When the Ideal Does Not Compute: Nonideal Theory and Fairness in Machine Learning

Otto Sahlgren [1]

[1] *Tampere University, Tampere, Finland*

Abstract

Recent studies have identified numerous shortcomings in the dominant methodology for evaluating and implementing fairness in machine learning, which have been explicitly or implicitly attributed to the field's "ideal mode of theorizing". Similarly to ideal theories of justice, fair machine learning methods have been alleged to be unsuitable or practically irrelevant as tools for evaluating and enhancing justice in circumstances characterized by structural injustice and feasibility constraints. This has been taken to indicate the need for a methodological turn in the field, similarly to the turn towards 'nonideal theory' in normative political philosophy. Drawing on philosophical literature on ideal and nonideal theory, this paper examines what a nonideal mode of theorizing in fair machine learning would look like in practice. The key contribution of the paper is an outline of six possible nonideal approaches which are connected to established conceptions of nonideal theory in political philosophy and illustrated with examples from recent research on fair machine learning. The paper then suggests that the dominant "ideal" approach and the six different kinds of "nonideal" alternatives are not fundamentally incompatible. They are grounded in diverging theoretical and practical aims and therefore address different kinds of questions, employ different kinds of idealizations. Still, they offer mutually complementary perspectives to evaluating and promoting fairness in (socio)technical machine learning systems. Overall, the paper contributes to current debates on fair machine learning methodology by bridging literatures in political philosophy and fair machine learning and by depicting a broader landscape of methodologies at the field's disposal.

Keywords

algorithmic fairness, fair machine learning, ideal theory, methodology, nonideal theory, political philosophy

1. Introduction

From domains such as medical care and finance to criminal justice and education, decision-makers increasingly use predictive models trained with machine learning (ML) algorithms to generate assessments and predictions, which often inform consequential decisions about individuals' access to important goods and opportunities. Predictive models may inherit and learn problematic biases from their training data, including but not limited to biases that track past and present injustice, introducing risks for unfair discrimination and inequitable outcomes in real-life prediction-based decision-making [1]. Fair ML research (sometimes called *algorithmic fairness*) has sought to address these issues, proposing numerous measures and methods for evaluating and implementing values such as fairness, equality, and justice in predictive models and decision-making algorithms [2, 3]. However, critical studies have highlighted numerous problems that pertain to the application of dominant fair ML methods as means to address unfairness and other algorithmic wrongs [4–10]. These problems raise important *methodological* questions about their relevance and usefulness in assessing, preventing, and rectifying potential injustices in real-life settings. Following Sina Fazelpour's and Zachary Lipton's [4] example, this paper situates these critiques and shortcomings in debates surrounding *ideal* and *nonideal modes of theorizing* in normative political theory (see [11–

EWAF'24: European Workshop on Algorithmic Fairness, July 01–03 2024, Mainz, Germany.
EMAIL: otto.sahlgren@tuni.fi
ORCID: 0000-0001-7789-2009

13]). In particular, it provides an account of how nonideal modes of theorizing could be harnessed in the fair ML, outlining six approaches and their connections to the philosophical literature on ideal and nonideal theory.

The paper is structured as follows. Section 2 begins with a brief overview of fairness criteria, metrics, and debiasing techniques, and then proceeds to contextualize methodological debates in fair ML against the backdrop of discussions surrounding non/ideal modes of theorizing in normative political theory. There, I further bridge the fair ML and political theory literatures and their respective debates, demonstrating connections between (i) the dominant methodologies in both fields, (ii) critical arguments that have been advanced against those approaches, and (iii) proposals for methodological reorientation in both fields. Section 2 concludes with the observation that, though numerous political theorists have called for a methodological turn towards *nonideal theory*, little consensus has been achieved on the exact nature of nonideal theory and its relation to ideal theory (see [11, 12]). This lack of consensus raises the corresponding "pragmatic question of what precisely a *non-ideal* approach [to fair ML] might look like in practice" [4, p. 62; italics added]. Furthermore, it brings into question the extent to which the theoretical resources (e.g., extant fairness criteria) and practical prescriptions (e.g., bias mitigation strategies) offered by dominant fair ML approaches remain relevant or applicable in nonideal decision-making settings (if at all). This paper answers both of these questions. The primary contribution of this paper is found in Sections 3 and 4 where, drawing on philosophical literature around non/ideal theory, I outline six different ways of doing 'nonideal theory' in fair ML which correspond to existing conceptions regarding the nature and aims of nonideal theory. I describe the motivations of each approach and illustrate them with examples from the fair ML literature. The outlined approaches notably depart from the dominant fair ML methodology in different ways. The three approaches described in Section 3 reimagine how the ideals operationalized by fairness metrics are specified and employed to guide ML model evaluation and fairness-enhancing intervention design. Three further approaches described in Section 4 theorize fair ML from a "fact-sensitive" perspective, seeking to refrain from idealized modelling assumptions and to address questions that arise in decision-making settings with nonmarginal noncompliance, for instance. Section 5 draws on Alan Hamlin and Zofia Stemplowska's account [11] to propose a taxonomy of theory that clarifies the relationship between ideal and nonideal approaches to fair ML. They are not fundamentally distinct but rather form a continuum of approaches to implementing values (or ideals) in (socio)technical systems. Though different approaches are characterized by different aims, they can complement one another. The final section summarizes the paper's contributions.

2. The methods of fairness in nonideal circumstances

Most work on fair ML follows a pattern which can characterized roughly as follows: First, a set of criteria for fairness is adopted or proposed and operationalized into a corresponding set of fairness metric(s) which the decision-maker then uses to evaluate available ML models. Second, if unfair bias is identified according to the metric(s), the decision-maker implements a debiasing technique to mitigate or eliminate that bias and, therefore, to secure the model's fairness. Noting this pattern, Sina Fazelpour and Zachary Lipton argue that there is "a connection between the recent literature on fair machine learning and the ideal approach in political philosophy" and suggest that the shortcomings of the dominant approaches, as documented in a number of critical examinations, "reflect broader troubles faced by the ideal approach" [4, p. 57]. This paper further bridges the two literatures by showing connections between (i) the methodologies of dominant approaches in both fields, (ii) critical arguments advanced against those approaches, and (iii) calls for methodological reorientation in both fields. This section will take first steps in this regard by introducing dominant approaches to fairness in ML, the so-called non/ideal theory demarcation, related debates in political theory, and by briefly outlining the previously mentioned connections. The shortcomings of ideal approaches are also discussed in more detail in Sections 3 and 4.

2.1. Fairness in machine learning

Research on fairness in ML has proposed numerous criteria for fairness which express conditions that a predictive model should satisfy to meet some broader notion or principle of *fairness* (see [2, 3]). In this literature, 'fairness' is often understood in terms of non-discrimination, but also in other terms of other egalitarian notions and justice [14]. For this reason, I will throughout this paper use the term 'fairness' as it is used in the fair ML literature, where the term encompasses a broad range of normative egalitarian considerations. Fairness criteria are generally formalized as a property of the joint distribution of the features included in the model (X), sensitive attributes such as 'race' or 'gender' (A), predictions viewed as numeric scores or classifications ($\hat{Y}$), and the "ground truth" labels (Y) representing the outcome, property, or behavior that is being predicted. Existing fairness criteria come in different flavors, representative of different ways to model fairness in prediction settings. Table 1 mentions three popular classes of fairness criteria and provides some examples of metrics for binary classification tasks, where classifications are denoted with $\hat{Y} = [0, 1]$, actual outcomes ("ground truth") with $Y = [0, 1]$, sensitive attributes with $A = [0, 1]$, and decision subjects with I.

Table 1
Fairness criteria and examples of metrics for binary classification tasks.

