On Prediction-Modelers and Decision-Makers: Why Fairness Requires More Than a Fair Prediction Model

Teresa Scantamburlo[1,*], Joachim Baumann[2,3,*] and Christoph Heitz[3,*]

[1]*Ca' Foscari University of Venice, European Centre for Living Technology, Italy*
[2]*University of Zurich, Switzerland*
[3]*Zurich University of Applied Sciences, Switzerland*

Abstract

This paper addresses the ambiguous relationship between prediction and decision in the field of prediction-based decision-making. Many studies blur these concepts, referring to 'fair prediction' without a clear differentiation. We argue that distinguishing between prediction and decision is crucial for ensuring algorithmic fairness, as fairness concerns the consequences on human lives created by decisions, not predictions. We clarify the distinction between the concepts of prediction and decision, and demonstrate how these two elements influence the final fairness properties of a prediction-based decision system. To this aim, we propose a framework that enables a better understanding and reasoning about the conceptual logic of creating fairness in prediction-based decision-making. Our framework delineates different roles, specifically the 'prediction-modeler' and the 'decision-maker,' and identifies the information required from each to implement fair systems. This framework facilitates the derivation of distinct responsibilities for both roles and fosters discussion on insights related to ethical and legal requirements.
This is an extended abstract based on the full paper published in AI & SOCIETY [1].

Keywords

prediction-based decision, algorithmic fairness, ethical decision-making, human-in-the-loop

1. Introduction

Algorithmic fairness has emerged as a important topic within the Machine Learning (ML) research community during recent years [2, 3], attracting attention not only from a technical standpoint but also from philosophical, political, and legal perspectives [4, 5]. This literature builds upon established scholarship investigating the limits of classification systems and power asymmetries in data collection practices [6]. Concerned with the consequences of prediction-based decisions on individuals and groups from a social justice viewpoint [7], the discourse on algorithmic fairness has been focusing on the fairness of prediction models, which represent the core of ML research [8, 9, 10, 11, 12]. Given this focus, it is unsurprising that much of the debate has revolved around how prediction models can cause unfairness.

We argue that the prediction model as such cannot be the reason for unfairness. Rather, it is the *usage* of the prediction model within its specific context that leads to unfairness. For instance,

EWAF'24: European Workshop on Algorithmic Fairness, July 01–03, 2024, Mainz, Germany

[*] All authors contributed equally.

teresa.scantamburlo@unive.it (T. Scantamburlo); baumann@ifi.uzh.ch (J. Baumann); christoph.heitz@zhaw.ch (C. Heitz)
0000-0002-3769-8874 (T. Scantamburlo); 0000-0003-2019-4829 (J. Baumann); 0000-0002-6683-4150 (C. Heitz)

the racial discrimination attributed to the COMPAS tool's recidivism risk model [13] is not an inherent feature of the model itself. Discrimination arises only when judges make decisions based on the COMPAS risk scores. Consequently, the connection between the characteristics of a prediction model, such as its false-positive or false-negative rates, and the potential harm to specific societal groups, like African Americans in the case of COMPAS, hinges on the assumption of how the model's outputs translate into tangible consequences for people's lives.

As a prototypical case of how prediction models are implemented in real-world applications, we focus our analysis on *prediction-based decision systems*, in which the outcomes of ML prediction algorithms are leveraged to make decisions impacting human lives. We imagine a (human or automated) decision-maker who is taking decisions on people or for people, who makes decisions about or for individuals, while these decisions are informed by a prediction regarding certain characteristics of the people involved. This scenario encapsulates many of the cases commonly associated with discussions on algorithmic fairness, including banks making loan decisions based on repayment predictions, companies making hiring decisions based on job performance prediction, or universities making admission decisions based on academic achievement predictions.[1]

In such prediction-based decision systems, we may distinguish two conceptually different functions: First, we have the function of *prediction*, performed by a prediction model that processes individual data of a person to produce a prediction of a target variable associated with this person, which is not known to the decision-maker at the time of decision-making. This prediction might come in the form of a score, a probability, or a point prediction. Second, we have the function of *decision*, which is informed by the prediction, but in nearly all cases also influenced by additional parameters. For example, in of a bank's loan decision, not only the repayment probability but also the interest rate and the bank's business strategy may be decisive parameters. This idea has been studied in so-called cost-sensitive learning problems [14]. However, it remains unclear how incorporating a fairness requirement alters the cost-sensitive approach and how prediction and decision functions interact in this process.

