Canadian Journal of Philosophy (2022), 52: 1, 44–60
doi:10.1017/can.2021.24

ARTICLE

Algorithmic Fairness and the Situated Dynamics of Justice

Sina Fazelpour[1,2]*, Zachary C. Lipton[3] and David Danks[4,5]

[1]Department of Philosophy and Religion, Northeastern University, Boston, MA, USA, [2]The Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA, [3]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA, USA, [4]The Halıcıoğlu Data Science Institute, University of California, San Diego, CA, USA and [5]Department of Philosophy, University of California, San Diego, CA, USA
*Corresponding author. Email: s.fazel-pour@northeastern.edu

Abstract
Machine learning algorithms are increasingly used to shape high-stake allocations, sparking research efforts to orient algorithm design towards ideals of justice and fairness. In this research on algorithmic fairness, normative theorizing has primarily focused on identification of "ideally fair" target states. In this paper, we argue that this preoccupation with target states in abstraction from the situated dynamics of deployment is misguided. We propose a framework that takes dynamic trajectories as direct objects of moral appraisal, highlighting three respects in which such trajectories can be subject to evaluation in relation to their (i) temporal dynamics, (ii) robustness, and (iii) representation.

Keywords: Algorithmic fairness; algorithmic bias; politics of artificial intelligence; distributive justice; fairness

1. Introduction

The adoption of machine learning (ML) algorithms for both automating and informing consequential decisions has emerged as a prominent concern, sparking public debate and a vibrant field of research investigating justice and fairness in algorithmic decision making (Angwin et al. 2016; Barocas and Selbst 2016; Danks and London 2017). In this growing body of research on fair ML, many works seek to orient algorithmic design towards the values of justice and fairness. In particular, value specification proceeds by constructing "fairness metrics" intended to quantify statistical disparities concerning the predictions or predictive performance of algorithms between protected groups[1] that we might think ought not to exist were we to find ourselves in an "ideally fair" target state. Value implementation, in turn, involves technical modifications in algorithmic design that are purported to be "fairness-enhancing" or "justice-seeking" insofar as they can alter performance in order to satisfy the chosen fairness metrics as assessed on a given dataset. In section 2, we argue that the standard methodology of fair ML research is a particular instance of a broader methodology in ML practice that explicitly considers values only in relation to evaluation metrics used to assess and optimize predictive performance with respect to a static dataset. These fairness metrics thus support the values of stakeholders in the same ways as traditional evaluation metrics (e.g., simple accuracy). In this framing, the locus of moral appraisal[2] remains restricted to

[1]A term borrowed from the Civil Rights Act of 1964 in the United States, which species several protected attributes, including sex, race, and ethnicity.

[2]We follow Hansson (2013) in using the term "moral appraisal" to cover both moral evaluation and moral prescription (including interventions undertaken to satisfy these prescriptions).

the properties of predictive models' outputs with respect to static snapshots of historical data at hand, in isolation from social and organizational context.

When deployed in complex social systems, however, the situated behavior of allocation policies is also shaped by the interdependencies and shifting dynamics of the social contexts in which the policies are embedded, including the real impacts of decisions on their subjects, the incentives that they communicate to future subjects, and the feedback loops that emerge as decisions and input information become coupled via their interaction with a shared environment. As a result, moral appraisals that focus narrowly on the properties of allocation outputs—algorithm-based or otherwise—in isolation from these broader dynamics can result in distorted evaluations, and so undermine our prescriptive capabilities. In section 3, building on a nascent technical literature that has begun to establish a longer-term perspective through theoretical economic models, simulation studies, and empirical analyses, we argue that, faced with deployment dynamics, the standard fair ML methodology cannot provide practical guidance and that its naive application, even when intended to remediate injustices, can just as easily cause them.

The challenges raised by deployment dynamics are pervasive. They undermine the epistemic and ethical underpinnings not only of the standard fair ML strategy, but also, and more generally, of outcome-based moral appraisals that abstract away from these dynamics or else seek to absorb them into static target states. Yet, any adequate framework for evaluating and regulating decision procedures must contend with the situated dynamics of justice-seeking trajectories—the ways that interventions on technical systems influence the evolution of the particular social systems in which they are embedded. We thus consider two possible reorientations. In section 4, we critically evaluate a proposal due to Elster (2013; see also 1992). According to this proposal, our epistemic limits in reliably anticipating and, so, proactively planning for the consequences of social dynamics should motivate us to scale back our ambitions. Instead of anchoring our evaluation and design efforts to the achievement of some prespecified target state for decision outcomes, we ought to focus on the decision procedures as direct objects of moral appraisal.

In section 5, we argue for an alternative framing that takes dynamic trajectories themselves as direct objects of moral appraisal. We draw out three critical ways in which dynamic trajectories can be subject to moral appraisal with respect to: (i) the temporal dynamics of trajectories, (ii) robustness of trajectories, and (iii) the representation of trajectories. Our analysis of these considerations suggests that trajectories ought to be regarded as direct objects of moral appraisal, not secondary considerations to be absorbed in assessments of static states or contingent products of a procedure.