Class of fairness criteria and examples of metrics
Statistical Criteria. Fairness is defined as parity in a model performance statistic between a set of comparison classes. The equalized performance metric depends on the definition.
• *Statistical Parity* [15]: The likelihood of receiving a positive prediction is probabilistically independent of the value of A. Formally: $P(\hat{Y} = 1 \mid A = 0) = P(\hat{Y} = 1 \mid A = 1)$.
• *Predictive Parity* [16]: The likelihood of receiving a true positive prediction is probabilistically independent of the value of A. Formally: $P(\hat{Y} = 1 \mid Y = 1, A = 0) = P(\hat{Y} = 1 \mid Y = 1, A = 1)$
• *Equalized Odds* [17]: The likelihood of receiving a false positive or false negative prediction is probabilistically independent of the value of A. Formally: $P(\hat{Y} = 1 \mid Y = 0, A = 0) = P(\hat{Y} = 1 \mid Y = 0, A = 1)$ and $P(\hat{Y} = 0 \mid Y = 1, A = 0) = P(\hat{Y} = 0 \mid Y = 1, A = 1)$.
Counterfactual and Causal Criteria. Fairness is operationalized in terms of counterfactual expectations or causal pathways between model features and the classification.
• *Counterfactual Fairness* [18]: The classification $\hat{Y}$ should be insensitive to permuting the value of A. Formally: $\hat{Y}_i(A = 0) = \hat{Y}_i(A = 1)$ for all individuals I.
Similarity-based Criteria. Fairness is defined in terms similar decision subjects receiving similar classifications.
• *Fairness Through Awareness* [15]: Individuals I who are similar in terms of a pre-defined set of model features X (where X excludes A) receive similar classifications $\hat{Y}$. Precise formalization depends on the applied measure of similarity (or distance).

For instance, *counterfactual* and *causal criteria* define fairness in terms of (un)acceptable dependencies between model features and the output. *Statistical criteria*, in turn, define fairness in terms of parity between compared groups with respect to some statistical measure of model performance (see Table 2). Whereas the former class of criteria depicts an ideal set of causal relationships between input features and the output that a fair predictive model should exhibit, the latter class operationalizes, albeit in different ways, the notion that information about individuals' sensitive attributes should not affect their treatment, where treatment is understood in terms of the received (correct or incorrect) classifications or predictions. In practice, the preferred set of formal fairness criteria is then further specified into a set of fairness metrics which the decision-maker uses to evaluate, compare, and select between available ML models. Unfair bias that is identified by employing the metric(s) is commonly eliminated, or at least mitigated, with the help of debiasing techniques. Existing techniques can be

distinguished into three classes based on their points of intervention: (i) pre-processing techniques balance or resample the model's training data, (ii) in-processing techniques adjust or constrain the learning algorithm, and (iii) post-processing methods intervene directly on the model outputs (see [2]).

Table 2
Confusion matrix of model performance measures.

	Positive actual class ($Y = 1$)	**Negative actual class ($Y = 0$)**
Positive predicted class ($\hat{Y} = 1$)	**True positive (TP):** $\hat{Y} = 1, Y = 1$ Positive predictive value: TP / (TP + FP) True positive rate: TP / (TP + FN)	**False positive (FP):** $\hat{Y} = 1, Y = 0$ False discovery rate: FP / (TP + FP) False positive rate: FP / (FP + TN)
Negative predicted class ($\hat{Y} = 0$)	**False negative (FN):** $\hat{Y} = 0, Y = 1$ False omission rate: FN / (TN + FN) False negative rate: FN / (TP + FN)	**True negative (TN):** $\hat{Y} = 0, Y = 0$ Negative predictive value: TN / (TN + FN) True negative rate: TN / (TN + FP)

A growing body of critical studies has documented various shortcomings related to the application of fair ML methods as means to assess and secure non-discrimination, equality, and justice in ML-supported decision-making [4–10]. In some critiques, the shortcomings are explicitly traced to the *ideal* mode of theorizing employed by dominant approaches. For instance, Fazelpour and Lipton [4] argue that fair ML engages in "small-scale" ideal theorizing and Davis et al. [6] argue that dominant approaches operate under a misguided "algorithmic idealism". To contextualize these claims and the identified shortcomings of standard approaches, we need to take a detour into debates on non/ideal theory in political philosophy.

2.2. Non/ideal theory on larger and smaller scales

John Rawls famously distinguishes the theory of justice into two complementary parts: The first one is ideal theory which "assumes strict compliance and works out the principles that characterize a well-ordered society under favorable circumstances" [19, p. 245]. If properly defined, the principles constitutive of 'perfect justice', Rawls suggested, specify both a *normative target* towards which societies should aspire and provide a set of *evaluative standards* in light of which current institutions and social arrangements can be assessed (see also [20]). The second part, nonideal theory, starts from more realistic assumptions (e.g., nonmarginal noncompliance to the principles) and assumes less favorable societal conditions (e.g., historical injustice). It seeks to explain how actual (unjust) societies could be made more just, proposing principles that should govern "adjustments to natural limitations and historical contingencies" (e.g., accommodations for people with disabilities) and responses to injustice (e.g., retributive justice, just warfare) [19, p. 246]. There is an order of primacy here, Rawls suggests: once ideal theory fixes an ideal conception of justice that would regulate societal arrangements in ideal circumstances, nonideal theory can devise justifiable principles for addressing problems that occur in less ideal societies.

Note that dominant approaches to fair ML conform primarily to the first, ideal part of the Rawlsian demarcation. The adopted set of fairness criteria for predictive models are commonly treated as specifying or operationalizing some broader 'ideal' (or general notion) of fairness, and therefore as providing both (i) a *yardstick* for evaluating predictive models and their expected distributive patterns (i.e., a fairness metric) and (ii) a *normative target* they should satisfy or at least approximate (i.e., a set of fairness requirements). Upon evaluation, "[t]he magnitude of (dis)parity measured by a given fairness metric is taken to denote the degree of divergence from the ideal for which that metric is supposed to be a formal proxy" [4, p. 59]. If an unfair disparity is detected, a debiasing technique is typically applied to produce a new ML model that maximizes overall model performance subject to the satisfaction of the metric(s) [5, p. 48]. Deviations from the 'ideal' are, in other words, treated as instances of *pro tanto* unfairness or injustice which can and should be addressed, at least in part, by

closing the discrepancy between the available predictive model and one that satisfies the preferred metric(s).

Rawls' writings sprouted a number of parallel and ongoing debates regarding the non/ideal theory demarcation as well as the relevance and practical usefulness of ideal theory for the purpose of enhancing justice in the real world. Rawls' demarcation has been subject to heated debate (see [11, 12]) and some theorists reject the idea that ideal theory is somehow theoretically primary in relation to nonideal theory (e.g., [21]). For a large part, the following debates have revolved around claims that ideal theory has restricted use in actual, nonideal circumstances, if at all. For instance, ideal theory has been argued to distort our understanding of justice by misrepresenting actual injustices such as racial oppression [22] and to provide misguided or unattainable prescriptions about how we should secure justice when 'perfect justice' is infeasible [20, 23–25]. In addition, some argue that knowing what constitutes 'perfect justice' is both unnecessary and insufficient for determining what is (un)just in our actual circumstances [20]. In response to these claims, many political theorists have since called for a methodological turn towards political theory which would forego or somehow depart from the central tenets of ideal approaches (e.g., [20]). In particular, theorists have proposed a turn towards *nonideal* theory which would be "realistic" and "fact-sensitive", action-guiding and feasible to implement, and capable identifying and addressing past and present injustices (e.g., [22, 23] and see also [11, 12]).

Many of the concerns expressed in regard to standard fair ML methods align with the previously mentioned problems that supposedly plague ideal modes of theorizing. The connection is explicit in certain works which note that the characteristically ideal modes theorizing of fair ML research will "always be inadequate in a context that is fundamentally unjust" [6, p. 2] and provide "ineffective solutions to current injustices" [4, p. 58]. Further critical studies on fair ML do not explicitly reference the non/ideal theory debate, but advance arguments which closely resemble ones that have been directed against ideal theories of justice in the political domain. For instance, critiques of idealization and unreasonable abstraction are commonly employed against paradigmatic theories of justice (see [11, 12, 22, 23]) but also more recently against dominant approaches to fair ML [4, 6, 8, 10]. These shortcomings and critical arguments will be discussed in detail in Sections 3 and 4. For now, it is useful to also note parallels between political theorists' proposals to shift the focus towards nonideal theory and recent proposals seeking to reorient fair ML research and practice, though, again, the latter proposals are not always explicitly advanced under the label of 'nonideal theory'. Indeed, the latter kinds of proposals are grounded in similar views as the former, suggesting that fair ML should be approached from a "realistic" [8], "fact-sensitive" [4] or "sociotechnical" perspective [10], on the one hand, and stating that the methods employed to evaluate and design predictive models and decision-making algorithms should be equipped to address injustices both past and present, on the other [4, 6, 9].