This paper is primarily concerned with on group fairness, the most established and widely adopted category of fairness. Group fairness seeks to prevent systematic disadvantages in algorithmic decisions based on sensitive attributes (such as gender, age, or race) [15, 2]. While there are also other types of fairness, such as counterfactual fairness [16], individual fairness [11], and procedural fairness definitions [17], these are not addressed in this study.

2. The relation between predictions and decisions

In popular narratives of algorithmic decision-making, the distinction between the idea of decision and that of prediction seems to be blurred. Neologisms like 'fair prediction' [18] or 'fairness-aware learning' [10] have become familiar within the ML community, inadvertently promoting the notion that fairness is an intrinsic property of prediction models. This conflation of concepts does not necessarily stem from an explicit ideological stance, and some studies

[1] It is important to note that other scenarios, such as recommender systems where predictions inform individuals making decisions about themselves, also exist. While the findings of this paper may not be directly applicable in such contexts, they could serve as inspiration for future research.

clearly specify that fairness is a characteristic pertinent to decision rules [19]. Nonetheless, formal characterizations often apply fairness criteria to the prediction model (e.g., the classifier), presupposing that the decision equates to the prediction's outcome (e.g., see [20, 21, 22]).

Such formulations suggest that the relation between prediction and decision is fixed and given, implying that a specific prediction directly leads to a specific decision. However, this is not true in many realistic examples, where the optimal decision depends on the prediction as well as on other parameters, as is explicitly acknowledged by the idea of cost-sensitive learning [14]. Thus, qualifying a prediction as 'fair' is misleading unless we explicitly assume how a prediction is transformed into a decision. In general, the fairness attribution applies more properly to the entire system (i.e., the combination of prediction and decision rules) rather than to the prediction alone.

Abstract formalization facilitates the overlap between the concepts of prediction and decision. For example, in classification tasks, the objective of prediction—to choose from among several options—can be seen as a special type of decision-making. From this viewpoint, the functions performed by a prediction model and a decision-maker appear similar. However, moving beyond mathematical simplifications to consider ethical implications reveals that decision-making encompasses more than selecting alternatives; it involves actions that affect individuals and the environment. In other words, decisions change the *status quo*, thus bearing consequences for the decision-maker, the decision subjects, and potentially the broader environment. In contrast, a prediction in itself has no direct impact; its influence on decision-making is enabled only through a policy or decision rule that outlines the consequences of future actions.

Take, for instance, a bank that decides whether to approve loan applications based on the predicted likelihood of repayment. Approving a loan has a tangible impact, offering the recipient enhanced financial flexibility and new purchasing opportunities. This benefit is withheld from applicants who are denied. Apparently, the prediction algorithm influences the decision, but the prediction itself is not what creates (un)fairness, it is the decision specifying how to use the prediction estimate. Note that even if the decision is fully determined by the prediction – a case which is rarely met – the question of whether the prediction algorithm is fair or not is conditioned on the assumed relation between prediction and decision rule. This is why we conceptually suggest to clearly distinguish between the two elements of prediction and decision, which both are ingredients of any prediction-based decision system, whether it be fully automatic or also influenced by humans. Most importantly, distinguishing these concepts encourages a broader examination of algorithmic decision-making as a process embedded in social constructs, reflective of value judgments and power asymmetries.

3. Responsibilities of prediction-modelers and decision-makers

For studying the interaction of prediction and decision, we introduce a framework allowing us to distinguish the tasks and responsibilities of two different roles: The role of the 'prediction-modeler,' and the role of the 'decision-maker.' Drawing on decision-theoretic concepts, we may think of two different agents, one being responsible for the prediction model and the other one being responsible for the decision-making. Our motivation for distinguishing these roles is not only fed by the theoretical analysis of how predictions are converted into (un)fair treatment as discussed above, but also by the practical observation that these two roles are often split organizationally and covered by different people, different departments, or even different companies. From a responsibility standpoint, the decision-maker is responsible for the decisions, and hence their consequences. Conversely, prediction-modelers have their own set of responsibilities. They are responsible for creating the basis for a good decision, which consists in (a) delivering a meaningful and robust prediction (e.g., think of transparency and safety requirements in [23]), and (b) supplying the decision-maker with all necessary information to address fairness and other relevant ethical obligations (refer to accountability and fairness considerations in [23] and the requirements outlined by [24]).

The rationale for differentiating these roles stems from their distinct objectives. Prediction-modelers focus on achieving high prediction performance, such as accuracy, which poses challenges in the context of consequential decisions (see, e.g. [25] and [26]).

