Using Fairness Metrics as Decision-Making Procedures: Algorithmic Fairness and the Problem of Action-Guidance

Otto Sahlgren [1]

[1] *Tampere University, Tampere, Finland*

Abstract

Frameworks for fair machine learning are envisioned to play an important practical role in the evaluation, training, and selection of machine learning models. In particular, fairness metrics are meant to provide responsible agents with actionable standards for evaluating ML models and conditions which those models should achieve. However, recent studies suggest that fair ML frameworks and metrics do not provide sufficient and actionable guidance for agents. This short paper outlines the main content of a working paper wherein I draw lessons from philosophical debates concerning action-guidance to build a conceptual account that can be applied to analyze whether and when fair ML frameworks and metrics can generate determinate evaluations of fairness and actionable prescriptions for model selection.

Keywords

Algorithmic fairness, fair machine learning, action-guidance, moral philosophy, political philosophy, practical ethics

1. Introduction

The fair machine learning ("fair ML") research community has proposed numerous frameworks which are supposed to help responsible agents (e.g., ML practitioners and algorithm auditors) identify and mitigate unfair bias in ML models. Most existing fair ML frameworks (and software toolkits designed to support their implementation) include at least two components: first, a formal definition and a corresponding metric for fairness and, second, a bias mitigation method that can be applied to improve ML models in terms of fairness [3, 8, 16]. Fair ML frameworks are envisioned to play an important practical role in ethical decision-making throughout the ML system pipeline. In particular, a common vision is that fairness metrics provide responsible agents with evaluative standards that can be applied to evaluate and compare available ML models, and also define absolute conditions that ML models should satisfy to count as 'fair' or 'just' (see [3, 16]). However, recent studies identify many limitations to using fair ML frameworks in practical contexts: evaluating model fairness can require considerable expertise and contextual value judgments from agents, and improving fairness can be infeasible or result in unintended, bad consequences [3, 4, 5, 7]. For these and other reasons it has been argued that statistical frameworks for fair ML, for instance, "do not offer sufficient practical guidance and can lead to misguided mitigation strategies" when applied in real-life contexts [3, p. 60].

The working paper outlined here provides a philosophical examination of problems that relate to the informativeness and usability of fair ML frameworks in practical contexts. I approach these problems from the perspective of philosophical debates concerning normative principles and theories' capacity to guide agents' action in real-life circumstances of ethical evaluation and decision-making [1, 2, 12, 14]. The paper consists of two parts which I will outline below: The first part (outlined in Section 2) surveys problems that have been identified in connection to the informativeness, usefulness and

[1] EWAF'23: European Workshop on Algorithmic Fairness, June 07–09, 2023, Winterthur, Switzerland
EMAIL: otto.sahlgren@tuni.fi
ORCID: 0000-0001-7789-2009

application of fair ML frameworks in practical contexts of model evaluation and training. The second part (outlined in Section 3) draws on Richard North's work on action-guidance [12] to develop an account of action-guidance for *proxy constructs* that operationalize normative principles or theories (e.g., evaluation metrics, ethics codes, and ethical design guidelines). The discussion focuses primarily on action-guidance in the evaluation and improvement of model fairness, but the proposed account also has potential to illuminate debates concerning the limits of formal approaches to ethical evaluation [6, 15] and the usefulness of principled approaches to AI ethics [9, 11, 13].

2. The Problem of Action-Guidance

Numerous problems can hinder or prevent the (correct) application of fair ML frameworks and software fairness toolkits in the ML system pipeline. For example, agents tasked with ensuring model fairness can lack the expertise required for implementing fair ML methods [5, 8]. Many frameworks also require agents to resort to their own contextual judgment when evaluating and improving model fairness [4]. A lack of comparative guidance regarding how fairness metrics and bias mitigation methods should be selected and applied introduces risks for their misuse and misapplication [8]. Moreover, problems can arise even when responsible agents know when and how fair ML methods should be applied in theory. There can be legal and organizational constraints standing in the way of fairness-sensitive design, such as restrictions on data access [5]. Due to emerging trade-offs, satisfying fairness objectives can also be practically infeasible or lead to outcomes that are otherwise undesirable [3, 7].