2.3. The methodological turn in fair machine learning and its challenges

Debates surrounding non/ideal theory are infamously obfuscated and lacking in consensus. Here, I highlight two open questions that are relevant also for debates on the methodology of fair ML. On the one hand, political theorists have struggled to specify what exactly distinguishes nonideal theory from ideal theory, if anything. Nonideal theory has traditionally been defined negatively, as departing from some particular defining feature(s) of ideal theory, but some argue that existing demarcations do not capture the complexity of the surrounding debates and concerns (see [11, 12]). Though extant proposals to reorient fair ML methodology are not always explicitly motivated by nonideal theory, most of them employ a similar strategy: they identify and seek to address some particular gap(s) or shortcoming(s) of the dominant approach (e.g., infeasibility, incapacity to handle past and present injustice). This strategy can be beneficial in and of itself, but it leaves open the "pragmatic question of what precisely a non-ideal approach [to fair ML] might look like in practice" [4, p. 62]. On the other hand, political theorists have not achieved consensus on whether nonideal theory *requires* or perhaps even *needs* ideal theory (i.e., what is the relation between ideal and nonideal theory). Some argue that nonideal theory is all there is (e.g., [21]). Others claim that ideal theory does not equip us with the

theoretical means to address injustice specifically in nonideal circumstances (e.g., [20, 22]). Still others suggest that nonideal theory need not entirely dispense with ideals; rather, it must understand their functions differently in such circumstances (e.g., [26, p. 5]). In particular, it might be that "we need to interpret the ideal-theoretical principles in a context with nonideal circumstances" or instead "develop a new set of nonideal principles of justice […] by adding layers of relevant facts from the nonideal world to the ideal theory, using the 'theoretical resources' that are available in the ideal theory" [27, p. 348]. Similarly, then, if fair ML is due a methodological turn, we must address the question of whether and to what extent the theoretical resources (e.g., fairness notions and fairness criteria) and practical prescriptions (e.g., bias mitigation strategies) offered by dominant approaches are relevant or applicable in nonideal decision-making settings.

In the remainder of this paper, I outline answers to both questions. Drawing on both critical examinations of fair ML methods and philosophical literature on non/ideal theory, I describe four general approaches to theorizing fair ML from a *nonideal* perspective (summarized in Table 3 below). I contextualize their motivations, examine their central features, and demonstrate how they track existing demarcations of non/ideal theory and justifications for nonideal theory in political philosophy (see [11, 12]). The first three options (Section 2) depart from the dominant ideal approach in terms of how metrics for evaluating predictive models are specified and/or how they function in the praxis of model evaluation and improvement. The fourth set of options centers on the modelling assumptions employed upon evaluating and implementing fairness, aligning with the idea that nonideal theory is "fact-sensitive" or "realistic" and thus contrasts with ideal approaches which employ considerable methodological abstractions and idealizations. I will describe three possible and mutually compatible ways in which an approach to fair ML might be considered 'fact-sensitive' (Section 4). There may be further conceptions of nonideal theory applicable to fair ML; I aim to only provide a neutral and general description of some prominent ones. Furthermore, I do not claim that the works on fair ML cited throughout the discussion actually advocate the discussed conceptions of nonideal theory, though I will mention some works that do.

Table 3
Possible nonideal approaches to fairness in machine learning contrasted with ideal approaches.

Ideal mode	Nonideal mode
Positive approach: Identifying and implementing requirements for fairness, equality, and justice (or related values).	**Negativist (or critical) approach:** Identifying and mitigating *un*fairness, *in*equality, or *in*justice (or related wrongs).
Perfectionist / End-state approach: Identifying the fairest possible model and implementing it, often via a one-shot debiasing intervention.	**Comparativist approach:** Identifying and implementing the *comparatively* fairest model within the feasible set. **Transitional approach:** Long-term improvement (or satisfaction) of fairness metrics by identifying and implementing feasible that satisfy *transitional* fairness requirements.
Fact-insensitive approach: Identifying and implementing fairness requirements and fairness-enhancing interventions under idealized assumptions and significant abstractions (e.g., strict compliance, absence of historical injustice).	**Fact-sensitive (or realist) approach:** Identifying and implementing fairness requirements and fairness-enhancing interventions under less idealized assumptions (e.g., past, present, or future partial compliance) and informed by comprehensive empirical models (e.g., sociotechnical context).

3. The evaluative and normative function of the ideal

Ideal theory has been attributed numerous labels ranging from 'perfectionist' or 'utopian' theory to 'transcendental institutionalism' and 'end-state' theory (see [11, 12]). These labels underscore that

paradigmatic theories of justice seek to answer the question of what constitutes 'perfect justice', usually by identifying the societal arrangement(s) that would realize justice so defined (e.g., [19]). They are often also attributed two views: first, that principles constitutive of 'perfect justice' provide the metric(s) for evaluating whether and to what extent social arrangements realize justice and, second, that they specify a normative goal which should guide decisions between available arrangements or justice-enhancing reforms (e.g., [20, 24]). Recall here that dominant fair ML methodologies operate under similar understandings, often proposing or adopting some notion and corresponding measure(s) of fairness which then functions as an evaluative standard and a target to be achieved through debiasing (see [4, 28]). For these reasons, they are subject to similar shortcomings as ideal theories of justice [4]. For instance, just as implementing a perfectly just societal arrangement can be infeasible given our current circumstances (e.g., [20]), satisfying the kinds of parity requirements prescribed by most fairness criteria is oftentimes unattainable in realistic modelling settings due to tradeoffs [16, 29]. 'Perfectly fair' models which satisfy all relevant metrics are often simply beyond reach, in other words, and hence decision-makers lack actionable guidance for determining which measures of fairness should give [4]. The following subsections describe three ways of re-envisioning how evaluation metrics are specified and how they function *qua* regulative 'small-scale ideals' in nonideal settings of model evaluation, improvement, and selection. The described methodologies for fair ML correspond to *negativist*, *comparativist*, and *transitional* understandings of nonideal theory, respectively (Table 3).

3.1. Negativism: Explaining and mitigating *un*fairness

In political theory, the non/ideal theory distinction has been defined as a distinction between (a) theories of justice that seek to identify and implement societal arrangement that realize full justice and (b) theories which seek to explain and mitigate *in*justice, respectively. The latter conception equates nonideal theory to *negativist theory* (or *critical theory*) grounded in the notion that we can "recognize the existence of a problem before we have any idea of what would be best or most just" [26, p, 3]. Nonideal theory so defined does not develop or seek to implement a positive ideal of justice (or its constitutive principles), but instead seeks to explain and mitigate salient *in*justices, often focusing on *local* as opposed to *global* improvements to justice [11, pp. 51–52].

A corresponding negativist methodology for fair ML research and practice would omit *positive* ideals (e.g., models of fairness in ML), including their supposed action-guiding functions in the evaluation and design of ML models. Rather, it would rather seek to explain what constitutes *un*fairness, *in*justice, or other algorithmic wrongs, and develop and implement effective approaches to mitigating or eliminating them. A closest example along these lines comes from Hutchinson and Mitchell who suggest that fair ML research should de-center efforts to formalize and achieve fairness and instead prioritize the development of "methods to *explain* and *reduce* model unfairness by focusing on the causes of *unfairness*" [30, p. 57].

Proponents of the negativist methodology might find it compelling for two reasons. First, negativist approaches are capable of producing actionable insights about how ML models could be improved in terms of justice. Positive ideals (e.g., fairness criteria) are not required to identify and correct *un*fair or *un*just distributive patterns in predictions or otherwise wrongful model behaviors. A negativist approach can make do with a theory which explains whether a particular disparity or model behavior is unjustifiable, for instance, and why it warrants mitigation, accordingly. Second, negativism also aligns with public discussions and activism around social justice which are often centered on correcting particular *in*justices rather than realizing some comprehensive political ideal [27].