Before we proceed, a clarification is in order. In the next sections, we discuss a predominantly technical approach to addressing societal concerns about algorithmic bias—namely, in terms of developing and enforcing fairness measures. This does not mean, however, that there are no other proposals about how we ought to respond to these worries (Fazelpour and Danks 2021; Jobin, Ienca, and Vayena 2019). For example, some have suggested organizational changes such as diversifying the technical workforce. Others have argued for the need for more fundamental changes to our social practices. Indeed, in certain tasks, we might decide against the deployment of machine learning altogether. However, as will become clear, the challenges that we identify are not unique to standard fair ML methods. Rather, fair ML simply provides a salient case study for thinking about the challenges that must be addressed when evaluating and justifying proposed allocation procedures in complex, interactive social systems. These challenges remain whether or not the procedure is algorithm-based.

2. Fairness by design: Standard methodological perspective

In this section, we first situate the standard methodology of fair ML by showing how explicit discussions of values in the design and development of ML algorithms tend to be framed in terms of questions about the choice of evaluation metrics. In doing so, we clarify both the considerations that are abstracted away from and the type of questions that might be explored in this framing. We then

https://doi.org/10.1017/can.2021.24 Published online by Cambridge University Press

discuss the extension of this standard methodology to concerns about justice, fairness, and nondiscrimination.[3]

2.a Machine learning in allocation-decision pipelines

Many allocative decisions are guided by beliefs about the likelihood of some event of interest. For example, decisions to grant loans are typically guided by beliefs concerning the likelihood of repayment. Similarly, hiring decisions reflect beliefs about a job candidate's likelihood to be successful (by some measure). The promise of ML tools in these domains owes to their ability to mine associations from large collections of historical data to produce models that can reliably estimate the likelihood of some event (the target) given the available context (the features). These systems are typically developed using supervised learning. Suppose each individual i is characterized by some feature vector x_i and target value y_i. Given a dataset $\{(x_1,y_1),\ldots,(x_n,y_n)\}$ sampled from a distribution characterized by probability density function $p(x,y)$, the goal is to produce a model $f(x)$ that can successfully (by some measure) predict the target value for previously unseen examples drawn independently from $p(x,y)$. Crucially, the framework's validity, in a statistical sense, rests on the assumption that the underlying distribution is truly fixed.[4]

Consider a university developing an ML algorithm to predict first-year student attrition. A natural choice for the target y might be a binary value 0 or 1 indicating whether the student returned to enroll for their second year. Corresponding input features x might include the student's major(s), permanent address types (in or out of state), GPA for each semester, high school GPA, SAT score(s), financial aid status (student loan, scholarship, or neither), etc. Indeed, university administrations already use such predictive models to inform a variety of decisions to guide the allocation of scarce resources such as admissions, academic support, counseling, and financial aid (Delen 2011; Herr and Burt 2005).

However, decisions about how (if at all) to use these predictions depend on reasoning and facts that are external to the statistical procedures and particular features that are used in the algorithm. Critically, appropriate and responsible use depends on: (i) the decision of interest being translatable (in some sense) into a statistical prediction problem; (ii) precisely which features to collect, which target to predict, and how to measure them based on available data sources; (iii) how these predictions are operationalized to drive actual decisions; and, as we will discuss in the coming sections, (iv) the dynamics of deployments (see Mitchell et al. 2021; Fazelpour and Danks 2021).

In practice, developers of prediction systems tend to focus on the narrow set of value judgments that can be expressed as evaluation metrics, abstracting away from these challenges of formulation, measurement, and deployment (Selbst et al., 2019). In this narrower perspective, the data are taken for granted, the deployment context ignored, and the central choice is simply which statistic (among those computable from the observed data) to select as the standard of evaluation and objective for optimization. This choice often depends on the aims and values of decision-makers and is typically formalized in terms of quantitative evaluation metrics that are defined in terms of various properties of the distribution of observed data.

In the simplest case, for example, we might measure a model's classification performance by its overall predictive accuracy in terms of the proportion of examples that the model correctly

[3]Prima facie, research in fair ML might be seen as an instance of a larger body of work on value-oriented or value-sensitive design (Aizenberg and van den Hoven 2020; Friedman and Hendry 2019). In this literature, value specification and implementation tends to proceed by means of a gradual, empirically informed, iterative translation of an abstract value into more context-sensitive norms that can promote the value throughout an information processing pipeline, and ultimately become concrete sociotechnical design requirements that could support those norms in deployment. In contrast, as we will show, many works in fair ML tend to adopt a narrower focus analogous to the choice of appropriate evaluation metrics in an engineering context, rather than value-sensitive design.

[4]Validity, more broadly construed, also depends on various other value judgements and assumptions discussed below, e.g., that the observed target variable is actually the outcome we care about, or that we agree on the correct measure of success.

https://doi.org/10.1017/can.2021.24 Published online by Cambridge University Press

classifies. In other applications, we might care not only about overall predictive accuracy, but also (or indeed more so) about how predictive errors are distributed across different types of examples. Consider again the case of predicting student attrition. If only 10 percent of students in a representative dataset drop out after the first year, then a model that simply predicts that everyone will return for their second year will achieve 90 percent accuracy, but provide no useful information. Clearly, this model fails to serve the interests and values of different stakeholders (e.g., universities, families, and policymakers) (Thammasiri et al. 2014). Universities, for instance, often care about the cost of attrition to their academic mission, reputation, and finances. False positive errors—instances where the model misclassifies a student as returning for enrollment when in fact they will drop out—matter differently than false negative errors—instances where the model erroneously predicts drop out but the student, in fact, returns. Supporting the aims and values of decision makers requires the use of other, more sophisticated evaluation metrics that offer insights into error distributions (e.g., precision, recall, F-score).

