.

[2] S. Grimm. Knowledge representation and ontologies. In *Scientific Data Mining and Knowledge Discovery*, pages 111–137. Springer, 2010.

[3] G. Loury. Why does racial inequality persist? Culture, causation, and responsibility. *The Manhattan Institute*, 2019.

[4] J. Pearl. *Causality*. Cambridge University Press, 2nd edition, 2009.

[5] A. Tversky and D. Kahneman. Extensional versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment. *Psychological Review*, 90(4):293, 1983.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).