3.2. Comparativism: Ranking feasible models

The non/ideal theory distinction has also been considered to reside between (a) theory that seeks to identify perfectly just societal arrangements and (b) theory that concentrates on ranking alternative societal arrangements [11, pp. 51–52; 12]. The latter conception equates nonideal theory with theory that applies the *comparativist methodology* originally proposed by Amartya Sen in response to two

shortcomings of what he calls "transcendental justice" [20]. On the one hand, Sen argues that knowing what constitutes 'perfect justice' is *insufficient* for evaluating and improving actual societies in terms of justice because feasible societal arrangements may be equally far from that ideal but in different respects[2] [20]. On the other hand, it is also *unnecessary* for making comparative assessments of justice because we can rank feasible societal arrangements without identifying the fully just societal arrangement(s) [20].

The problems noted by Sen also arise in the context of designing fair ML models, albeit at a smaller scale. For instance, empirical studies have demonstrated that any feasible ML model will in realistic decision-making settings exhibit *some* kind of disparity that violates some plausible measure of fairness [16, 29]. This means that "matching the ideal in some respect may only be possible at the expense of widening gaps in others" and that decision-makers will require a further "basis for deciding which among competing discrepancies to focus on" [4, p. 62]. Furthermore, the normative targets expressed by most fairness criteria (e.g., "zero statistical disparity") are not strictly speaking required to rank the *feasible* models that decision-makers *can* build; the evaluator only needs some metric for (un)fairness. In other words, knowing what constitutes an ideal distribution is unnecessary for the purpose of comparing available models in terms of their respective disparities (see [28]).

Sen's solution [20] is a comparativist justice methodology where feasible societal arrangements are compared pairwise to produce a ranking that can be used to identify the "second-best" option (see also [25]). An analogical comparativist approach to fair ML would (i) omit the notion that fairness criteria express strict normative targets that should guide the improvement and selection of models and instead (ii) use some set of positive (or negative) metrics to compare feasible models in terms of their (un)desirable distributive patterns or behaviors[3]. This kind of an approach has been defended by Lee and Floridi who propose "a new methodology that views fairness as a trade-off of objectives—not as an absolute mathematical condition—but in relation to an alternative decision-making process" [28, p. 166]. This comparativist methodology is not premised on the goal of satisfying a set of fairness criteria, in other words, but focuses on producing a model that is both feasible and "better on multiple dimensions in relation to any existing process or model" [28, p. 188].

3.3. Transitional fairness: Fairness as a long-term target

A third distinction that has been drawn defines ideal theory as *end-state theory* and nonideal theory as *transitional theory* (see [12, pp. 660–662; 13]). This distinction has gained salience due to critiques of ideal theory which observe that implementing a fully just societal arrangement under nonideal background conditions is often infeasible and prone to produce negative externalities [20; 25]. The conception of transitional theory *qua* nonideal theory aims to address these problems. It treats the ideal specified by ideal theory as a stable *long-term* normative target which should be realized eventually, even if doing so is currently not possible. The goal of transitional theory *qua* nonideal theory, then, is "to create a world in which the ideal theory can be applied" [41, p. 487]. To achieve this goal, it must explain what should be done here and now to improve justice relative to the ideal specified by ideal theory. Here, transitional theory recognizes that "our feasibility sets for political action are open to temporal variation" and that "what is not feasible now may become feasible in the future if we take steps to expand our [...] capabilities" [31, p. 47]. However, the justice-enhancing improvements available to decision-makers at a given time will, in nonideal circumstances, involve *some* unavoidable short- and long-term costs. Transitional theory must therefore identify feasible and effective courses of action which are also *permissible*: they must strike an appropriate compromise between the associated benefits and costs or balance them somehow [13]. This comprises the crux of *transitional fairness*.

A transitional approach to fair ML would treat fairness criteria as expressing requirements that ML models should ultimately satisfy or at least approximate, even if current models cannot. Here, "debiasing" is no longer conceptualized as a one-shot intervention, but instead as a long-term project consisting of a sequence of *incremental* (non-)computational interventions which improve fairness

[2] This is *the problem of the second-best*: "if one of the background social conditions assumed when analyzing a political ideal does not obtain, then the (normatively) best […] distributive profile does not necessarily satisfy the principles that characterize a fully just [one]" [25, p. 133].
[3] Even a comparative approach requires *some* standard against which justice is evaluated (see [27]). However, a comparative approach can proceed without a comprehensive and positive ideal of fairness or justice in its search for "second-best" options.

relative to the long-term target. This creates a need to assess available models' effectiveness in terms of contributing to the achievement of that target, on the one hand, but also to identify the short- and long-term costs and benefits associated with deploying a particular model at a particular point in time, on the other. Indeed, studies have observed that debiasing interventions that mitigate some group-level disparity can in the long run harm the group that initially benefitted from those interventions (e.g., [32]). A transitional approach must therefore consider whether interventions that reduce injustice in the short run may decrease overall justice down the line (see also [27, p. 160] and *vice versa*. Emerging work on "long-term" and "dynamic" fairness in ML in sequential decision-making settings (e.g., [5, 32]) provides important resources for such purposes, including computational tools for run-time fairness monitoring (e.g., [33]) which may prove useful in practice. However, it serves to note that the respective short- and long-term impacts of particular models may vary both synchronically (at a particular time t) and diachronically (through time $t, \ldots t_n$) due to changes in the modelled population or the deployment setting, including due to the deployment of the model [5, 10]. Transitional approaches must therefore also address substantive questions about *transitional fairness* (e.g., what is justifiable configuration of short- and long-term impacts on different groups) and incorporate novel methods (e.g., metrics) for evaluating the transitional fairness of particular models as means to improve overall fairness relative to the long-term target.

4. "Fact-sensitive" and "realistic" fairness

In political philosophy, a large part of the debate around non/ideal theory revolves around the notions of *abstraction* and *idealization* [11, 12] which refer to (i) the bracketing of complexity and (ii) the incorporation empirically false assumptions about the world or subject matter, respectively. Sure enough, abstraction and idealization are common in both empirical and normative domains of theory. Frictionless surfaces are found practically nowhere in the real world, for instance, but theoretical physics models employ such idealizations for multiple theoretical and practical purposes. Paradigmatic ideal theories employ considerable abstractions and idealized modeling assumptions – such as the assumption that agents act according to the principles of justice presented in the theory [11, p. 49; 19] – for similar purposes. Reasonable critics of ideal theory have no problem with abstraction and idealization *per se*. Still, many of them argue that simplifying modeling assumptions – whether false or not, empirically speaking – can leave us with normative theory that misrepresents past or present injustices [22] and prescribes courses of action that prove ineffective or infeasible to follow in realistic settings [23]. Similar concerns have been expressed in critical studies on fair ML methodology [4, 7, 8] which examine how excessive abstraction can lead to "ineffective, inaccurate, and sometimes dangerously misguided" design prescriptions and technical debiasing interventions [10, p. 59].

The non/ideal theory distinction has been used to distinguish between *fact-insensitive* and *fact-sensitive* theory, accordingly, as well as between *utopian* and *realistic* theory (see [11, 12]). Recent studies have similarly proposed that fair ML methodology should be reoriented towards fact-sensitive [4] and realistic [8] directions. However, the distinction between abstraction and idealization has proven murky and even misleading in many cases. The key question that requires attention is rather whether specific kinds of simplifications and modelling assumptions are reasonable and appropriate in a particular context of normative theorizing [11, pp. 50–51]. Indeed, political philosopher Laura Valentini notes that, "[s]ince the relevant facts will vary on a case-by-case basis, it is almost impossible to come up with a *general* rule prescribing what the correct level of idealization in normative theorizing should be" [12, p. 660]. Critical studies on fair ML have argued in similar vein that the field should further reflect on "what abstractions are reasonable, which simplifying assumptions are justified, and what formalizations are appropriate" [4, p. 62]. I will not attempt to provide a general rule for appropriate idealizations here. Instead, I will outline three potential instantiations of fact-sensitive or realistic theorizing in fair ML, drawing on different kinds of idealizations that have been criticized in the domain of political theory. Though I discuss these approaches separately, they are not mutually incompatible and indeed may overlap in numerous ways.

