**NEWSLETTER**
OCTOBER 2019
ISSUE 125



# FAIRNESS OF AI FOR PEOPLE WITH DISABILITIES: PROBLEM ANALYSIS AND INTERDISCIPLINARY COLLABORATION

Jason J.G. White,Educational Testing Service, Princeton, NJ, USA, jjwhite@ets.org

## ABSTRACT

There are several respects in which recent developments in machine learning-based artificial intelligence pose challenges of fairness for people with disabilities. In this paper, some of the central problems are identified, and briefly reviewed from a philosophical perspective motivated by a broad concern for social justice, emphasizing the role of ethical considerations in informing the problem analysis.

## 1. INTRODUCTION

The opportunities and risks posed by AI, particularly machine learning, for people with disabilities demand an unprecedented collaboration among researchers and practitioners with expertise in fields including applied ethics, human rights law, disability studies, privacy, and artificial intelligence. The response to AI thus demands formation of communities that depart from the configurations of interest which have so far operated at the intersection of technology and disability, and which have predominantly focused on problems of accessibility.

Equally, there is a need to take advantage of commonalities in both the nature of the problems encountered and in potential solutions with other communities who risk being marginalized by algorithmic bias. To facilitate this effort, analyses of the problems should be developed that clearly delineate similarities and differences of the position of people with disabilities with respect to other populations, thus establishing a framework for cooperation and intervention in wider efforts to shape development and implementation of AI in the service of social justice. In the discussion that follows, observations are put forward that contribute to the problem analysis from a broadly philosophical standpoint, accompanied by brief comments regarding relevant current developments in the field.

## 2. PROBLEMS OF UNDERREPRESENTATION

The performance of machine learning systems is shaped by their training data. If specific populations are underrepresented in the training data, then the results of applying a machine learning model may be inaccurate. For example, if individuals with speech-related disabilities are not taken into account in training a speech recognition system, its performance in recognizing their utterances is likely to be poor. This may be regarded as a similar problem to that encountered by linguistic minorities. Apart from the social and technical challenge of accumulating sufficient data from small and diverse populations, and of developing learning algorithms that are less data demanding, there arise fundamentally important questions of privacy. Clearly, obtaining data from people with disabilities is a necessary element of any strategy to achieve fairness for these populations in the operation of machine learning models. This data collection is often necessary in order to make such systems accessible and usable. However, in many instances, these data are also apt to disclose a person's disability status, whether directly or indirectly, creating opportunities for discriminatory behavior if knowledge of a person's disability is misused [11]. This risk amplifies the difficulty suggested by Lazar et al. [7] that the harnessing of technology to overcome practical challenges created by living with a disability may also increase vulnerability to privacy violations, contrary to a human right recognized in international law. Moreover, some individuals, perhaps indeed due to their disability, are unable to provide voluntary and informed consent to data collection, or may at least encounter great difficulty in understanding the choice presented to them [1].
Ethically informed guidance is needed in such cases. More generally, the ethics of achieving appropriate privacy protections while ensuring fairness by enabling people with disabilities to be sufficiently represented in training data merit further investigation. As is true of the population in general, individuals with disabilities can be expected to benefit from privacy-preserving technical measures such as differential privacy, federated learning, and homomorphic encryption, as long as they are thoughtfully and effectively implemented.

Researchers and practitioners with disabilities also need to be engaged in the design, development and application of AI-based technologies, whether in the formulation of standards and policies, or in contributing to individual implementations. Failure to involve people with disabilities (for example in participatory design and evaluation efforts) constitutes a second form of underrepresentation. Although this problem occurs in technological projects generally, its significance in the current context should not be underestimated.

[1] *A recent study concluded that the textual complexity of privacy policies of popular web sites already exceeds what average readers can reasonably be expected to understand. Those with cognitive disabilities affecting reading comprehension may be disadvantaged even further.*

## 3. MAKING DECISIONS ABOUT PEOPLE WITH DISABILITIES

The use of machine learning systems to automate decision-making tasks that would previously have been subject to

human judgment is a matter of current concern among researchers and in public discourse. These AI applications, especially as used in the public sector, can also be regarded as contributing to a trend that has been described as treating code as law - of substituting algorithmic processes for the more contextually sensitive and interpretive judgments that occur in the application of legal rules by human adjudicators [4]. Biases in the design of machine learning models and in training data can result in decisions that discriminate against specific categories of persons, including categories based on disability. Indeed, algorithmic decision-making may compound societal injustices encountered elsewhere. For example, a person's history of prolonged unemployment may be attributable, in part, to discrimination on grounds of disability; and it may inappropriately influence the output of an algorithm for screening job applications, thus exacerbating the person's disadvantage.