The previously described problems are notably emblematic of the general "gap [that] exists between the theory of AI ethics principles and the practical design of AI systems" [10, p. 239]. In the working paper outlined here, I contextualize these problems against the background of philosophical debates surrounding the problem of *action-guidance* [1, 2, 12, 14]. These debates center on the question: can normative principles (or theories) inform and guide agents' moral conduct in real-life circumstances? Competing views have been presented about whether and when moral principles can offer actionable information to agents [1, 12, 14]. There is also underlying disagreement about what kinds of information should be considered "action-guiding" in the first place [2, p. 555]. To address questions related to action-guidance in fair ML, I draw particularly on debates on ideal theory in political philosophy, wherein theorists have debated whether principles that are constitutive of justice in ideal circumstances can be used to derive prescriptions for action in actual, non-ideal circumstances ([2, 12] and see also [3]). This contextualization is motivated by the widely held notion that definitions and metrics for model fairness constitute *proxy constructs* that operationalize egalitarian principles of justice (see [3, 6, 16]). I suggest that the fair ML research community would benefit from an account of the conditions under which proxy constructs (e.g., fairness metrics) can guide agents' action.

3. An Account of Action-Guidance for Proxy Constructs

The second part of the paper develops an account of action-guidance for proxy constructs, prominent examples of which include ethics codes, ethical design guidelines, evaluation metrics and standards that operationalize moral principles or theories. I will focus primarily on fairness metrics and fair ML frameworks more broadly (*qua* proxies that operationalize egalitarian principles of justice) to demonstrate the applicability of the proposed account, to analyze previously identified problems with action-guidance in fair ML, and to draw lessons for designing actionable fair ML frameworks.

The proposed account owes much to Richard North's view according to which principles of justice are action-guiding when they can be used as decision-making procedures by the constituency of those principles [12]. North argues that action-guiding principles of justice indicate (upon correct application) whether actions available to citizens are just or unjust, and that citizens are able to derive from such principles a coherent prescription for action. North's account suggests that two kinds of factors are important for action-guidance: On the one hand, there are certain properties or elements that the principle(s) should possess to ensure that they generate determinate, consistent, and coherent evaluations. On the other hand, the agent applying the principle should also possess the "beliefs and abilities needed to derive a prescription from that principle and [to] act in conformity with that prescription" [12, p. 81]. It is consistent with this account that we can disagree about what is just to

begin with, and that other matters (e.g., other values than justice) can bear on the overall deontic status of the actions available to citizens.

From North's account [12] one can derive at least seven more specific conditions for action-guidance: (1) the principle(s) should explain what makes an object or act right or good (or wrong or bad); (2) the principle(s) should generate identical evaluations of deontic status for identical cases; (3) the principle(s) should indicate an object or act as either permissible, impermissible, or required; (4) the principle(s) should be applicable to a sufficiently broad set of objects or acts; (5) the prescriptions derived from the principle(s) should cohere with empirical facts about the world; (6) the agent should know whether they belong to the constituency of the principle(s); and (7) the agent should be able to derive a correct prescription for action by applying the principle(s). If these conditions are satisfied, North suggests, principles of justice can be used to derive action-guiding prescriptions even in most cases where "ideally just" actions are unavailable to citizens; or, at the very least, they will help agents identify what kinds of *secondary duties* apply to them [12, p. 87].

I draw on North's account to develop an account of action-guidance for proxy constructs. The proposed account can be summarized as follows:

> A proxy construct *F** that operationalizes a target principle *F* is action-guiding if and only if agent *A* belonging to the constituency of *F* can use *F** as a decision-making procedure to derive a prescription for how to comply with *F* in context *C*.

I conceptualize action-guidance as a relation that obtains between a proxy construct *F** and an agent *A* when (i) *F** indicates the deontic statuses of actions available to the agent *A* in light of the principle *F* operationalized by *F** and (ii) agent *A* can derive from *F** a prescription for action that is consistent with *F* in the context of action *C*. The conditions (1–7) described above can be translated into more detailed requirements that specify what is required action-guidance to obtain in this general sense. I divide these requirements into *Construct Requirements* and *Agent Requirements*: The former, Construct Requirements, are based on conditions (1–5) as described above. They specify what kinds of features or elements a proxy construct *F** should possess to properly execute its evaluative function. The latter, Agent Requirements, are based on conditions (6–7). They comprise broader requirements for the beliefs and skills that agents should (be reasonably expected to) possess to be able to apply the proxy construct correctly and to derive a prescription for action from that proxy. Agent Requirements are moderately sensitive to variation in agents' beliefs, skills, resources, and it is reasonable to expect that different audiences tasked with applying proxy constructs (e.g., model developers and domain-experts applying software fairness toolkits [5, 8]) can require varying amounts and types of information.