4.1. Avoiding upstream idealization

Charles Mills' influential critique of ideal theory observes that theories of justice often neglect "actual historic oppression and its legacy in the present, or current ongoing oppression" [22, p. 168]. If theory starts from an ideal model of the world, Mills suggests, it ends up "abstracting away from realities crucial to our comprehension of the actual workings of injustice in human interactions and social institutions" [22, p. 170]. In these passages, Mills highlights that justice-enhancing interventions informed not by an accurate description of the actual, nonideal world but instead by an idealized world that could be can potentially reproduce the injustices and inequalities it seeks to address[4]. Critical studies on fair ML notably observe similar problems. For instance, Herington [9] demonstrates that many standard measures of fairness can be satisfied even when model predictions track historical injustice, and that they are oblivious to cases where a sensitive attribute has an effect on model predictions via an unmodelled variable. In similar vein, Fazelpour and Lipton [4] note that statistical measures of fairness are indifferent to the *history* and *causes* of specific group-level disparities, and hence provide insufficient evidence for decision-makers to determine whether a given disparity should be mitigated.

The problems implicated here can be understood as resulting from so-called *upstream idealization*: the bracketing or misrepresentation of (facts concerning) past injustice and its effects on the present population and/or the observed data. The term 'upstream' is used to capture the notion that the (potentially inappropriate) modelling assumptions in these cases concern factors that causally and/or temporally precede the observed ML model. The problems identified with regard to upstream idealization suggest the need for fact-sensitive fair ML understood specifically as an approach that seeks to both *recognize* and *correct* past and present injustice and their continuing effects. In contrast to many fair ML methods which may simply reproduce "structural and historical stratifications as they manifest in computational code" [6, p. 4], this approach seeks to develop an empirically grounded understanding of the causal pathways that produce the stratifications which manifest as disparities at the level of model outputs [4, p. 62]. Furthermore, the prescribed debiasing interventions are in this approach guided by the explicit aim of repairing past injustice and eliminating its influence on current populations and, therefore, the predictions that are made based on models of those populations [6].

4.2. Avoiding downstream idealization

What can be called *downstream idealization* refers to the bracketing or misrepresentation of (facts concerning) the factors that mediate the realization of the specified ideal. Downstream idealization is notably characteristic to research on fair ML, where dominant approaches do not seek to "model the entire system over which a social criterion, such as fairness, will be enforced" [10, p. 60] but rather specify and implement the value(s) at the level of the technical (sub)systems (e.g., software) or their underlying components (e.g., datasets and predictive models). This narrow focus may contribute to an insufficient understanding of downstream factors which can affect the system's effectiveness in terms of realizing fairness in practice. For example, it is commonly assumed for purposes of simplification that the predictions produced by a fair predictive model will also translate into fair *decisions* or *outcomes* for decision subjects and other stakeholders. However, this assumption does not necessarily hold since human users of ML systems can misinterpret system outputs or introduce novel biases, for instance, creating a discrepancy between predictions and decisions [34].

One way to approach fair ML from a fact-sensitive or realistic perspective then is to minimize potentially problematic downstream idealizations that might factor into models of fairness or the practice of identifying the (expected) effects of ML models or particular fairness-enhancing interventions. In other words, the level of abstraction is extended to ensure that ML models are evaluated and (re)configured based on an empirically informed understanding of factors that affect whether and how model outputs translate into concrete harms and benefits for decision subjects. The extension of the abstraction boundary is recognized as crucial in particular because implementing fairness (or other values) at the level of a technical system (e.g., securing fair predictions) is typically

[4] This does not mean that the project of value implementation can or should be pursued without reference to ideals understood broadly as normative or evaluative conceptions of what we ought to do in some general sense. The cited passages from Mills are explicitly directed towards the empirical world-models on which normative theories are premised (see [22, p. 166]).

insufficient for realizing fairness at a broader level of practical significance (e.g., securing fair decisions or outcomes for individuals, minimizing unfair group-level inequalities) [40]. Multiple factors (e.g., user behaviors, incentive structures, and "chilling effects" created by system deployment) are likely to mediate a predictive model's actual impact in a given context, however, which means that the relevant empirical facts to be taken into consideration depend partly on the social context of the model's use. Here, harnessing interdisciplinary perspectives and methodologies (e.g., human-computer interaction, sociotechnical systems studies) becomes crucial to ensure that both the set of feasible fairness-enhancing interventions (e.g., technological but also non-technological interventions) as well as their expected impact can be properly identified and assessed (see also [4, 8, 10]). This idea aligns closely with an analogical position held by many nonideal political theorists: "social science does not enter the picture *after* we have specified directive principles" but rather "enters alongside moral theory to help with the specification of directive principles suited for nonideal circumstances" [24, p. 445].

4.2.1. Theorizing fairness under partial compliance

A final but perhaps most common distinction is between ideal theory *qua* strict compliance theory and nonideal theory *qua* partial compliance theory (e.g., [19]; cf. [23]; see also [11, 12]). In ideal theory, the so-called *strict compliance assumption* is often defended on grounds that, unless we control for noncompliance upon comparing alternative principles of justice, we cannot ascertain that those principles actually contribute to the full and stable realization of justice [19]. However, critics have noted that justice-enhancing reforms or interventions which are designed without consideration of non-compliant agents are liable to prove ineffective or misguided in actual, nonideal settings where everyone *cannot* or simply *do not* act as they ought to [23]. In addition, strict compliance theory does not provide (sufficient) solutions to numerous challenges that arise solely in nonideal circumstances. For instance, "what constitutes just punishment" or "what is a justifiable way of mitigating deep social inequalities" are questions that simply do not arise under strict compliance. Yet in a nonideal world where everyone does not do their fair share, it seems securing justice (or other values) may require some agents to do *more* (or perhaps *less*) than their fair share[5]. Partial compliance theory *qua* nonideal theory abandons the strict compliance assumption, meaning all relevant agents (e.g., citizens and institutions) are not expected to adhere to the demands of justice *qua* normative content of the proposed ideal theory, and asks how we can and should enhance justice given the expectation of nonmarginal noncompliance.

Dominant fair ML approaches align with strict compliance theory: they assume that the satisfaction of the preferred fairness metrics provides an end-to-end guarantee of fairness in the observed setting (see [10]). However, as noted above, fair predictions might be insufficient as means to realize just outcomes in case accurate predictions track unjust structural inequalities or because downstream factors (e.g., biased system users) skew the outcomes. Furthermore, noncompliance can also have broader undesirable effects at the level of an observed population. For instance, decision-makers who implement unfair decision policies can obstruct the realization of an equitable distribution of goods or opportunities between (sub)groups in a given population, even when other decision-makers do their fair share in terms of implementing fairness constraints on their respective models [36]. This underscores that coordinated and collective action is often necessary to ensure desirable outcomes, but also raises the question of whether and to what extent compliant decision-makers (or other agents) are required to pick up the noncompliant decision-makers' slack, as it were. In other words, fair ML *qua* strict compliance theory can fail to provide achievable normative prescriptions regarding in situations where noncompliance is expectable, and also fail "to delineate the responsibilities of current decision-makers in a world where others fail to comply with their responsibilities" [4, p. 60].

Partial compliance theory provides a general framework for theorizing fair ML in nonideal circumstances (e.g., [36]; see also [23]). Here, the evaluation and design of debiasing interventions is

[5] For example, Richard North describes three general conceptions concerning distributive duties under partial compliance [35, pp. 81–86]: In the first view, agent *A*'s failure to fulfill their duty towards a subject *C* does not impose extra demands on agent *B* who does their fair share by fulfilling their duty towards *C*. In the second view, *A*'s failure to comply creates a duty for *B* to do *A*'s share in addition to their own. A last view maintains that *A*'s failure to comply effectively cancels *B*'s duty to comply.

conducted under the expectation that some (socio)technical factor(s) such as noncompliant agent(s) or malfunctioning technologies obstruct the realization of the desired outcomes in an observed decision-making setting or population. This requires, first, identifying and anticipating (sources and causes of) noncompliance ranging from technical problems (e.g., bugs and glitches) to user-level problems (e.g., user-errors, cognitive bias, external attacks that affect ML models' predictions) and finally organizational problems (e.g., "fairness-washing"). Second, it requires formulating effective and normatively justifiable responses to noncompliance and its expected effects. These responses can similarly range across a number of technical and social levels. For instance, promoting compliance among ML developers may require improving their access to evaluation benchmarks and technical tools (e.g., suitable datasets), domain-expertise, and other resources required for the effective detection and mitigation of unfair bias (e.g., [37]). On broader scales (e.g., multi-agent decision-making contexts), it may instead become necessary to develop means to coordinate actions between agents (e.g., decision-makers and regulators) so that unacceptable disparities can be mitigated even if some decision-makers fail to do their fair share in this regard.

