The Case for Correctability in Fair Machine Learning

Mattia Cerrato[1], Alesia Vallenas Coronel[1], Marius Köppel [1]

[1] *Johannes Gutenberg-Universität, Saarstraße 21, Mainz 55122, Germany*

Abstract

Beyond a statistical account of group-based fairness, individual fairness approaches in machine learning have been historically motivated by the notion of "similar treatment" — individuals with similar data should be treated similarly. In this paper, we propose to extend the notion of individual fairness so to implement a fair machine learning process more generally. Our focus is on the concept of correctability as introduced by Leventhal: the ability of people to challenge allocative decisions. During the machine learning process, however, allocative decisions are also made at the data collection step. We therefore argue that affordances for data recourse are necessary to obtain correctability, and thus, a fair machine learning process. While searching for a technical implementation, we claim that correctability should be facilitated by regulatory means. We discuss possible approaches considering anti-discrimination law and the AI Act.

Keywords
Individual fairness, procedural fairness, correctability

Research on "algorithmic fairness" in machine learning has focused on the properties of individual and group fairness. The latter term refers, informally, to the property of avoiding discrimination against historically disadvantaged groups on egalitarian grounds [1]. Group fairness techniques in e.g. classification seek to assign certain outcomes in a balanced fashion across different groups of people, defined over the characteristics protected by anti-discrimination law. However, these measures are limited to a statistical account of the concept of fairness, and may thus neglect edge cases or outliers.

Individual fairness, on the other hand, is motivated by the principle of "similar treatment", mandating that similar individuals should be treated similarly. In machine learning terms, this implies that people with similar *data* should receive similar decisions. The concept is due to Dwork et al. [2] who also claim inspiration to Rawls and egalitarianism [3]. Other authors such as Binns [4] have pointed out a connection to classical conceptions of justice as consistency, while Fleisher [5] discusses the limitations of individual fairness as a general, "correct" or "primary" notion of fairness, in contrast to previous proposals [7].

Our proposal is to extend the notion of individual fairness so that it may include the notion of fair decision process and we more in general seek a possible definition for the concept of fair machine learning process. Our inspiration is the concept of procedural fairness in law and philosophy [6]. More specifically, we identify correctability as an underexplored feature of the machine learning process and put forward a constructive proposal so to account for it in automated or assisted decision making via machine learning.

We first take correctability in the meaning proposed by Leventhal [8], that is, the ability of people to challenge allocative decisions. Here previous work in the field of "fair ML" has modelled the task of "algorithmic recourse" as allowing certain individuals to challenge decisions depending on how close they are to the decision boundary of a classification model [9]. Other authors have put forward proposals to evaluate and optimise models so to maximise the opportunity and easiness of *individual* recourse

EWAF'23: European Workshop on Algorithmic Fairness, June 07–09, 2023, Winterthur, Switzerland
EMAIL: mcerrato@uni-mainz.de (A. 1); alvallen@uni-mainz.de (A. 2); mkoeppel@uni-mainz.de (A. 3)

[15]. A recent systematic review of techniques in algorithmic recourse [16] has highlighted different characteristics of these methodologies while analysing them under the lens of *counterfactual explainability*. A counterfactual explanation is an actionable modification in an individual's data which would have implied a positive allocative decision.

In contrast with these proposals, our view is that allocative decisions are implicitly made at every step of the machine learning loop (see Figure 1), and not only when a certain hypothesis class and objective function are chosen, or when the system is employed in the real world. While affordances for recourse of individual model decisions may lead to correctable *models*, we would define a correctable machine learning *process* as the situation in which **i)** individuals are able to recourse against all steps of the machine learning process and **ii)** the developers and data curators seek to actively include the insights gained from the recourse procedure in future cycles of the ML loop. We note that a similar distinction has been previously drawn by Venkatasubramanian and Alfano [17, Section 3.1], who posited that counterfactual explanations may amount only to a "recourse narrowly defined" whereas a correctable "appeal" procedure would require e.g. rectifying incorrect features and data.

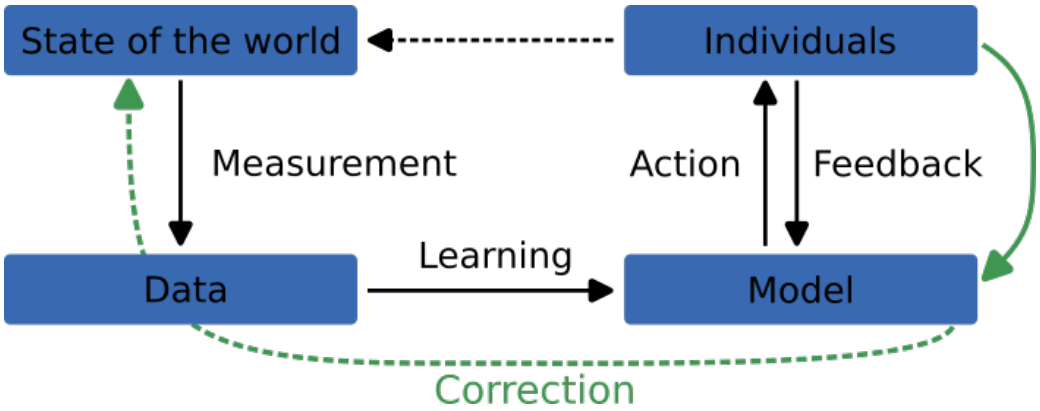

Figure 1: a schematic of the ML loop, involving the measurement and learning steps which create the data and the model respectively. This visualisation of the ML process is due to Barocas, Hardt and Naranyanan [10]. Our proposal to obtain a correctable ML process is to have the appeals made against individual decisions be propagated into the earlier steps of the process, including data collection in particular.

