Tech Otherwise • CSCW '20

A Discussion of Harm Reduction for Algorithmic Fairness: Lessons from the Opioid Crisis

Christopher Caulfield

Published on: Oct 16, 2020

DOI: 10.21428/93b2c832.4ec985c8

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

A Discussion of Harm Reduction for Algorithmic Fairness: Lessons from the Opioid Crisis

By Christopher Caulfield

There has been a mushrooming of calls for a more caring and careful design and implementation of algorithms as they have increasingly mediated the experiences and life opportunities of people in diverse areas including healthcare, employment, provision of loans (Dixon & Gellman, 2015), government benefits, and criminal justice (Eubanks, 2018). Amongst these calls, the principles of harm reduction have been enlisted to aid in the theorization of algorithmic fairness, a development this author views as both promising and fraught (Altman, Wood, & Vayena, 2018). The harm reduction approach to public policy and healthcare has a long history in drug treatment of saving lives by focusing on the individual and communal well-being of the most vulnerable and marginalized groups, meeting people where they are at without judgment or coercion, prioritizing the voices of people with lived experience during the development of policies and programs designed to serve them while attending to systemic barriers to care including class, race, gender, and disability (Collins et al., 2012). This paper offers an extension and modification of recent work by Altman et al. which integrates a harm reduction framework into algorithmic design and implementation, but by failing to integrate the voices of those adversely impacted by algorithmic design, does not go far enough to protect the vulnerable and marginalized and fails to incorporate the valuable insights of people with lived experience being processed by algorithms. A fuller harm reduction approach to designing algorithmic fairness makes space for situated ethnographic research of users processed by algorithms and involvement of those users in iterative redesigns.

In outline, Altman et al. prescribe that “algorithmic fairness must take into account the foreseeable effects that algorithmic design, implementation, and use have on the well-being of individuals” (Altman et al., 2018, p. 2). They borrow a process for using counterfactual framework for causal inference from statistics and computer science, and assert that certain patterns of foreseeable harms are unfair, namely, “An algorithmic decision is unfair if it imposes predictable harms on sets of individuals that are unconscionably disproportionate to the benefits these same decisions produce elsewhere. Also, an algorithmic decision is unfair when it is regressive, i.e., when members of disadvantaged groups pay a higher cost for the social benefits of that

decision” (Altman et al., 2018). Importantly, the framework offered by Altman et al. seeks to avoid disproportionate harms on minority groups and group punishment. I extend this harm reduction approach to algorithmic fairness by arguing that in addition to ex-ante statistical modeling of probable harms, algorithmic fairness can be strengthened by ex-post analysis of the harms arising from algorithmic design, prioritizing the voices of people with lived experience using the algorithms that are designed to serve them, a key feature of harm reduction work within the opioid crisis (Drainoni et al., 2019). Ex-post analysis of algorithmic harms would ethnographically collect and prioritize the voices of people who have experienced the algorithm in question, especially the most marginalized and vulnerable including minorities, low-income, and those with disability.

In the case of drug treatment for opioid use, there are numerous examples of successfully using ethnographic research, embedded within affected communities, to uncover and articulate the predictable harms imposed on people by a criminal (in)justice system that imbricates the medical system, enrolling clinicians as agents of surveillance and punishment, placing them on ‘addiction trajectories’ much as an algorithmic logic might place one on an unfair trajectory to imprisonment (Knight, 2015; Raikhel & Garriott, 2013). This paper calls for analogous work in evaluating the actual (not just foreseeable) harmful outcomes of algorithmic design and implementation, by embedding ethnographic researchers into historically marginalized groups of users of the algorithm in question in order to prioritize their uniquely situated voices while designing and implementing algorithms that affect them. Such a richly situated social scientific approach to algorithmic fairness would result not only in fewer disproportionate harms for marginalized and disadvantaged groups, but also a better user experience for the full range of society by improving the reliability of fair outcomes and thereby trust in, and compliance with, the sociotechnical systems of which the algorithms are but one part.

References

- Altman, M., Wood, A., & Vayena, E. (2018). A harm-reduction framework for algorithmic fairness. *IEEE Security & Privacy*, 16(3), 34–45. IEEE.
- Collins, S. E., Clifasefi, S. L., Logan, D. E., Samples, L. S., Somers, J. M., & Marlatt, G. A. (2012). Current status, historical highlights, and basic principles of harm reduction.
- Dixon, P., & Gellman, R. (2015). *The scoring of America: How secret consumer scores threaten your privacy and your future*. *World Privacy Forum*, April 2.

- Drainoni, M.-L., Childs, E., Biello, K. B., Biancarelli, D. L., Edeza, A., Salhaney, P., Mimiaga, M. J., et al. (2019). “We don’t get much of a voice about anything”: Perspectives on photovoice among people who inject drugs. *Harm reduction journal*, 16(1), 61. Springer.
- Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin’s Press.
- Knight, K. R. (2015). *Addicted.pregnant.poor*. Critical global health : evidence, efficacy, ethnography. Durham: Duke University Press.
- Raikhel, E., & Garriott, W. (2013). *Addiction Trajectories*. Duke University Press.