Are algorithms necessarily biased? The case of causal search algorithms

Abstract (500 words, excluding references):

Science and technology studies (STS) and philosophy of technology (PT) scholars understand algorithms as mathematical tools that are given imperatively and implemented to facilitate or replace human decision-making (Hill 2015). They can point to a long list of "biased" or "value-laden" algorithms that have led to ethically problematic decision-making in criminal justice (Angwin et al 2016), medicine (O'Reilly-Shah et al 2020), computer vision (Klare et al 2012), hiring (Garcia 2016) etc. Some survey studies conclude that algorithmic bias or value-ladenness is "inescapable", "inevitable" (Mittelstadt et al 2016) or "systemic" (Drozdowski et al 2020). A popular explanation is that algorithms are specified by developers of configured by users with desired outcomes in mind (Diakopoulos 2015).

The paper takes issue with the modality involved in this conclusion. It claims that desired outcomes do *not* necessarily influence the specification or configuration of algorithms, and that algorithms are *not* necessarily biased or value-laden. The paper derives its claim from the fact that "debiasing" or mitigating approaches (Zhou et al 2022) involving various "fairness" metrics (Mitchel et al 2019) wouldn't make sense if algorithmic bias and value-ladenness were necessary. The paper also derives its claim from a case study of algorithms that receive little attention in the STS and PT literature, even though they are widely used in the sciences to make predictions under interventions: causal search algorithms.

Causal search algorithms proceed by operating on data sets that provide values for preselected variables, by forming complete undirected graphs connecting these variables, and by testing for conditional independence relations to eliminate unnecessary edges and to direct the remaining ones (Spirtes et al 2001). The success of this procedure depends crucially on the satisfaction of the assumption of the absence of confounders: an arrow departing from *X* and directed into *Y* will fail to stand for a relation of causal dependence if both *X* and *Y* causally depend on a (confounding) variable *Z* that is not included in the set of preselected variables. The selection of variables might be thought to be the (most plausible) stage, at which values or bias enter the procedure. The paper will argue, however, that values or bias do not necessarily enter the procedure (at this stage).

It will argue, more specifically, that practicing statisticians tend to understand variable selection as "art" or "lore" (Welsh 1986), and that "artful" variable selection is not necessarily biased or value-laden. The paper will also argue that in the sciences, variable selection is governed by important (though defeasible) heuristics or default rules, and that these rules have important connections with principles of causal explanation (Woodward 2016). The rules require that the selected variables

- be well-defined targets for interventions,

- have unambiguous effects on other selected variables,
- can be manipulated independently,
- be causally specific,
- strongly correlate with the variables to which they are causally connected,
- take positions in stable relations of causal dependence,
- take positions in relations that do not include unexplained causal cycles or correlations among exogenous variables or residuals.

References:

Angwin, J. et al (2016). “Machine Bias: there’s software used across the country to predict future criminals. And it’s biased against blacks.” *Pro Publica*.

Diakopoulos, N. (2015). “Algorithmic Accountability: Journalistic investigation of computational power structures.” *Digital Journalism* 3(3): 398-415.

Drozdowski, P. et al (2020). “Demographic Bias in Biometrics: A Survey on an Emergent Challenge.” *IEEE Transactions on Technology and Society* 1(2).

Garcia, M. (2016). “Racist in the Machine: The Disturbing Implications of Algorithmic Bias.” *World Policy Journal* 33(4): 111-117.

Hill, R. K. (2015). “What an algorithm is.” *Philosophy & Technology* 29(1): 35-59.

Klare, B. F. et al (2012). “Face Recognition performance: Role of demographic information.” *IEEE Transactions on Information Forensics and Security* 7(6): 1789-1801.

Mitchel, S. et al (2019). “Prediction-based decisions and fairness: a catalogue of choices, assumptions, and definitions.” https://arxiv.org/pdf/1811.07867.pdf

Mittelstadt, B. D. et al (2016). “The ethics of algorithms: Mapping the debate.” *Big Data & Society* July-December: 1-26.

O’Reilly-Shah, V. N. et al (2020). “Bias and ethical considerations in machine learning and the automation of perioperative risk assessment.” British Journal of Anaesthesia 125(6): 843-6.

Spirtes, P., Glymour, C., Scheines, R. (1993). *Causation, Prediction and Search*. New York: Springer.

Welsh, R. E. (1986). “Comment.” *Statistical Science* 1, 403-5.

Woodward, J. (2016). “The problem of variable choice.” Synthese 193: 1047-1072.

Zhou, N. et al (2022). “Bias, Fairness and Accountability with Artificial Intelligence and Machine Learning Algorithms.” *International Statistics Review* 90(3): 468-80.