Data-Driven Discrimination and (Dis)Respect

Extended abstract

Sahlgren Otto

Tampere University
otto.sahlgren@tuni.fi

1 Introduction

Data-driven discrimination (or 'algorithmic discrimination') has been established as a central area of legal and ethical concern in the field of artificial intelligence (cf. Barocas & Selbst, 2016; West, Whittaker & Crawford, 2019). Automated decision-making systems utilizing machine learning and statistical models have been found in many cases to discriminate against individuals, disadvantaging people especially on the basis of sensitive traits (or legally protected characteristics), such as 'race'. To mention a paradigmatic example, the COMPAS system used by U.S. courts to predict recidivism was shown to systematically favor white defendants over black defendants in its predictions (cf. Dressel & Farid, 2018).

There is, however, considerable discrepancy in the public and scholarly discourse on both (i) how discrimination may arise in data mining and use of automated decision-making systems (cf. Barocas, 2014), and (ii) what specifically makes a given instance of data-driven discrimination morally objectionable. Philosophical theories concerning the ethics of discrimination have long examined these questions in the abstract. This study explores the connections between the discourse on data-driven discrimination and existing theories of ethics of discrimination – those grounding the moral wrongness of discrimination in the more general wrong of disrespect, in particular (cf. Eidelson, 2015; Hellman, 2008).

A special focus is on Eidelson's (2015) pluralist theory in which disrespect is understood as a deliberative failure to recognize the normative weight of a person. According to this theory, discrimination is intrinsically wrong, if it is, when an agent engaging in discrimination will fail to treat a person as equal in value in comparison to others or undermine her autonomy in the process. Some discrimination – prominently, instances of statistical discrimination (cf. Schauer, 2017) – will be contingently wrong, if at all. Specifically, statistical discrimination will be wrong in virtue of the broad harms it (re)produces, such as stigmatization, alienation and segregation. Moreover, differing from theories that identify only discrimination on the basis of socially salient traits as wrongful discrimination, if at all (cf. Lippert-Rasmussen, 2006; 2014), Eidelson's view involves no such notion. What matters is whether attending to a given trait in engaging in discrimination constitutes disrespect for an individual's personhood.

This study addresses two questions. First, what types of discrimination can be involved in data mining and automated decision-making? As is noted in the literature, discriminatory harms may arise via different routes in data mining and use of automated decision-making systems (cf. Barocas & Selbst, 2016). Specifying these routes will help in determining points of intervention. Second, should paradigmatic cases of 'algorithmic discrimination' be identified as instances of discrimination involving disrespect? For example, if a facial recognition program misclassifies faces of people of color disproportionately in comparison to others, are the former wrongfully discriminated against? If one adopts a deliberative failure conception of disrespect, the answer will hinge on whether an agent has weighed the interests of the affected individuals sufficiently in conducting such treatment, showing due regard for their standing as autonomous persons.

2 Conclusions

This account entails that one should not focus only on legally protected attributes in considering (wrongful) discrimination in data mining and automated decision-making. We should pay attention, rather, to whether and how human dignity is given due regard in the process. It is argued that, if one adopts Eidelson's view, the use of sensitive data should not be understood as prima facie morally objectionable. This may seem problematic, as so-called algorithmic bias may reflect disproportionalities between groups that are a result of structural discrimination and historical oppression. When structural disadvantage is (re)produced, however, statistical discrimination may be wrongful irrespective of the use of sensitive information.

Automated decision-making is taken by some to inherently disrespect a person by considering only "what correlations can be shown between an individual and others" (Kaminski, 2019, p.10). In other words, automated decision-making allegedly fails to respect a person's individuality. Drawing on Eidelson's account (2015), it is argued that respect for individuality in automated decision-making necessitates using data that reflects how people have exercised their autonomy and acknowledging the limits to predicting individuals' behavior. Automated decision-making will be disrespectful in virtue of undermining one's autonomy if due regard is not shown for individual histories or if a person's future actions are taken to be "determined by statistical tendencies" (Eidelson, 2015, p.148). However, it is argued that while automated decisions may be disrespectful of individuals' autonomy in this sense, they will not likely be discriminatively so. Lastly, some possible shortcomings of this account are considered.

The study contributes to the discussion on discrimination in data mining and automated decision-making by providing insight into both how discrimination may take place in novel technological contexts and how we should evaluate the morality of automated decision-making in terms of dignity and respect.

Acknowledgements

Assistance provided by Professor Arto Laitinen is greatly appreciated. I also extend my thanks to everyone involved in the research project ROSE (Robots and the future of welfare services).

References

1. Barocas, S. (2014). Data mining and the discourse on discrimination. Data Ethics Workshop, Conference on Knowledge Discovery and Data Mining.
2. Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104, 671– 732.
3. Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. Science advances, 4, 1.
4. Eidelson, B. (2015). Discrimination and Disrespect. Oxford: Oxford University Press.
5. Hellman, D. (2008). When is Discrimination Wrong? Oxford: Oxford University Press.
6. Lippert-Rasmussen, K. (2006). The badness of discrimination. Ethical Theory and Moral Practice, 9, 2, 167-185.
7. Lippert-Rasmussen, K. (2014). Born Free and Equal?: A Philosophical Inquiry into the Nature of Discrimination. Oxford: Oxford University Press.
8. Schauer, F. (2017). Statistical (and Non-Statistical) Discrimination. In K. Lippert-Rasmussen (ed.), The Routledge Handbook of the Ethics of Discrimination (pp. 42–53). New York: Routledge.
9. West, S.M., Whittaker, M., & Crawford, K. (2019). Discriminating Systems: Gender, Race and Power in AI. AI Now Institute. Retrieved from https://ainowinstitute.org/discriminatingsystems.html.