Given an appropriate evaluation metric, developers can use a variety of interventions in the process of algorithmic design to optimize model performance for that metric. Such interventions might target the data preprocessing stage (e.g., using sampling methods that artificially balance cases of attrition and nonattrition); the learning stage (e.g., introductions of differential costs directly into learning); or the postprocessing stage (e.g., modification of decision thresholds) (Guo et al. 2008). In general, optimizing for different metrics will produce different solutions, and the choice to optimize a particular metric will reduce performance according to other metrics. By itself, the formalism of ML cannot tell us which evaluation metric (if any) is appropriate, or how exactly to settle tradeoffs among competing desiderata. Rather, it can only inform us about the existence of tradeoffs and potential ways of balancing them. Developers must make value choices about which evaluation metrics are most ethically, socially, and politically defensible.

So far, we have refrained from discussing specific issues of justice and fairness in allocation. This is intentional, as we seek to draw attention to the general form of value-sensitive appraisal that is typical in the evaluation and design of predictive models. This general form is surprisingly constrained: (i) the evaluations and interventions tend to focus narrowly on the predictive model (and its immediate input and output); (ii) the main focus is on quantitative evaluation metrics that depend only on the model and the statistical properties of the available data; and (iii) the resulting assessments rely on the assumption that the data distribution is static (i.e., what we have seen in the past is statistically representative of what we will see in future deployment). In these ways, values are explicitly considered in only a narrow slice of the design, development, and deployment process.

2.b The role of fairness metrics in fairness by design

While mainstream work in fair ML appears to address a categorically different set of concerns (typically, equity among subpopulations of interest), it nevertheless adopts the same methodological framework as more conventional ML work. Specifically, they restrict the universe of possible objectives to statistics that can be estimated on the test set for a given dataset. Just as with conventional supervised learning, the job of the fair ML practitioner is to decide which metric to designate as the evaluative standard and what methods to apply to optimize that metric. Of course, the fair ML practitioner focuses on different metrics that are sensitive to disparities across subpopulations, but the basic methodology remains the same.

Consider, for example, the much-publicized discussion of bias in risk assessment models that inform judges in making bail and sentencing decisions across many counties and states (Stevenson and Doleac 2019). These models are designed to predict the risk of recidivism—operationalized as likelihood of rearrest within a fixed period after release—based on defendants' features such as age, sex, and number and severity of prior offenses. Although the set of features often excludes race, Angwin et al. (2016) demonstrated that some widely used risk assessment models exhibit significant disparities in the distribution of predictive errors across different racial groups. Specifically, among

https://doi.org/10.1017/can.2021.24 Published online by Cambridge University Press

the set of defendants that were in fact not rearrested upon release, black defendants were found to be almost twice as likely as white defendants to be misclassified as "high risk." Alternately, in cases where defendants were, in fact, rearrested, white defendants were found to be almost twice as likely to be misclassified as "low risk." Importantly, such disparities will not be detected if models are evaluated exclusively via aggregate metrics that do not quantify disparities across groups. The use of different, more sophisticated evaluation metrics can help in these cases.

The initial wave of research in fair ML thus focused primarily on these novel fairness metrics—additional formal evaluation criteria intended to capture the extent to which a given model satisfies certain desiderata concerning justice, fairness, and antidiscrimination (e.g., Zafar et al. 2017; Hardt, Price, and Srebro 2016; Grgić-Hlača et al. 2018). In the dominant statistical approach, these metrics are functions of the distribution of predictions (conditional on observed features) across different groups. In practice, these metrics are typically expressed in terms of disparities in either the difference or the ratio of some quantity (e.g., the fraction predicted positive, predictive accuracy, false positive rate, false negative rate) assessed separately on two subpopulations.

Even if we detect a groupwise disparity as measured by some fairness metric, the proper response is far from straightforward. Disparities detected at the modeling stage are often too coarse to tell us definitively whether we actually have an ethical problem or what its source(s) might be. To be clear, disparity measures can be useful. For example, a detected disparity may prompt us to investigate its causes, but that investigation must proceed in the real world and not merely at the level of datasets, models, and statistical metrics. Problems can arise because of actual bias in the world, or from one of the many choices made prior to model development (e.g., how we collect data or operationalize key concepts), or through many other pathways (Danks and London 2017).

Crucially, fairness metrics often play a further practical role beyond this (imperfect) diagnostic function. In particular, they serve as target states for corrective interventions that seek to align model performance with the metric. Similar to techniques for optimizing model performance with respect to traditional evaluation metrics, these "fairness-enhancing" interventions can take a variety of forms, including methods that alter data preprocessing procedures (e.g., Kamiran and Calders 2012), learning objectives (e.g., Zafar et al. 2017), or decision thresholds (e.g., Hardt, Price, and Srebro 2016). The core idea behind all these interventions, however, is to produce a new model that maximizes predictive accuracy subject to the satisfaction of some chosen fairness metric.[5] In practice, the variety of targets and modes of intervention means that there are various trajectories to achieving the target state encoded in a metric on a given dataset. That is, there is often a multiplicity of "fair" models (according to some metric) that while indistinguishable with respect to that metric might differ substantially in other respects (Chouldechova and G'Sell 2017).[6]

This type of fair ML faces a number of challenges, however. Many situations have been identified in which decisions that optimize proposed fairness metrics violate common notions of justice and fairness (Dwork et al. 2012; Lipton, Chouldechova, and McAuley 2018; Fazelpour and Lipton 2020). Moreover, we cannot (in general) simultaneously eliminate all disparities as quantified by these metrics. As the much-publicized impossibility results demonstrate, tradeoffs among many metrics

[5]The role of these techniques for improving fairness measures appears to be underappreciated by philosophers. For example, Glymour and Herington (2019) claim that it is structurally impossible in certain cases to satisfy measures that require groupwise parity in error rates, regardless of the quality of our measurements. However, their method (using causal graphical models) requires coherence assumptions that are violated by almost all the aforementioned techniques for satisfying fairness metrics. As a result, the general claim by Glymour and Herington is mistaken, although their more specific claim—if we do not introduce statistical biases that violate these assumptions, then no amount of better measurements suffices to resolve the detected disparities—is correct.