Our present focus is broadly on data recourse and designing how individuals may meaningfully interact – and challenge – the data collection process. We identify here three separate affordances for recourse:

1. Factual recourse. After revising the collected information, an individual notices that some feature x_i is factually incorrect, and asks the data curator for correction. As an example, the data may be outdated due to time delays between the data collection process and the decision undertaken by the ML model.
2. Parameter recourse. Some feature x_i has a negative effect on the decision; nonetheless, the individual maintains that its value should be interpreted as a positive in their case. We note that this recourse regards both the data collection and learning steps of the ML loop.
3. Contextual recourse. The individual puts forward the claim that some feature x_i which has a negative effect should be interpreted considering some other piece of information, e.g. another feature x_{i+1} which had not been seen during the data collection process.

We argue that none of these three affordances allow for a straightforward ML implementation strategy. In the following, we discuss some of the limitations of existing approaches when dealing with the situations described above. Furthermore, we outline some directions for possible future research seeking to fulfill correctability of machine learning processes in a deeper sense. We do not propose

here, however, a specific algorithmic intervention strategy to implement factual, parameter or contextual recourse.

Factual recourse may seem to require only a simple database correction. However, factual/imputation errors in the data may also reveal deeper issues with the data collection process and possibly measurement bias against certain individuals or groups [11]. Implementing a strategy for parameter recourse would require case-by-case reasoning to understand whether the data, or some subset of the data, supports the claim that a certain feature value should be interpreted as a positive in place of a negative. Some level of logical reasoning is also required in contextual recourse, with the added challenge that individual users may not be able to provide the additional contextual information in a machine-readable format.

Apart from possible technical solutions, ensuring that an individual's interaction with the different steps of the ML loop is procedurally fair may be facilitated by a regulatory approach. In this context, an approach based on anti-discrimination law would grant subjective rights to the affected individuals while designing complementary enforcement mechanisms. Unlike in the analogue world, the abstract, more subtle and less intuitive unequal treatment in a ML context is regularly not perceived at all, neither by the developers or users nor by the disadvantaged persons [12]. Hence, to recognise disadvantages and to be able to demonstrate the correlations between the decision at issue and a protected characteristic, those affected would also have to be able to understand the data on which the decision is based, along with how it is processed by the relevant ML model [13]. Against this background, we argue that corresponding rights to information and obligations to provide reasons need to be included in anti-discrimination law, while safeguarding the economic interests of the developers. As EU anti-discrimination law so far only covers specific civil law transactions and only by certain actors, the scope of application may need to be extended in a "situational regard" to ML as argued by Wolff [12].

The European Commission's proposal for an AI Act [14] as a regulatory approach under administrative law operates with licensing and supervisory elements and does not seem to entail ways for individuals to interact with the machine learning process. It focuses on the bilateral relation between the developer and the user of a ML model. We argue that a comparison with the GDPR may enable individual recourse by further developing and adapting the data subject rights as constituted in Art. 12 et seq. to automated or assisted decision making via ML.

References

[1] Binns, Reuben. "Fairness in machine learning: Lessons from political philosophy." *Conference on fairness, accountability and transparency*. PMLR, 2018.

[2] Dwork, Cynthia, et al. "Fairness through awareness." *Proceedings of the 3rd innovations in theoretical computer science conference*. 2012.

[3] Rawls, John. *Justice as fairness: A restatement*. Harvard University Press, 2001.

[4] Binns, Reuben. "On the apparent conflict between individual and group fairness." *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 2020.

[5] Fleisher, Will. "What's fair about individual fairness?." *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 2021.

[6] A. Tschentscher, Prozedurale Theorien der Gerechtigkeit, Nomos, Baden-Baden, 2000.

[7] Friedler, Sorelle A., Carlos Scheidegger, and Suresh Venkatasubramanian. "The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making." *Communications of the ACM* 64.4 (2021): 136-143.

[8] Leventhal, Gerald S. "What should be done with equity theory? New approaches to the study of fairness in social relationships." (1976).

[9] Sharma, S.; Gee, A. H.; Paydarfar, D.; and Ghosh, J. 2021. FaiR-N: Fair and Robust Neural Networks for Structured Data. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society.
[10] Barocas, S.; Hardt, M.; and Narayanan, A. 2019. Fairness and Machine Learning. fairmlbook.org. http://www.fairmlbook.org.
[11] Jacobs, Abigail Z., and Hanna Wallach. "Measurement and fairness." *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 2021.
[12] Wolff, Daniel. "KI-Biases im Gesundheitswesen." *DuD 1* (2023): 37-41.
[13] Barocas, Solon, and Andrew D. Selbst. "Big Data's Disparate Impact." *Cal L Rev 104* (2016), 671-732.
[14] European Comission. *Artificial Intelligence Act: Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative Act.* COM/2021/206 final. 2021.
[15] Ustun, B.; Spangher, A; and Liu Y. 2019. *Actionable Recourse in Linear Classification.* In: Conference on Fairness, Accountability and Transparency, January 29-31, 2019, Atlanta, GA, USA. ACM, New York, NY, USA.
[16] Verma, S.; Boonsanong V.; Hoang, M.; Hines, K.E.; Dickerson, J; and Shah, C. *Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review.* ArXiv Preprint. ID: arXiv:2010.10596v3. 2022.
[17] Venkatasubramanian, S; and Alfano, M. 2020. *The philosophical basis of algorithmic recourse*. In Conference on Fairness, Accountability and Transparency, January 27-30, 2020, Barcelona, Spain. ACM, New York, NY, USA.