5. Fairness in machine learning: A landscape of theory

To answer the first question of this paper (see Section 2.3), I have outlined six general ways to approach fairness in ML from a nonideal perspective. Though further approaches are possible, the options outlined here capture central concerns that critical studies have expressed in relation to mainstream fair ML methods and also align with established positions in philosophical debates on non/ideal theory. Now, I turn to the second question: what is the relationship between ideal and nonideal modes of theorizing in the context of fair ML? I argue there is no categorical distinction between them, and that one's reasons for favoring a particular approach to fair ML depend on one's aims and the question(s) one is seeking to answer. Different approaches have different kinds of benefits and limitations, and there is room for both more and less ideal approaches. Here, I draw on Alan Hamlin and Zofia Stemplowska's account [11] which distinguishes between the so-called *theory of ideals* and *theory of institutional design* and suggests that debates surrounding non/ideal theory concern the latter domain which comprises a multi-dimensional continuum from ideal to nonideal theories.

Hamlin and Stemplowska [11] suggest that demarcations of non/ideal theory have been unsuccessful for two reasons. First, debates on non/ideal theory often confuse two domains of theory: (i) the theory of ideals which seeks to *explain* or *describe* some value(s) or ideal(s) and their inter-relations and (ii) the theory of institutional design which is concerned with the *implementation* of some value(s) or ideal(s). Hamlin and Stemplowska suggest the debates on non/ideal theory concern the latter domain, noting that criticisms of ideal theory are "often couched in terms of worries about impracticability, and it is social arrangements rather than ideals that are subject to considerations of practicality" [11, p. 53]. Indeed, it is obvious that *any* normative theory has to assume *some* general notion of what is valuable (i.e., a theory of ideals) as its normative basis (e.g., for moral evaluation and prescription). Even *non*ideal theory requires at least an *incomplete* conception in this regard (e.g., a partial theory of justice or *in*justice). A second reason for the lack of success in demarcating non/ideal theory is that extant distinctions "focus on one (or a small number) of the set of relevant dimensions" in which theories of institutional design can differ from one another [11, p. 60]. Hamlin and Stemplowska suggest that no binary, categorical distinction can fully capture the complexity of the debates and instead argue that "the ideal/non-ideal distinction is better construed as a multi-dimensional continuum" [11, p. 52]. A given theory of institutional design is not exclusively ideal or nonideal but rather *more* or *less* ideal, and its 'idealness' depends on numerous factors that have been used to distinguish ideal theory from nonideal theory in the philosophical debates (e.g., assumed level of compliance, fact-sensitivity). The individual factors that constitute this multi-dimensional continuum are notably also *continuous* or *gradient* as opposed to simply binary. Fact-sensitivity, for example, comes in degrees: a theory is more or less fact-sensitive depending on the number of empirical facts that are incorporated as *constraints* in the proposed normative model or theory [11, p. 51]. The same thing can be said of other factors (or modeling assumptions) implicated in the non/ideal theory debate, such as the assumed level of compliance which can range from full compliance to widespread noncompliance.

If idealness comes in degrees as measured along multiple gradient dimensions or factors, the lingering question is: what level of idealness is appropriate in normative theory? Hamlin and Stemplowska [11] note that there is no simple answer. Theories located at different ends of this continuum have *different kinds of aims*, address somewhat *different questions* (e.g., justice under strict versus partial compliance), and produce *different theoretical resources*. The practical import of a particular theory depends on the kind of information it provides but also on the circumstances of the theory's application (e.g., a just versus unjust society). Broadly speaking, theories in the more ideal end employ idealizations and considerable abstractions (e.g., assume fewer feasibility constraints) because they seek to produce normative prescriptions that are highly consistent with the ideals they are seeking to bring about. Strict compliance theory, for instance, idealizes agents' expected behavior in order to avoid baking in concessive bias towards a nonideal *status quo* where noncompliance is common [12, pp. 656–660]. Theories in the less ideal end generally start from (more) realistic assumptions to provide less-than-fully-just but actionable prescriptions and to address practical and moral challenges that arise only in circumstances of infeasibility and injustice. Partial compliance theory, for instance, assumes some realistic level of noncompliance among the relevant agents (e.g., decision-makers) precisely to identify what kinds of obstacles and tensions arise when everyone does not do their fair share, and to produce normative prescriptions that are not oblivious to those obstacles and tensions.

In what follows, I use Hamlin and Stemplowska's [11] taxonomy of normative theory to classify different approaches to fairness in ML (broadly construed). I then explain how the multi-dimensional continuum conception of the theory of institutional design clarifies the relationship between standard approaches to fair ML and the alternatives described in this paper and also makes room for the application of both more and less ideal modes of theorizing in the field.

5.1. Algorithmic fairness and fair machine learning

Drawing on Hamlin and Stemplowska's [11] taxonomy, I distinguish between two (sub)domains of theory which I will call (i) algorithmic fairness *qua* subdomain of the theory of ideals and (ii) fair ML *qua* a subdomain of the theory of institutional design, respectively.

On the one hand, I understand *algorithmic fairness* as denoting a domain of theory that examines, elucidates, and seeks to define values or ideals as they concern prediction-based decision-making procedures (e.g., fairness notions but also other related values). This domain of theory is not directly concerned with the practical implementation of the values, principles, and ideals it deals with. It is rather concerned with the *nature* of those ideals (e.g., necessary and sufficient conditions), their *moral justifications* (e.g., arguments for and against particular notions of fairness), and their *inter-relations* (e.g., normative priorities, commensurability, conceptual compatibility). Current literature in this domain deals primarily with what Rawls' calls *imperfect procedural justice* [19] though it also frequently draws from philosophical theories of non-discrimination, equality, distributive justice ([14, 39, 40]). Indeed, many philosophers conceptualize fairness criteria roughly as (approximations of) normative requirements that regulate imperfect decision procedures in general (e.g., [38, 39]). Here, predictive models are notably viewed only as *instantiations* of imperfect procedures; that is, technological systems are of interest to theory in this domain because they inform or execute prediction-based decisions. I emphasize this because, while a theory of algorithmic fairness might discuss predictive models and software to make its point, as it were, it need not assume that fairness in a broad sense *is* purely a property of the technical system itself [8, 10].

On the other hand, I understand *fair ML* as circumscribing a subdomain of the theory of institutional design which deals specifically with the (socio)technical implementation of algorithmic fairness (as specified by some theory of algorithmic fairness *qua* theory of ideals). Work in this subdomain does not seek to explain or elucidate algorithmic fairness. It is instead concerned with the practical implementation of algorithmic fairness, purporting to guide and empirically test how it could be implemented or realized in technical or sociotechnical systems. Literature in this area explores a wide range of questions, including but not limited to how algorithmic fairness should be operationalized in practical contexts [3], how debiasing interventions affect model predictions and performance and what kinds of trade-offs arise as a result (e.g., [15, 16, 29]), how human users actually use predictive models [34], and how developers use practical resources for fairness-sensitive design

[37]. I argue that it is this subdomain of theory where we find a similar multi-dimensional continuum as described by Hamlin and Stemplowska [11].