[6]For example, two models might both satisfy equal distribution of error rates across two protected groups and exhibit the same overall accuracy, but nonetheless differ in terms of particular cases where these errors occur. This multiplicity of "fair" models might be seen as a particular case of general "predictive multiplicity" in algorithmic design with respect to evaluation metrics (Marx, Calmon, and Ustun 2020).

https://doi.org/10.1017/can.2021.24 Published online by Cambridge University Press

are inevitable, with perfect parity simultaneously achievable only in highly contrived circumstances (Chouldechova 2016; Barocas and Selbst 2016; Kleinberg, Mullainathan, and Raghavan 2016). Even if we determine which fairness metrics (if any) actually align with societal desiderata, significant work remains. In particular, the fair ML practitioner typically must balance the utility of the model (assessed by some conventional metric of performance) against potential fairness disparities. Most mainstream technical work on fair ML thus puts aside normative considerations in favor of the constrained optimization problems that arise from various choices of utility and fairness metrics.

In light of disagreements about the appropriate fairness metric(s), one line of philosophical work has aimed to make explicit the normative underpinnings of the ideal target states (i.e., the optima of the fairness metrics) (Leben 2020; Hellman 2019; Glymour and Herington 2019; Binns 2018). This work draws upon theories of distributive justice, as well as legal views on antidiscrimination, to clarify the normative principles favoring various target states and, thereby, provide a reasoned basis for resolving the disagreements sparked by impossibility results. If one could perhaps establish a desired target for "fairness," then that could guide choices about which fairness metric to adopt in particular cases.

Nevertheless, these more sophisticated analyses all proceed within the basic framing of fair ML methodology, focusing on the statistical properties of the predictive model as assessed on a given historical dataset. Ultimately, however, the primary concern in devising fairness-enhancing strategies is not about these past cases, but instead to assess the impact of allocations on individuals and groups who will be affected by the deployment (at scale) of these models in the future. As discussed above, given the current framing of work in ML (in general) and fair ML (in particular), assessments on historical cases transfer to these future cases only if the relevant population characteristics are assumed to remain static. As we will argue in the next section, however, the very introduction of predictive models into complex social systems can set in motion dynamics that alter these characteristics in critical ways, thus decoupling fairness-enhancing interventions (devised in a static setting) from their intended targets in the real world.

3. The challenge of dynamics

Two core shortcomings of the standard approach are that its analyses are static and local. They are static in the sense that they focus myopically on whatever snapshot of historical data constitutes the available dataset. Any analysis that relies exclusively on such data necessarily ignores both the mechanisms by which the data came to be, and those by which decisions feed back into the environment, setting incentives that influence the behavior of those subject to decisions, and thus altering the future distribution of data. The analyses tend to be local in that they focus myopically on a single decision-maker (agent, organization, or institution), assuming that the conclusions deduced from their data alone will provide sufficient guidance about what actions should be taken in social contexts involving multiple, interacting decision-making agents.

Several recent technical papers have demonstrated how this static and local focus can produce mistaken diagnoses and specious guidance concerning justice-seeking interventions. Even in simple models, conclusions can break down or even reverse after accounting for interdependencies among the relevant social actors and institutions (Milli et al. 2019; D'Amour et al. 2020). Interventions that proceed through a static and local lens can lead to societal harm in the long run, even as assessed by the very same target ideal and metrics that they were intended to optimize (Liu et al. 2018; Heidari, Nanda, and Gummadi 2019b).

Consider Liu et al. (2018), who examine a hypothetical lending scenario that takes into account not only a (static) dataset of applicants (consisting of their attributes, the lending decisions, and the eventual observed defaults), but also a model of the process by which today's lending decisions and the associated defaults might influence long-term credit scores, impacting future access to credit. They find that considering just one step in a relatively straightforward model, short-term satisfaction of fairness criteria from a static perspective can (depending on the model parameters) lead to

https://doi.org/10.1017/can.2021.24 Published online by Cambridge University Press

either improvement, stagnation, or decline for the group that the practitioner intends to protect. While the goal is to widen access to credit, whether the intervention achieves this benefit or harms the protected group depends precariously on knowledge of how the environment evolves that cannot be determined based on the static data alone (see also D'Amour et al. 2020).

In another study, Lipton, Chouldechova, and McAuley (2018) examine a category of algorithms that they dub Disparate Learning Processes (DLPs). These algorithms aim to simultaneously satisfy two of fair ML's ideal parity conditions: blindness and demographic parity. These two conditions are often motivated via the US Civil Rights Act of 1964. Blindness requires that an ML model not depend directly on the protected trait, while demographic parity requires that members of the disadvantaged group receive favorable decisions at the same frequency as members of the advantaged group. Rather than allow the model itself to access the protected attribute as an input (which would violate blindness), DLPs incorporate the protected feature into the training scheme (e.g., as a constraint when choosing model parameters). While the resulting models may indeed satisfy these two parity conditions, Lipton et al. demonstrate that DLPs can produce models that are clearly problematic if one pauses to consider the social incentives set by the learned model.