In connection with these manifestations of algorithmic bias, people with disabilities have much in common with other populations who are at risk of unjust treatment, specifically on grounds of gender, race, national or ethnic origin, or other categories. The challenge, then, is to address the problem in general terms, while understanding and appropriately taking into account the specific ways in which people with disabilities are affected, including identification of the variables correlated with disability that may induce biases in machine learning systems. Amid this analysis, the overarching question of which decisions it is morally appropriate to automate should remain at the forefront of the discussion.

A recent proposal by Altman et al. [1] for evaluating the fairness of algorithms used in decision-making illustrates the combination of empirical and normative problems that justify a philosophically informed interdisciplinary approach. The framework calls for a counterfactual analysis of the effects on human well-being across the life course of various situations (for example, different choices in the design or implementation of an algorithm, and different roles for human judgment in making the decision), with a view to over all harm reduction. Application of this approach in the context of people with disabilities would impose a substantial burden of foresight upon evaluators, requiring them to judge the harms and benefits that would be incurred in the lives of very diverse individuals under multiple counterfactual assumptions. The empirical difficulties are magnified by the heterogeneous experiences and life circumstances of people with disabilities, which introduce a risk of making inaccurate simplifying generalizations based on group membership. This observation supports the conclusion by Trewin [11] that diversity among people with disabilities poses a challenge for achieving fairness in machine learning systems.

The examples developed in Altman et al. [1] in the context of evaluating recidivism risk in criminal justice fall short of capturing the full complexity of relevant counterfactual scenarios. For instance, in comparing algorithmic decision-making with the alternative of human adjudication, the authors note the relevance of research on human fallibility, and data concerning the error rate of judges. In the parallel case of disclosure of disability status in applications for employment, the human error rate appears to be substantial[2], suggesting that use of an algorithm capable of delivering better than typical human performance on this measure would be justified on grounds of fairness. However, one should also consider the effects of interventions designed to reduce prejudice in human decision-making, such as introducing disability awareness training. As the relevant counterfactual situations are multiplied and the available evidence becomes more sparse, the credibility of the counterfactual analysis is reduced.

Furthermore, disability creates significant differences in the capacities of individuals to convert social goods, including wealth and income, into actual well-being, and leads to variations in a person's need for resources. These differences complicate the interpersonal comparisons required by the counterfactual analysis, and, according to Nussbaum [9, § 3 V], are among the reasons why disability poses a profound challenge to leading theories of social justice, notably the contractarian account proposed by Rawls [10]. Nussbaum argues that the relative position of individuals in society cannot be reduced to numerical measures approximated by wealth and income. For example, she maintains, the redesign of the built environment needed to make it accessible to people with physical disabilities, ensuring them dignity and self-respect, is not equivalent to and cannot be addressed by solving a problem of income or wealth inequality.

The counterfactual framework for algorithmic fairness also depends on the controversial moral assumption that fundamental conditions for human well-being can be traded off among individuals and groups in society by applying an analysis that strives to achieve an equitable distribution of harm and benefit that avoids disproportionately disadvantaging minority groups. In developing a form of the capabilities approach to social justice, Nussbaum [9, § 3 V] warns of the limits of such trade-offs, maintaining that fundamental capabilities needed for a fully human life are not fungible: deficits in one aspect cannot be compensated for by benefits in another. Philosophical criticisms of cost-benefit analysis echo this concern, emphasizing not only the problem of incommensurability of costs and benefits, but also the question of whether interpersonal aggregation (weighing harm to one person or group against benefit to another) is morally defensible [6]. A full treatment of the issues raised here lies beyond the scope of the present paper.

## 4. WHOSE PROBLEMS DOES AI SOLVE?

A crucial question of fairness that is quite independent of issues of algorithmic bias concerns whose problems are addressed by emerging AI technologies, and whose are marginalized. The resurgence of public interest in AI in recent years has been driven by advances in deep neural networks that have a growing variety of applications. Many of the most striking successes are in fields that have great potential to enhance the autonomy of people with disabilities and to overcome access barriers, for example, computer vision and natural language processing [8]. There is a risk, however, of devoting inadequate resources to the development of applications of machine learning which meet the needs of people with disabilities, and of failing to invest sufficiently in the necessary research and collaboration to actualize the promise that AI offers. In effect, addressing other societal problems while devoting comparatively little effort to creating AI applications that satisfy needs deriving from disability would arguably itself constitute a form of injustice. Broadening the discussion beyond issues of discrimination and accessibility, narrowly conceived, and asking how AI can contribute for example to the health, employment, education and welfare of people with disabilities is consistent with the notion reflected in international law of civil and political rights, on the one side, and economic, social and cultural rights, on the other, as unified and mutually independent principles[2].