Our framework[2] is grounded in a decision-theoretic analysis of prediction-based decision-making, assuming binary outcomes for both the decision D and the decision-critical unknown variable Y, and links to existing literature that conceptualizes fairness as a decision-theoretic problem (see, for example, [27] and [28]). The analysis demonstrates that rational decision-making is an optimization problem, reliant on the decision-maker's utility function and the prediction $P(Y = 1)$. Accordingly, the ideal output of a prediction model for a decision-maker is a probabilistic model, in particular a calibrated score. Interestingly, this does not change if a fairness constraint is added to the decision problem – the only change is in the decision rule.

These results suggest that to achieve fairness while still optimizing for a decision-maker's goal, the task of the prediction-modeler is not to deliver a 'fair' prediction model, but to deliver not-skewed estimation of the outcome of interest. On the other hand, the task of the decision-maker is to combine the prediction with their goal achievement.

We also analyze the necessary interaction between the two roles to enable the development of fair decision systems. The decision-maker should communicate details about the target variable Y and group attributes to the prediction-modeler, who, in turn, must relay information on model performance, calibration functions, and group-specific baseline distributions.

This study focuses on human decision-makers at the final stage of the AI decision process. We know there are other important decision-makers involved earlier, like human annotators and trainers. While their roles are also important, discussing all human influences is beyond the scope of this paper.

4. Discussion and conclusions

Our findings underscore that different actors bear varying responsibilities towards the collective aim of creating a fair system.[3] In algorithmic decision-making under group fairness constraints, these obligations translate into specific pieces of information that each role is expected to deliver to the other. The deliverables suggested in our framework are not optional. They can be derived

[2] We focus on a post-processing approach to fairness. For a more extensive discussion of how our approach relates to pre-processing and in-processing techniques see the full article [1].

[3] Here, we focus more specifically on professional responsibility, that is, the set of obligations based on a role played in a certain context. Responsibility, of course, extends beyond roles, and for a broader discussion see [29].

from the very nature of the decision problem and establish a strong interdependence between the roles involved.

While we acknowledge that the ultimate responsibility for ensuring fairness in decisions lies with the decision-maker, their ability to address fairness issues depends heavily on the work of the prediction-modeler. Without a minimum set of essential information regarding both the training population and performance characteristics of the prediction model, a decision maker cannot guarantee fair decisions.

The interdependence between the roles recalls the problem of creating meaningful communication channels among parties involved in the design and implementation of artificial intelligence (AI) systems. Specifically, it clarifies the need for a structured exchange of information critical for achieving fairness in prediction-based decision-making.

The line between organizations that design and those that operate decision-making systems can be unclear and complicated. For example, when banks work with credit agencies, it can be hard to separate their roles in ensuring fairness – credit agencies might use banks' preferences in their models. This highlights the need to look at fairness in decision-making not just in theory but also by considering the actual roles and institutions involved.

References

[1] T. Scantamburlo, J. Baumann, C. Heitz, On prediction-modelers and decision-makers: why fairness requires more than a fair prediction model, AI & SOCIETY (2024) 1–17. doi:10.1007/s00146-024-01886-3.

[2] S. Barocas, M. Hardt, A. Narayanan, Fairness and Machine Learning, fairmlbook.org, 2019.

[3] M. Kearns, A. Roth, The Ethical Algorithm: The Science of Socially Aware Algorithm Design, Oxford University Press, Inc., USA, 2019.

[4] R. Binns, Fairness in Machine Learning: Lessons from Political Philosophy, Technical Report, 2018. URL: http://proceedings.mlr.press/v81/binns18a.html.

[5] S. Barocas, A. D. Selbst, Big Data's Disparate Impact, California Law Review 104 (2016) 671–732. URL: http://www.jstor.org/stable/24758720.

[6] G. C. Bowker, S. L. Star, Sorting things out: Classification and its consequences, MIT press, 2000.

[7] D. K. Mulligan, J. A. Kroll, N. Kohli, R. Y. Wong, This thing called fairness: Disciplinary confusion realizing a value in technology, Proceedings of the ACM on Human-Computer Interaction 3 (2019) 1–36.

[8] D. Pedreschi, S. Ruggieri, F. Turini, Discrimination-Aware Data Mining, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '08, Association for Computing Machinery, New York, NY, USA, 2008, p. 560–568. URL: https://doi.org/10.1145/1401890.1401959. doi:10.1145/1401890.1401959.

[9] T. Calders, S. Verwer, Three Naive Bayes Approaches for Discrimination-Free Classification, Data Min. Knowl. Discov. 21 (2010) 277–292. URL: https://doi.org/10.1007/s10618-010-0190-x. doi:10.1007/s10618-010-0190-x.

