The Perils of Objectivity: Towards a Normative Framework for Fair Judicial Decision-Making

Andi Peng
Adaptive Systems and Interaction Group
Microsoft Research
andipeng@microsoft.com

Malina Simard-Halm
Institute of Criminology
University of Cambridge
mjs296@cam.ac.uk

ABSTRACT

Fair decision-making in criminal justice relies on the recognition and incorporation of infinite shades of grey. In this paper, we detail how algorithmic risk assessment tools are counteractive to fair legal proceedings in social institutions where desired states of the world are contested ethically and practically. We provide a normative framework for assessing fair judicial decision-making, one that does not seek the elimination of human bias from decision-making as algorithmic fairness efforts currently focus on, but instead centers on sophisticating the incorporation of *individualized* or *discretionary bias*—a process that is requisitely human. Through analysis of a case study on *social disadvantage*, we use this framework to provide an assessment of potential features of consideration, such as political disempowerment and demographic exclusion, that are irreconcilable by current algorithmic efforts and recommend their incorporation in future reform.

CCS CONCEPTS

• **Social and professional topics** → **Government technology policy**; • **Applied computing** → *Law*.

KEYWORDS

criminal justice; decision-making; risk assessment; bias; fairness

ACM Reference Format:
Andi Peng and Malina Simard-Halm. 2020. The Perils of Objectivity: Towards a Normative Framework for Fair Judicial Decision-Making. In *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES '20), February 7–8, 2020, New York, NY, USA.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3375627.3375869

1 INTRODUCTION

The prison and its penumbra of control have become spectacles of social and economic inequality [1]. The increased deployment of algorithmic risk assessment tools to aide judicial decision-making has been found to consistently reflect—and exacerbate—stereotypes related to race, class, and gender [2, 4]. This indicates an issue and an irony—that the ostensible instruments of "justice" have come in fact to deepen the contours of historic disadvantage. The technical foundations of algorithmic design necessitate the existence of an ideal state for system builders to optimize for—one that does not and, we argue, should not exist in the real world. In this paper, we contend that current risk assessment efforts in criminal justice, including work on algorithmic fairness, unproductively seek to eliminate human bias from decision-making. We offer a normative framework for assessing fair judicial decision-making that acknowledges the value of *discretionary bias*, not its elimination.

We begin by reviewing the history of criminal risk assessment and the failures of earlier systems designed to predict the "dangerousness" of criminals [5]. We detail *selective incapacitation theory* as a flawed ethical framework for assessing risk assessment tools, stipulating the high cost of false positive prediction errors in system adoption [3]. We argue that the present use of algorithmic tools in the criminal justice system represents the next chapter in a long-running trend of risk mitigation which unwisely undermines the spirit, if not the letter, of individual liberty.

We then analyze a growing body of work on algorithmic fairness, which seeks to rectify algorithms' unequal treatment of different groups. We review the challenges faced by the computer science community in optimizing for fairness as a technical metric. Despite developments, we argue that algorithmic attempts at achieving parity will remain ineffectual—and likely counteractive—so long as there is no consensus on a desired world optimum.

Next, we offer a normative framework for assessing what constitutes fair judicial decision-making—one that considers central, and often irreconciable, philosophical principles of justice, procedural legitimacy, and the benefits of judicial discretion. We contend that improved human *discretionary bias*, not its elimination, from decision-making processes is paramount in systems where complex and evolving normative aims exist. We assess the case study of *social disadvantage* to highlight how optimization functions fail to capture features such as oppression and demographic exclusion.

We conclude that only human discretion can fully assail the moral quandaries underwriting the justice system. As such, we offer directions of reform in search of a more comprehensive and interdisciplinary notion of fairness in judicial decision-making—and consider the role that algorithms should play within it.

REFERENCES

[1] Michelle Alexander. 2012. *The new Jim Crow: mass incarceration in the age of colorblindness.* The New Press.
[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: there's software used across the country to predict future criminals, and it's biased against blacks. *ProPublica* (2016).
[3] Harvard Harvard Law Review. 1982. Selective incapacitation: reducing crime through predictions of recidivism. *The Harvard Law Review Association* (1982).
[4] Cathy O'Neil. 2016. *Weapons of math destruction: how big data increases inequality and threatens democracy.* Crown Books.
[5] Peter Scott. 1977. Assessing dangerousness in criminals. *British Journal of Psychiatry* (1977).

AIES '20, February 7–8, 2020, New York, NY, USA

ACM ISBN 978-1-4503-7110-0/20/02.
https://doi.org/10.1145/3375627.3375869