In short, the only way for the model to satisfy demographic parity in a demographic-blind fashion is to rely on proxies for the demographic information. Thus, seemingly irrelevant features that happen to be correlated with gender might become important features for driving ostensibly gender-blind admissions decisions. Because blindness is typically motivated via appeals to Disparate Treatment doctrine (in title VII of the US Civil Rights Act), which explicitly forbids intentional discrimination on the basis of proxy variables, the resulting models satisfy the technical definition of blindness while violating the legal principle that it was intended to formalize.

Particularly relevant in our case is the precise proxies relied upon. Using real computer science admissions data, where the target variable reflects historical admissions decisions, and applying DLPs, Lipton et al. find that the learned models rely heavily on proxies like subfield of study. In subfields of study (e.g., human computer interaction) that already enjoy greater gender balance, individuals are more likely (absent demographic info) to be women and, thus, are upweighted by DLPs, while individuals applying to male-dominated fields (even if they are in fact women) are more likely (absent demographic info) to be men and, thus, disfavored by DLPs. Thus, in practice, it is in precisely those subfields where women are least represented that they are hurt by the proposed intervention. Moreover, the admissions score distorts the landscape of incentives, encouraging individuals (both men and women) to misreport their preferred field of study. Importantly, if individuals respond to this incentive, then the intervention would lose its power to achieve gender parity.

Finally, the static lens can create the illusion of impossibilities (the aforementioned irreconcilability of various parity conditions) by neglecting the fact that institutional interventions are typically part of more complex sequences of choices that unfold over time. Crucially, in many cases where it appears that fairness metrics cannot be reconciled instantly, they might nonetheless be satisfied simultaneously in the long run. For example, as Hu and Chen (2018) demonstrate through a simple model of the labor market, a short-term intervention with a demographic-aware policy may bring about long-term equilibria that can be sustained via demographic-blind policies.

Social dynamics also pose a challenge to analyses adopting the standard fair ML strategy because of the myopic focus of such analyses on a single decision-maker in ways that neglect the behavior of other relevant agents in the environment. Examining issues of partial compliance within the context of fair ML, Dai, Fazelpour, and Lipton (2020) employ simulation as a tool to model a hiring setting in which candidates stream onto a job market involving multiple potential employers.[7] In this

[7]In each turn, candidates apply to one among a set of employers and they exit the market if they are hired. Candidates also exit the market after a specified number of rounds if they are never hired.

https://doi.org/10.1017/can.2021.24 Published online by Cambridge University Press

market, the distribution of scores characterizing the perceived skills of employees differs across demographics.[8] In this setting, Dai et al. (2020) explore the consequences of partial compliance by varying the fraction of employers that comply with a version of demographic parity in their hiring policies. While employers do not communicate directly, they interact via the dynamics of the job market. An employee hired by one employer is not available to other employers in subsequent rounds. These interactions are more pronounced once one models the strategic behavior among the applicants, wherein applicants can incorporate knowledge of their group membership as well as the relevant job-market statistics to choose whether to apply to a compliant (versus a noncompliant) employer.

Dai et al. (2020) find that when members of each group make strategic decisions about where to apply based on the fraction of their demographic that were hired by each category of employer, then at equilibrium, the compliant and noncompliant employers might appear to be performing equally well vis-a-vis a naive local and static application of demographic parity, even when the compliant employers are hiring far more candidates from the disadvantaged group.[9] Moreover, when viewed at the level of social hiring statistics, whether the fraction of fairness-conscious employers translates into commensurate aggregate benefits depends precariously on a number of factors. In particular, a version of demographic parity that appears effective in static analyses becomes highly fragile in circumstances of partial compliance.[10]

Importantly, as Dai et al. (2020) note, versions of demographic parity that appear more robust to dynamics of partial compliance are not without their own undesirable externalities. Specifically, the differential incentives sort disadvantaged candidates into compliant employers and the advantaged candidates to noncompliant ones, effectively segregating the workforce. Thus, tradeoffs can emerge between the value of integration (with all the attendant benefits of diversity) (Anderson 2010) on one hand, and progress towards demographic parity. Importantly, however, these considerations cannot be accessed via the local lens adopted by the standard fair ML methodology.

Together, these works highlight the fact that we should consider the longer-term, situated impacts of proposed policies on the social systems in which they are embedded. Doing so often requires knowledge of contingent features of the world in order to assess whether some proposed intervention achieves its intended aim (without thereby resulting in undesirable externalities). Two situations can have the same statistics but exhibit radically different dynamics. Thus, any fair ML ideal (e.g., parity metric) may assess these situations identically, even when opposite interventions are called for. Note that these challenges due to social dynamics are not limited to allocation decisions that incorporate ML tools. Fazelpour and Lipton (2020), for example, suggest that these complications are encountered by certain modes of ideal theorizing about justice—which they argue is instantiated in fair ML approaches—that take as their starting point the specification of ideal target states in abstraction from contextual factors of deployment. Moreover, as we will see in next section, the prevalence of complications due to dynamics is widely appreciated in many local allocation settings.