5.2. The ideal–nonideal continuum of fair machine learning

Fair ML as defined above forms a continuum spanning from more to less ideal approaches. Standard approaches to fairness in ML often enforce some notion of algorithmic fairness in an available predictive model. Most such approaches are arguably located at the *more ideal* end of the described continuum. The guiding idea is that, "[i]deally, the sensitive attribute will have a causal connection to neither the model variables, the classification, nor the target property" [9, p. 287] and an intervention should be performed to ensure that the model's predictions align with, or at least approximate, this ideal pattern. This also highlights that standard approaches tend to focus primarily on the *technological* implementation of values. Indeed, they typically employ considerable abstractions to isolate the technological system from the broader sociotechnical context and rely on idealized modeling assumptions concerning, for example, the population of interest (e.g., a lack of background injustice), the technical system and its components (e.g., accurate training data), and the system deployment setting (e.g., compliant decision-makers *qua* system users). This introduces problems, as was discussed above, but it is crucial to note that idealization and abstraction are important methodological tools which serve the highly specific aims of normative modeling in these approaches. Abstractions and idealizations provide control over contingent and external factors that characterize realistic, nonideal systems and settings (e.g., pervasive inequalities, data inaccuracies, misuse of the predictive model). By focusing on an idealized and isolated system, these approaches can produce theoretical resources (e.g., fairness metrics) and practical prescriptions (e.g., debiasing strategies) that are more likely to be consistent with a notion of algorithmic fairness that is regulative in a more ideal system or decision-making settings (cf. principles of justice that characterize a just society [19]).

The approaches described in Sections 3 and 4 are located in the *less ideal* end of the continuum. They recognize that the theoretical resources and prescriptions produced by more ideal approaches are not applicable (at least directly) in more realistic, nonideal (socio)technical systems. They seek to provide guidance in settings with realistic constraints and complex dynamics, and where decision-makers encounter a wide array of moral conflicts and tradeoffs as a result. However, each of the described approaches departs from the standard, more ideal approaches in a distinct way. Broadly speaking, negativist, comparativist, and transitional approaches (Section 3) all start from the assumption that, in a nonideal decision-making setting, a realistic model will exhibit some *prima facie* undesirable inequalities (and therefore be less-than-fully just). But they also represent different approaches to identifying a feasible "second-best" option (e.g., a predictive model, decision procedure, or a distributive pattern that should be preferred when the ideal notion of algorithmic fairness cannot be realized). Negativist approaches prioritize the elimination of salient instances of *un*fairness, for instance, whereas transitional approaches seek permissible means to improve fairness relative to a long-term target. The three fact-sensitive approaches, in turn, are unified in their focus on incorporating more realistic empirical assumptions and constraints into their respective fairness evaluations and prescriptions for intervention strategies. Still, each of them omits *specific kinds* of idealizations and abstractions based on *specific reasons*. Some approaches explore fairness from a *socio*technical perspective to produce evaluations and prescriptions that are desirable and feasible even when social factors and complex causal dynamics of the real-life decision-making setting are factored in *qua* constraints on value implementation. Other approaches omit idealizations instead to identify and address moral and practical conflicts which simply do not arise in idealized (socio)technical systems. Upstream fact-sensitive approaches theorize what is fair or just in a nonideal world shaped by past injustice (e.g., [4, 6]), for instance, whereas partial compliance approaches theorize what constitutes a fair or just decision procedure in a setting where group-level outcomes are affected by the unjust or unfair actions of some decision-makers (e.g., [36]).

'Ideal' and 'nonideal' approaches to fair ML thus form a complex continuum, where a particular approach operates at some fixed level of abstraction (e.g., technical–sociotechnical) and relies on some theory of ideals. That theory may be *positive* (e.g., a theory of algorithmic fairness) or *negative* (e.g., a theory of algorithmic *un*fairness). It may have a broader or more restricted *scope of application* (e.g.,

it may apply universally or only to a particular domain) and it may be more or less *complete* (e.g., it may include every relevant principle characterizing the ideal or only a subset of them) [27, p. 344]. But a particular approach is also likely to be more ideal in one sense and less ideal in another, and combinations of different approaches are possible (e.g., transitional fairness under partial compliance). I emphasize that I do not assume that critics of standard approaches to fair ML subscribe to a binary view of the non/ideal theory demarcation. Rather, I suspect that some critiques of mainstream methods in the field might be grounded in different conceptions about the theoretical and practical *aims* of theorizing in this area. For example, Fazelpour and Lipton state in the beginning of their insightful critique that "[t]he purpose of developing these [fair ML] tools is to ensure that ML-based decision systems yield allocations that are just in a world that is plagued by systematic injustices" [4, p. 57]. But this does not capture the diversity of theoretical and practical aims that drive work in the field. For one, research on algorithmic fairness (*qua* theory of ideals) may simply seek to explain what constitutes fairness in prediction without any immediate interest in matters of practical implementation. But even studies that *do* seek to provide practical guidance in nonideal circumstances can diverge in their more specific aims, research questions, methodologies, and abstractions and idealizations. Furthermore, the choice between ideal and nonideal perspectives often hinges on the relative weight assigned to *desirability* (or "perfection") versus *feasibility* (or "realism") and this choice comes with tradeoffs [11, 12]. If we approach fairness from an ideal-theoretical perspective, the produced theoretical resources, tools, and practical prescriptions may be consistent with ambitious ideals that are worth pursuing. But they offer less guidance for present, nonideal circumstances where such ideals are beyond reach. Nonideal approaches are better positioned to offer such guidance, including by being sensitive to past and present patterns of inequality and injustice. Still, actionable prescriptions can come at the cost of losing sight of the ideal. There is no guarantee that all compromises and "second-best" solutions improve justice in the long term, or that such compromises are even consistent with ideal-theoretical models of justice (see [13, 27, 41]).

In light of these complex tradeoffs, I argue that "more ideal" methodologies for implementing fairness in (socio)technical systems persist in their theoretical and practical relevance, but they should be supplemented with different kinds of "less ideal" approaches which address their shortcomings and limitations. Indeed, I maintain that the application of different perspectives and methodological frameworks can be highly beneficial in practice. This would involve exploring different sets of empirical modeling assumptions (e.g., levels of fact-sensitivity) and other factors, such as intervention points (e.g., data, learning algorithm, user) and timeframes (e.g., short- versus long-term improvement). By moving between (different kinds of) ideal and nonideal perspectives, one can produce a more detailed picture of available fairness-enhancing interventions and the effects they produce in different kinds of potential scenarios. For example, some interventions might mitigate some particular unfair disparity quite robustly in numerous counterfactual scenarios, while others might remove that disparity entirely but only in highly restricted cases [5, p. 13].

6. Concluding remarks

This paper continued a discussion initiated by Sina Fazelpour and Zachary Lipton' insightful critique of ideal modes of theorizing in fair ML [4], further exploring the connection between debates concerning fair ML methodology and debates concerning non/ideal theory in political philosophy. The primary contribution of the paper is an outline of six "nonideal" approaches to fair ML which can be applied to evaluate and implement fairness in (socio)technical systems in nonideal circumstances, where small-scale ideals of fairness in prediction "do not compute" due to feasibility constraints, systemic noncompliance, or the presence of background injustice and pervasive inequalities. Drawing on Hamlin and Stemplowska's [11] account, I proposed a taxonomy of theory that elucidates the relationship between ideal and nonideal approaches to fair ML. I argued, first, that they are distinct from the theory of algorithmic fairness and, second, that they are not fundamentally distinct but rather form a continuum. Different approaches can also complement one another in practice. It can be beneficial to assess available fairness-enhancing interventions (technical and sociotechnical) under different kinds of modeling assumptions and constraints, and in different counterfactual scenarios. The depicted landscape of methodologies for evaluating and implementing fairness in (socio)technical ML systems presents fruitful avenues for future research, including

research where different approaches are combined. The proposed account also highlights the need to continue theorizing algorithmic fairness (and other values and ideals relating to prediction-based decision-making) – no approach can operate without *some* theory of ideals. Currently, the field's operative ideals are, alas, often underspecified and/or underinclusive[6]. Still, such problems are not addressed by doing more nonideal theory and less ideal theory – rather, they are addressed by doing *more* and *better* theory.

References

[1] S. Fazelpour, D. Danks, Algorithmic bias: Senses, sources, solutions, Philosophy Compass, 16(8) (2021). https://doi.org/10.1111/phc3.12760.