3.1. Applying the Account to the Case of Fair Machine Learning

Fair ML frameworks are proxy constructs that are meant to support the evaluation and improvement of ML models in terms of fairness (and other broadly egalitarian principles). Fairness definitions provide the evaluative standards and normative targets for agents in these tasks [3, 16]. Dozens of competing fairness definitions and corresponding metrics are currently available – including statistical, counterfactual, and similarity-based metrics [16]. However, to properly examine action-guidance in the context of fair ML, we have to bracket disagreement about specific metrics and instead focus on the commonalities that can be found across different metrics and frameworks, such as their formal nature.

The proposed account suggests that a fairness definition (or metric) is action-guiding to an agent (*qua* proxy construct) if that agent can apply the metric to identify whether the available ML models are impermissible, permissible, or required in light of the target principle operationalized by the metric. Insofar as evaluations generated by the metric track those of the target principle, the agent should be able to draw from the proxy a prescription for action that is consistent with that principle. This general account can also be used to specify more detailed Construct Requirements (CR) and Agent Requirements (AR) for action-guiding fair ML frameworks (or individual fairness metrics):

(CR1 *Explanation*): The metric(s) should track properties constitutive of (un)fairness according to the principle(s) operationalized by the metric(s).

(CR2 *Consistency*): The metric(s) should generate identical evaluations for identical ML models.

(CR3 *Determinacy*): The metric(s) should (i) indicate ML models as either permissible, impermissible, or obligatory in light of the operationalized principle(s), and (ii) order available ML models accordingly.

(CR4 *Scope*): The metric(s) should apply to a well-defined and sufficiently broad class of ML models and use-cases.

(CR5 *Coherence*): The prescriptions derived from the metric(s) should cohere with empirical facts about the world.

(AR1 *Constituency*): The agent should (be reasonably expected to) know whether the principle(s) operationalized by the metric(s) apply to them in the context of deployment *C*.

(AR2 *Affordances*): The agent should (be reasonably expected to) possess the beliefs, skills, and resources required to apply the metric(s) correctly.

In the working paper, I detail these requirements and use them to analyze problems related to action-guidance in fair ML, some of which were mentioned in Section 2. I will briefly mention some examples.

Consider CR1. This requirement states that a fairness metric *F** should articulate and track the features that are morally relevant according to the substantive principle (or theory) of fairness *F* for which *F** constitutes a proxy. This means, among other things, that *F** should provide an agent with means to distinguish between *fair* and *unfair* discrimination (or *just* and *unjust* inequalities) in model predictions. Notably, most existing fairness metrics violate this condition by default due to their strictly formal and "anonymous" nature: they require the agent to identify and operationalize the comparison classes (e.g., social groups) that "matter" from the perspective of fairness [3, 4]. While the formal nature of fairness metrics allows agents to tailor them according to different substantive conceptions of (un)fairness, the proposed view suggests that purely formal metrics are not action-guiding unless actually coupled with a robust, substantive conception of fairness.

Consider also CR3, which states that action-guidance requires evaluations of model (un)fairness to be unambiguous and determinate (see also [2, 12]). An implication of this requirement is that the metric *F** should have a clear threshold that specifies whether and when the level of fairness exhibited by an ML model is (im)permissible. Determining such a threshold is a delicate moral exercise, but a threshold is nonetheless required for the metric to be action-guiding – otherwise the agent has to again resort to contextual judgment. In addition, CR3 also suggests that fair ML frameworks which prescribe multiple fairness objectives (and corresponding metrics) should explicate a second-order decision-rule that agents can apply to resolve trade-offs that arise between those objectives. For instance, a framework could prescribe that a given fairness objective is lexically dominating, meaning that agents should always prioritize its satisfaction over the satisfaction of other objectives (e.g., maximizing accuracy).

Lastly, note that fair ML frameworks have been argued to lack the capacity to provide practical guidance for agents in part because they remain silent about the relevant agents' responsibilities (e.g., who is obliged to mitigate some specific and persistent instance of structural injustice or social inequality) [3]. Requirements CR4 and AR1 are aligned with this claim: to be action-guiding, fair ML frameworks should define and articulate the applicatory scope of fairness objectives proposed therein (CR4) and agents should also know (or be able to find out) whether and when those objectives should bear on their choices between available ML models (AR1).