4. Taking dynamics seriously: Procedures as objects of moral appraisal

Any effort to incorporate dynamics into our ethical and policy reasoning must acknowledge the difficulty of that task. For example, Elster (1992) describes the many ways in which the dynamics of

[8]Dai, Fazelpour, and Liptons' (2020) analysis and conclusions do not depend on whether the disparities are due to historical discrimination or bias in the assessment itself.

[9]This is due to Simpson's paradox type of issue that arises for local and static evaluations that ignore confounding causes of how the observed data can change as a result of an agent's actions (see also D'Amour et al. 2020).

[10]That is, the social benefits (in terms of the aggregate proportion of disadvantaged group hired across employers) of adopting this version of demographic parity quickly drops as the number of noncompliant employers increases, such that k percent compliance across employers does not bring about k percent of social benefits that one might observe under full compliance.

https://doi.org/10.1017/can.2021.24 Published online by Cambridge University Press

deployment—via partial compliance, incentive effects, …—can impair our ability to anticipate the consequences of adopting different allocation schemes in settings such as admissions, military drafts, and layoffs. According to Elster, the challenge of "anticipating and identifying such [dynamic] effects suggest that it is … naive to believe that allocative schemes could be fine-tuned so as to take full account of them" (166). In fact, Elster (2013) goes further, arguing that the complex dynamics of deployment will introduce a novel type of indeterminacy: even if we agree on what constitutes a desirable outcome,[11] we lack the means for verifying whether (or how often) policies designed in advance will bring about that outcome in practice. This type of indeterminacy is not only due to the practical impossibility of anticipating and proactively responding to all potential environmental interactions, but also because of (potentially value-laden) disagreements and limitations that impact our other epistemic capabilities. For example, we are often uncertain about the appropriate causal model of the environment, but such models are required to estimate the likely consequences of a given policy.[12] Similarly, we often face knowledge gaps that limit our ability to evaluate the quality of various decisions.

Elster's concerns translate directly to the case of ML-based decision-making. The previous section surveyed works that showed how even *minimal* considerations of environmental dynamics can complicate fairness appraisals carried out in static settings. While these minimal considerations suffice for highlighting *that* the situated dynamics of interventions should be incorporated in moral appraisal, they are insufficient for suggesting *how* this might be done. That is, the findings do not translate into robust guidelines for proactively fine-tuning predictive models to mitigate potential complications, precisely because these works do not contend with the further challenges of disagreement over appropriate statistical and causal analyses of the environment (Chouldechova and Roth 2020).[13]

Epistemic limits to observational capacities provide further reason to question efforts that focus our ethical and policy evaluations on outcomes. How can we assess the quality of decisions or the predictions that guided them when many of the relevant outcomes are observed only after a long delay (or never at all)? For example, it may be decades before we learn whether a borrower defaults prior to full repayment. Moreover, for those applicants denied a loan altogether, we never have the opportunity to observe whether they would have defaulted in the counterfactual scenario. While lenders apply a variety of heuristics to handle this problem (which they call reject inference), some require strong, unverifiable assumptions and others proceed from no discernible principle at all.[14]

Although these concerns are particularly salient in fair ML, Elster (2013) has a much more general target. He argues that outcome-based views—i.e., those that seek to justify the adoption of a policy by reference to the "goodness" (in some sense) of its consequences—are simply untenable when we are working in the real world. As an alternative, Elster offers an "impure procedural" normative framework that takes decision procedures as the core object of moral evaluation and intervention. That is, particular policies and mechanisms are justified by the "goodness" of the decision-making procedure (rather than its consequences or outcomes). Elster defines the goodness

[11]In fair ML context, even if there were no disagreements about the choice of fairness metric.

[12]As Elster notes, this uncertainty about causal structure can open the door to post hoc justifications for any policy. In particular, "[t]o justify a policy to which one is attached on self-interested or ideological grounds, one can shop around for a causal or statistical model just as one can shop around for a principle. Once it has been found, one can reverse the sequence and present the policy as the conclusion" (2013, 5).

[13]As Chouldechova and Roth note, "the specific predictions of most models of this sort [that seek to incorporate aspects of dynamics] are brittle to the specific modeling assumptions made—they point to the need to consider long term dynamics, but do not provide robust guidance for how to navigate them" (2020, 86). Part of the issue is that most works focus on dynamics but retain a strongly outcome-based perspective.

[14]The algorithmic decision itself might be a cause of the prediction subject's later response, so we must make assumptions about those impacts. A real-world instance is the potential criminogenic impact of incarcerations, which significantly complicates efforts to assess accuracy about criminal sentencing and parole algorithms. For other examples of related "selective labels" problems, see Lakkaraju et al. (2017); Corbett-Davies and Goel (2018).

https://doi.org/10.1017/can.2021.24 Published online by Cambridge University Press

of a procedure negatively: a procedure is better to the extent that it eliminates or mitigates known obstacles to good decision-making (e.g., systematic cognitive and motivational biases, prejudices, …). It is an impure procedural view in that, even after having done our best to remove such obstacles, the results might not necessarily constitute desirable outcomes and, so, might need to be overturned. Overall, though, our epistemic humility (in light of the challenges raised by dynamics) requires that we morally appraise only the procedure and then simply "let the chips fall where they may" (12).