[10] T. Kamishima, S. Akaho, H. Asoh, J. Sakuma, Fairness-Aware Classifier with Prejudice Remover Regularizer, in: P. A. Flach, T. De Bie, N. Cristianini (Eds.), Machine Learning

and Knowledge Discovery in Databases, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 35–50.

[11] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel, Fairness through awareness, in: ITCS 2012 - Innovations in Theoretical Computer Science Conference, ACM Press, New York, New York, USA, 2012, pp. 214–226. URL: http://dl.acm.org/citation.cfm?doid=2090236.2090255. doi:10.1145/2090236.2090255.

[12] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, C. Dwork, Learning Fair Representations, in: S. Dasgupta, D. McAllester (Eds.), Proceedings of the 30th International Conference on Machine Learning, volume 28 of *Proceedings of Machine Learning Research*, PMLR, Atlanta, Georgia, USA, 2013, pp. 325–333. URL: https://proceedings.mlr.press/v28/zemel13.html.

[13] J. Angwin, J. Larson, S. Mattu, L. Kirchner, Machine bias, ProPublica, May 23 (2016) 139–159. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[14] C. Elkan, The Foundations of Cost-Sensitive Learning, in: Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2, IJCAI'01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001, pp. 973–978.

[15] R. Binns, On the apparent conflict between individual and group fairness, in: FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, Inc, New York, NY, USA, 2020, pp. 514–524. URL: https://dl.acm.org/doi/10.1145/3351095.3372864. doi:10.1145/3351095.3372864.

[16] M. J. Kusner, J. Loftus, C. Russell, R. Silva, Counterfactual Fairness, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

[17] N. Grgić-Hlača, M. B. Zafar, K. P. Gummadi, A. Weller, Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning, Proceedings of the AAAI Conference on Artificial Intelligence 32 (2018). URL: https://ojs.aaai.org/index.php/AAAI/article/view/11296.

[18] A. Chouldechova, Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments., Big data 5 (2017) 153–163. doi:10.1089/big.2016.0047.

[19] S. Corbett-Davies, S. Goel, The measure and mismeasure of fairness: A critical review of fair machine learning, arXiv preprint arXiv:1808.00023 (2018).

[20] M. B. Zafar, I. Valera, M. G. Rodriguez, K. P. Gummadi, Learning fair classifiers, arXiv preprint arXiv:1507.05259 1 (2015).

[21] A. K. Menon, R. C. Williamson, The cost of fairness in binary classification, in: S. A. Friedler, C. Wilson (Eds.), Proceedings of the 1st Conference on Fairness, Accountability and Transparency, volume 81 of *Proceedings of Machine Learning Research*, PMLR, New York, NY, USA, 2018, pp. 107–118. URL: http://proceedings.mlr.press/v81/menon18a.html.

[22] R. Berk, H. Heidari, S. Jabbari, M. Kearns, A. Roth, Fairness in Criminal Justice Risk Assessments: The State of the Art, Sociological Methods & Research 50 (2021) 3–44. URL: https://doi.org/10.1177/0049124118782533. doi:10.1177/0049124118782533.

[23] High-Level Expert Group on Artificial Intelligence, Ethics guidelines for trustworthy AI, Technical Report, European Commission, Brussles, 2019. URL: https://op.europa.eu/en/

publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1.

[24] European Commission, Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on AI and amending certain union legislative acts, Technical Report, Brussels, 2021. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206.

[25] S. Athey, The Impact of Machine Learning on Economics, in: A. Agrawal, J. Gans, A. Goldfarb (Eds.), The Economics of Artificial Intelligence: An Agenda, University of Chicago Press, 2019, pp. 507–552.

[26] S. Athey, Beyond prediction: Using big data for policy problems, Science 355 (2017) 483–485. doi:10.1126/science.aal4321.

[27] N. S. Petersen, An Expected Utility Model for "Optimal" Selection, Journal of Educational Statistics 1 (1976) 333–358. URL: https://doi.org/10.3102/10769986001004333. doi:10.3102/10769986001004333.

[28] R. L. Sawyer, N. S. Cole, J. W. L. Cole, Utilities and the Issue of Fairness in a Decision Theoretic Model for Selection, Journal of Educational Measurement 13 (1976) 59–76. URL: http://www.jstor.org/stable/1434493.

[29] Van de Poel Ibo and Royakkers Lambèr, Ethics, Technology and Engineering:An Introduction, Wiley-Blackwell, 2011.