The present situation regarding the use of AI to support independence and fulfillment for individuals with disabilities is mixed: on the one hand, some AI applications developed to address disability-related needs are now available. On the other hand, much of the practical benefit provided by popular AI-oriented tools, in particular dialogue-based personal assistants, tends to be accidental rather than a result of deliberate efforts to design with people with disabilities in mind. A strategy is needed that invests both in the accessibility and usability of applications developed for general populations,

and, as necessary, in the creation of specialized tools specific to meeting the diverse use cases that living with a disability presents.

[2]*For a discussion of this approach in connection with the rights of persons with disabilities, see Degener [5].*

## 5. CONCLUSION

Challenges of fairness for people with disabilities in AI systems lead to an interplay of issues in ethics, law and technology that necessitate an interdisciplinary collaboration among researchers and practitioners. Understanding the problems posed by machine learning technologies, as well as the opportunities they create, is crucial to effective intervention, as is the involvement of people with disabilities themselves in shaping policy, design and implementation. Further research and practice in this area should be informed by considerations of ethics, and also by an understanding of how the challenges as they manifest themselves for people with disabilities are related to the difficulties that these technologies have the potential to create for society at large, and for other populations who are at risk of unfair treatment.

## ACKNOWLEDGMENTS

The author gratefully acknowledges Mark Hakkinen (Educational Testing Service), and Clayton Lewis (University of Colorado Boulder) for providing valuable comments and discussion. Cara Laitusis and Nitin Madnani of Educational Testing Service thoughtfully reviewed the penultimate draft of the manuscript.

## REFERENCES

1. M. Altman, A. Wood, and E. Vayena (2018) A harm-reduction framework for algorithmic fairness. IEEE Security & Privacy 16 (3), pp. 34–45. Cited by: §3, §3.
2. M. Ameri, L. Schur, M. Adya, F. S. Bentley, P. McKay, and D. Kruse (2018) The disability employment puzzle: a field experiment on employer hiring behavior. ILR Review 71 (2), pp. 329–364. Cited by: §3.
3. S. I. Beacher and U. Benoliel (2019) Law in books and law in action: the readability of privacy policies and the gdpr. In Consumer Law and Economics, External Links: Link Cited by: footnote 1.
4. P. De Filippi and S. Hassan (2018) Blockchain technology as a regulatory technology: from code is law to law is code. arXiv preprint arXiv:1801.02507. Cited by: §3.
5. T. Degener (2016) A human rights model of disability. In Routledge Handbook of Disability Law and Human Rights, pp. 47–66. Cited by: footnote 2.
6. S. O. Hansson (2007) Philosophical problems in cost–benefit analysis. Economics & Philosophy 23 (2), pp. 163–183. Cited by: §3.
7. J. Lazar, B. Wentz, and M. Winckler (2017) Information privacy and security as a human right for people with disabilities. In Disability, Human Rights, and Information Technology, pp. 199–211. Cited by: §2.
8. Y. LeCun, Y. Bengio, and G. Hinton (2015) Deep learning. Nature 521 (7553), pp. 436–444. External Links: ISSN 0028-0836 Cited by: §4.
9. M. C. Nussbaum (2009) Frontiers of justice: disability, nationality, species membership. Harvard University Press. Cited by: §3, §3.
10. J. Rawls (1999) A theory of justice. revised edition, Harvard University Press. Cited by: §3.
11. S. Trewin (2018) AI fairness for people with disabilities: point of view. arXiv preprint arXiv:1811.10670. Cited by: §2, §3.

## ABOUT THE AUTHORS

Jason White is an Associate Research Scientist in the Accessibility Standards and Inclusive Technology Group at Educational Testing Service. In this role, he has participated in projects intended to enhance the accessibility of educational materials and assessments delivered via Web technologies to people with disabilities. He is also a long-standing contributor to standards-related activities, and currently serves as Facilitator of the Research Questions Task Force in the W3C's Accessible Platform Architectures Working Group.