There is much to commend about Elster's proposal. In particular, his recognition of the insufficiency of outcome-based justifications as well as their vulnerability to interactive dynamic effects, fragile (and potentially motivated) causal analyses, and strong epistemic limits all provide strong reasons to focus on institutional decision-making processes. This focus on various stages of the decision process stands in stark contrast with the debates surrounding the choice of (fairness or evaluation) metrics for ML-based decisions, which often abstract away from other value judgments that are embedded throughout the process of design and development. As noted by some authors (Selbst et al. 2019; Fazelpour and Lipton 2020), the standard approach in fair ML has resulted in a perilous solutionism that obscures objectionable choices made throughout the process by constraining the scope of normative reasoning to debates about fairness metrics.

Nonetheless, there are reasons to worry that Elster's impure procedural framework simultaneously goes too far and not far enough. On the one hand, as noted by Elster himself, sometimes even our best efforts toward securing the quality of decision-making procedures are not good enough. Appropriate overruling mechanisms should thus be in place to safeguard against this possibility. Yet, decisions to invoke these overruling mechanisms require the anticipation and evaluation of outcomes. The impure procedural framework arguably goes too far in its focus on procedures to the exclusion of outcomes and dynamics.[15] On the other hand, it is unclear whether Elster's procedural approach truly dispenses with outcome-based considerations or only changes the ways that they are incorporated. Consider the suggestion that systematic biases resulting from cognitive strategies such as availability and representativeness heuristics are in some sense problematic, and so should be nullified by a "good procedure." Implicit in this claim is the idea that these heuristics are supposed to approximate some standard (e.g., the standards of rational choice theory) from which they systematically diverge in certain scenarios. Characterizing such divergences as problematic is, therefore, not agnostic with respect to the conceptions of good decision outcomes.[16] That is, the impure procedural framework is still in a sense outcome-centric.

5. Taking dynamics seriously: Trajectories as objects of moral appraisal

While outcome-based attempts to absorb the consequences of dynamics into static target states is epistemically untenable, the sole focus on the procedures is also insufficient.[17] Instead, we propose to take dynamic trajectories as objects of moral appraisal in themselves, and not merely as agglomerations of static states or contingent products of a procedure. That is, we contend that the moral evaluation of a trajectory cannot be simply the sum of moral evaluations at each moment in time, nor the moral evaluation of the procedure that generated it. Rather, the moral evaluation of a trajectory depends (partly) on good-making properties of the trajectory as a whole. We thus turn now to three important ways in which full trajectories are subject to moral appraisal.

[15]Thankfully, the knowledge required here can be more coarse-grained than the type of information needed for proactively fine-tuning the policy to account for such possibilities. One can agree with Elster that we ought not be completely outcome-centric while disagreeing that we ought to be completely procedure-centric.

[16]Gerlsbeck (2014) raises this type of concern about Elster's approach.

[17]Of course, the choice to focus on procedures or outcomes need not be dichotomous (Meshelski 2016). In what follows, however, we consider trajectories as an additional object of moral appraisal.

https://doi.org/10.1017/can.2021.24 Published online by Cambridge University Press

5.a Temporal dynamics of trajectories

Justice-promoting interventions are often evaluated in a relatively static way: the action occurs at some time and then the outcome results at some indeterminate future time. When we consider trajectories as a whole, however, we see that such interventions are typically part of more complex sequences of choices. Actions are rarely irrevocable or unchangeable but can, instead, be adapted as we learn more about the relevant social systems, or as those systems change over time. Any particular choice (given a state at a particular moment) must be understood in relation to the many other choices and states in the trajectory. For example, our moral evaluations of a trajectory should acknowledge injustices along the path towards the ideal state, not simply assume that they are permissible transient costs (Valentini 2012). Moreover, social systems rarely reach an end state but rather continue to evolve over time, and so the language of "reaching the ideal target state" presupposes a false finality.

When we broaden our perspective to consider trajectories as objects of moral consideration in themselves, we immediately recognize that considerations of speed, efficiency, and related tradeoffs all become relevant for ethical evaluations. For example, a sequence of actions $A_1, \ldots, A_n$ might quickly lead to a good but not-quite-ideal state, while actions $B_1, \ldots, B_m$ slowly reach an almost-ideal steady-state. If we prioritize the short-run, then the A-sequence is presumably preferable; in the (sufficiently) long-run, the B-sequence is better. Moral evaluation of a trajectory thus depends on the relevant timescales, and the tradeoffs between success on different timescales. In particular, we must consider whether the short-run harms incurred under the B-sequence (relative to the A-sequence) are ethically defensible in light of the longer-term benefits. We do not advance a context-general solution here since any resolution of these tradeoffs will need to engage with issues of interpersonal comparison and moral entitlements: How do we weigh harms to individuals now versus those done to future individuals, and what can each generation claim as a legitimate moral entitlement?

Of course, one might address these questions without considering the trajectory as a distinct object of moral evaluation (and many ethicists and political philosophers have). One might, for example, hope to simply "integrate" the ethical benefits and harms at each time point to obtain an overall evaluation of a trajectory. Such an "integration" over a trajectory is far from straightforward, however, since they are not necessarily fixed objects; we typically have the ability to adapt our interventions over time, or repeatedly intervene in different ways. Indeed, our evaluation of trajectories invariably involves key ethical-epistemic tradeoffs. Given our status as epistemically limited agents, we might have to incur some ethical "cost" in the short-run to gain epistemic insights that enable us to do better ethically in the long-run, as raised in Mill's "experiments in living" (Anderson 1991; Mill 1892; Muldoon 2016). This tradeoff is also familiar in the context of biomedical research: we run clinical trials in the short-run even though we know that some people may be harmed (e.g., by failure to adopt a beneficial treatment or by the trial itself), since the epistemic reward enables us to act more ethically in the long-run. This type of ethical-epistemic tradeoff is ubiquitous, and not a special property of clinical trials (see Gaus 2016; Morton 2012). We almost never have all the relevant background knowledge to assess the options in front of us, and so we must consider whether to take some short-run actions to reduce uncertainty (including additional information search) even though that additional time might contribute to, or allow the continued existence of, moral injustices. The ethical value of a trajectory is not merely the sum of the ethical value at each moment in time, but rather must incorporate epistemic value that enables future ethical value. Hence, there is no single timescale or time frame over which we can integrate the momentary ethical values, at least not without specifying these complex ethical-epistemic tradeoffs.