4. Concluding Remarks

The working paper outlined here discusses problems with action-guidance in practical contexts of evaluating and improving ML model fairness from a philosophical perspective. I propose a general account of action-guidance for proxy constructs (i.e., operationalizations of moral principles or theories) that can be used to draw lessons for designing actionable fair ML frameworks and metrics, and to diagnose problems that arise in practical contexts of model evaluation and improvement. The proposed account is not without its limitations, and much remains to be said on the topics discussed here. The final version of the paper will also address more complex questions related to feasibility in fairness-sensitive design, such as ones concerning the identification of secondary duties that might arise when agents cannot evaluate and/or improve ML models in terms of fairness due to feasibility constraints.

Acknowledgements

I thank professor Arto Laitinen for providing thorough feedback and valuable comments on the working paper which I have outlined in this shorter version of the paper.

References

[1] Bales, R. Eugene. "Act-Utilitarianism: Account of Right-Making Characteristics or Decision-Making Procedure?". American Philosophical Quarterly 8.3 (1971): 257–265. https://www.jstor.org/stable/20009403.

[2] Chahboun, Naima. "Ideal Theory and Action-Guidance: Why We Still Disagree". Social Theory and Practice, 45.4 (2019): 549–578. https://www.jstor.org/stable/45276660.

[3] S. Fazelpour, Z. C., Lipton, Z. C., Algorithmic fairness from a non-ideal perspective, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, Association for Computing Machinery, New York, NY, USA, 2020, pp. 57–63. https://doi.org/10.1145/3375627.3375828.

[4] W. Fleisher, What's Fair about Individual Fairness?, in: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, Association for Computing Machinery, New York, NY, USA, 2021, pp. 480–490. https://doi.org/10.1145/3461702.3462621.

[5] K. Holstein, J. Wortmann Vaughan, H. Daumé, M. Dudik, H. Wallach, Improving fairness in machine learning systems: What do industry practitioners need?, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA, 2019, pp. 1–16. https://doi.org/10.1145/3290605.3300830.

[6] A. Z. Jacobs, H. Wallach, Measurement and fairness, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, 2021, pp. 375–385. https://doi.org/10.1145/3442188.3445901.

[7] J. Kleinberg, S. Mullainathan, M. Raghavan, Inherent trade-offs in the fair determination of risk scores. arXiv preprint, 2016. arXiv:1609.05807.

[8] M. S. A. Lee, J. Singh, The Landscape and Gaps in Open Source Fairness Toolkits, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA, 2021, pp. 1–13. https://doi.org/10.1145/3411764.3445261.

[9] Mittelstadt, Brent. "Principles alone cannot guarantee ethical AI." Nature machine intelligence 1.11 (2019): 501–507.

[10] Morley, Jessica, Anat Elhalal, Francesca Garcia, Libby Kinsey, Jakob Mökander, and Luciano Floridi. "Ethics as a Service: A Pragmatic Operationalisation of AI Ethics." Minds & Machines 31.2 (2021): 239–256. https://doi.org/10.1007/s11023-021-09563-w.

[11] L. Munn, The uselessness of AI ethics, AI Ethics (2022): 1–9. https://doi.org/10.1007/s43681-022-00209-w.

[12] North, Richard. "Principles as guides: The action-guiding role of justice in politics". The Journal of Politics 79.1 (2017): 75–88. https://doi.org/10.1086/687286.

[13] E. Seger, In defence of principlism in AI ethics and governance, Philosophy & Technology 35 (2022): 45. https://doi.org/10.1007/s13347-022-00538-y.

[14] Smith, Holly. "Using moral principles to guide decisions". Philosophical Issues 22 (2012): 369–386. https://www.jstor.org/stable/41683078.

[15] R. L. Thomas, D. Uminsky, Reliance on metrics is a fundamental challenge for AI, Patterns, 3.5 (2022). https://doi.org/10.1016/j.patter.2022.100476.

[16] S. Verma, J. Rubin, Fairness definitions explained, in: Proceedings of the International Workshop on Software Fairness (FairWare '18), Association for Computing Machinery, New York, NY, USA, 2018, pp. 1–7. https://doi.org/10.1145/3194770.3194776.