[2] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Computing Surveys (CSUR), 54(6) (2021) 1–35. https://doi.org/10.1145/3457607.

[3] S. Mitchell, E. Potash, S. Barocas, A. D'Amour, A., K. Lum, Algorithmic fairness: Choices, assumptions, and definitions, Annual Review of Statistics and Its Application, 8 (2021) 141-163. https://doi.org/10.1146/annurev-statistics-042720-125902.

[4] 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.

[5] S. Fazelpour, Z. C. Lipton, D. Danks, Algorithmic fairness and the situated dynamics of justice, Canadian Journal of Philosophy, 52(1) (2022) 44-60. https://doi.org/10.1017/can.2021.24.

[6] J. L. Davis, A. Williams, M. W. Yang, Algorithmic reparation, Big Data & Society, 8(2) (2021). https://doi.org/10.1177/20539517211044808.

[7] B., Green, L. Hu, The myth in the methodology: Towards a recontextualization of fairness in machine learning, in: Proceedings of Machine Learning: The Debates Workshop at the 35th International Conference on Machine Learning (ICML'18), 2018.

[8] B. Green, S. Viljoen, Algorithmic realism: expanding the boundaries of algorithmic thought, in: Proceedings of the 2020 conference on fairness, accountability, and transparency (FAT* '20), 2020, pp. 19–31. https://doi.org/10.1145/3351095.3372840.

[9] J. Herington, Measuring Fairness in an Unfair World, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, Association for Computing Machinery, New York, NY, USA, 2020, pp. 286–292. https://doi.org/10.1145/3375627.3375854.

[10] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, J. Vertesi, Fairness and abstraction in sociotechnical systems, in: Proceedings of the conference on fairness, accountability, and transparency (FAT* '19), January 29–31, Atlanta, GA, USA, 2019, pp. 59–68. https://doi.org/10.1145/3287560.3287598.

[11] A. Hamlin, Z. Stemplowska, Theory, ideal theory and the theory of ideals, Political Studies Review, 10(1) (2012) 48–62. https://doi.org/10.1111/j.1478-9302.2011.00244.x.

[12] L. Valentini, Ideal vs. non-ideal theory: A conceptual map, Philosophy compass, 7(9) (2012) 654–664. https://doi.org/10.1111/j.1747-9991.2012.00500.x.

[13] A. J. Simmons, Ideal and nonideal theory, Philosophy & Public Affairs, 38(1) (2010) 5–36. https://www.jstor.org/stable/40468446.

[14] R. Binns, Fairness in machine learning: Lessons from political philosophy, 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. 149–159. URL: http://proceedings.mlr.press/v81/binns18a.html.

[15] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel, Fairness through awareness, in: Proceedings of the 3rd innovations in theoretical computer science conference, ITCS '12,

[6] Fairness notions are often rather *vague* in their content (e.g., formally defined) and thus lend themselves to diverging interpretations and misapplication (e.g., "fairness-washing") [4]. They also tend to capture certain justice-related concerns (e.g., non-discrimination) but neglect others (e.g., structural injustice, non-comparative fairness) [4 ,6, 10]. These are genuine problems, but notably unrelated to the applied *mode* of theorizing (cf. [4, 6]). Underspecification and under-inclusion are general problems that can arise even when dealing with nonideal theory (e.g., a theory of justice in conditions of partial compliance can conceivably be highly abstract and fail to include important facets of justice).

2012, pp. 214–226. Association for Computing Machinery. https://doi.org/10.1145/2090236.2090255.
[16] A. Chouldechova, Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, Big data, 5(2) (2017) 153–163. https://doi.org/10.1089/big.2016.0047.
[17] M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning, in: Advances in Neural Information Processing Systems, vol. 29, 2016, pp. 3315–3323. http://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.pdf.
[18] M. J. Kusner, J. Loftus, C. Russell, R. Silva, Counterfactual Fairness, in: Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https: //proceedings.neurips.cc/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract. html.
[19] J. Rawls, A Theory of Justice, 1st ed. (reprint), Harvard University Press, 2005/1971.
[20] A, Sen, The Idea of Justice, Penguin Books, 2009.
[21] J. T. Levy, There is no such thing as ideal theory, Social Philosophy and Policy, 33(1-2) (2016) 312-333. https://doi.org/10.1017/S026505251600025X.
[22] C. W. Mills, "Ideal theory" as ideology, Hypatia, 20(3) (2005) 165–183. https://doi.org/10.1111/j.1527-2001.2005.tb00493.x.
[23] C. Farrelly, (2007), Justice in ideal theory: A refutation, Political studies, 55(4) (2007) 844-864. https://doi.org/10.1111/j.1467-9248.2007.00656.x.
[24] D. Wiens, Against ideal guidance, The Journal of Politics, 77(2) (2015) 433–446. https://doi.org/10.1086/679495.
[25] D. Wiens, Assessing ideal theories: Lessons from the theory of second best, Politics, Philosophy & Economics, 15(2) (2016) 132–149. https://doi.org/10.1177/1470594X15620343.
[26] E. Anderson, The imperative of integration, Princeton University Press, 2010.
[27] I. Robeyns, Ideal theory in theory and practice, Social Theory and Practice, 34(3) (2008) 341–362. https://doi.org/10.5840/soctheorpract200834321.
[28] M. S. A. Lee, L. Floridi, Algorithmic fairness in mortgage lending: from absolute conditions to relational trade-offs, Minds and Machines, 31(1) (2021) 165–191. https://doi.org/10.1007/s11023-020-09529-4.
[29] J. Kleinberg, S. Mullainathan, M. Raghavan, Inherent trade-offs in the fair determination of risk scores, in: C. Papadimitriou (ed.) 8th Innovations in Theoretical Computer Science Conference (ITCS 2017), Leibniz International Proceedings in Informatics (LIPIcs), vol. 67, 2017, pp. 43:1–43:23.
[30] B. Hutchinson, M. Mitchell, 50 years of test (un) fairness: Lessons for machine learning, in: Proceedings of the conference on fairness, accountability, and transparency (FAT* '19), January 29–31, 2019, Atlanta, GA, USA, 2019, pp. 49–58. https://doi.org/10.1145/3287560.3287600.
[31] P. Gilabert, Comparative assessments of justice, political feasibility, and ideal theory, Ethical Theory and Moral Practice, 15(1) (2012) 39–56. https://doi.org/10.1007/s10677-011-9279-6.
[32] L. T. Liu, S. Dean, E. Rolf, M. Simchowitz, M. Hardt, Delayed impact of fair machine learning, in: International Conference on Machine Learning, PMLR, 2018, pp. 3150–3158. https://proceedings.mlr.press/v80/liu18c.html.
[33] A. Ghosh, A. Shanbhag, C. Wilson, Faircanary: Rapid continuous explainable fairness, in: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, Oxford, United Kingdom, 2022, pp. 307–316. https://doi.org/10.1145/3514094.3534157.
[34] B. Green, Y. Chen, Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments, in: Proceedings of the conference on fairness, accountability, and transparency (FAT* '19), January 29–31, Atlanta, GA, USA, 2019, pp. 90–99. https://doi.org/10.1145/3287560.3287563.
[35] R. North, 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.
[36] J. Dai, S. Fazelpour, Z. C. Lipton, Fair machine learning under partial compliance, in: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, May 19–21, Virtual Event, USA, 2021, pp. 55-65. https://doi.org/10.1145/3461702.3462521.
[37] 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.
[38] B. Hedden, On statistical criteria of algorithmic fairness, Philosophy & Public Affairs, 49 (2021) 209–231. https://doi.org/10.1111/papa.12189.
[39] M. Loi, A. Herlitz, H. Heidari, Fair equality of chances for prediction-based decisions, Economics & Philosophy, (2021) 1–24. doi:10.1017/S0266267123000342.
[40] F. Beigang, On the advantages of distinguishing between predictive and allocative fairness in algorithmic decision-making, Minds and Machines, 32 (2022) 655–682. https://doi.org/10.1007/s11023-022-09615-9.
[41] R. S. Taylor. Rawlsian affirmative action, Ethics, 119(3) (2009) 476–506. https://doi.org/10.1086/598170.