We have deliberately refrained from using examples from fair ML in this section, as our conclusions are broader than that context. We contend that the scope of moral evaluation writ large must include the trajectories themselves, including the necessarily dynamic nature and

https://doi.org/10.1017/can.2021.24 Published online by Cambridge University Press

tradeoffs of justice-promoting actions in social systems. Nonetheless, the issues raised here clearly apply directly to evaluations, interventions, and policies within fair ML. We should not simply ask whether some alternative algorithm would, in this particular moment, lead to "fairer" judgments. Rather, we must consider whether, for example, our knowledge of the relevant social systems could be improved (by well-designed interventions) in ways that could lead to fairer judgments at future times.[18]

5.b Robustness of trajectories

The previous subsection highlighted the need for navigating complex value tradeoffs when considering possible sequential interventions. Of course, the world gets a say in what happens after our interventions. Moral appraisal of full trajectories is also required because of the complexity of the social and physical systems within which our actions occur. These systems consist of highly interdependent networks of agents, institutions, and norms, all capable of functioning as distinct causal actors. As a result, we will rarely be in a position to predict or anticipate exactly how a local corrective intervention will unfold in the actual world. Assessments of its likelihood of success, as well as the distinctive ways in which it might fail, must incorporate details of this broader system.

Many areas of (applied) ethics and political philosophy aim to address this complexity by determining the probabilities of each possible outcome, and then assessing the moral value of an intervention in terms of this probability distribution, rather than only the value of the intended state. However, calculation of such probability distributions requires knowledge of the actual structure of our interventions, the relevant causal structure of the broad social system, and the likelihoods of various possible perturbations of that system. While such knowledge might be possible in very limited circumstances, these assumptions are more appropriate for toy examples rather than real-world efforts. Indeed, as we saw in section 4, Elster (1992; 2013) took this complexity as reason to focus on the decision procedure.

We suggest a different response: the moral appraisal of a trajectory should depend in part on the *robustness* of various intervention strategies for reaching (or nearing) desired states. Philosophers of science have examined the importance of robustness for purposes of prediction, understanding, and control (Weisberg 2012; Woodward 2006). In general, an action is a robust cause of some outcome when the outcome is produced in a wide range of background conditions. For example, a hammer strike is a robust cause of a glass shattering, while a fingernail tap in just the right place is not. The former cause will succeed even if the strike occurs in a different place or the crystalline structure of the glass is slightly different, while the latter is highly dependent on the exact environmental conditions. The idea of robustness also applies to policies or sequences of actions, as some will be more effective at bringing about or maintaining some desirable outcomes. In everyday life, we typically prefer robust causes on pragmatic grounds, since they enable us to achieve our goals without requiring substantial knowledge or control over our environment. We propose here that we should also prefer robust causes on ethical grounds.

Robustness evaluation requires that we consider not only the (sequence of) actions, but also (our uncertainty about) the structure and stability of the world. In particular, our moral appraisals must include the social and environmental perturbations that are epistemically foreseeable or reasonable, as those determine the conditions under which some justice-promoting intervention or policy should be robust. For example, evaluation of efforts within fair ML to change hiring practices to reduce systemic discrimination might reasonably expect only partial compliance, but not zero compliance. There is obviously no "bright line" that demarcates the reasonable from the unreasonable in this step. Rather, the exact set of potential perturbations will depend on the evaluator's

[18]This consideration is particularly important for fairness measures that require causal knowledge, as we frequently are uncertain about important parts of the relevant causal structures.

https://doi.org/10.1017/can.2021.24 Published online by Cambridge University Press

knowledge (e.g., what they should reasonably know about the other actors in the system) and their pragmatic abilities (e.g., the aspects of the environment they could control).

In many situations, moral evaluation of a trajectory will require a tradeoff between robustness and perfection. For example, a trajectory τ_1 might involve interventions that robustly lead to not-quite-ideal states, while trajectory τ_2 results in an ideal state but is generated by nonrobust interventions (i.e., most "nearby" alternative trajectories are badly suboptimal). These types of tradeoffs are widespread in everyday pragmatic decision-making. In the context of ethics and political philosophy, the usual framing in terms of decision-making under risk or uncertainty fails to capture the complexity of robustness evaluations.

We close this section by highlighting a complication that we previously elided: Should the set of reasonable perturbations include predictable-but-morally-objectionable actions by others? That is, should the moral evaluation incorporate expectations that other agents will act in morally objectionable ways? On the one hand, their actions are expected and so should arguably be included. On the other hand, if I adjust my choice in light of their bad actions, then I potentially become complicit in the perpetuation of injustice.[19] As an example in fair ML, consider a hiring evaluation algorithm presented with two applicants, where candidate C_1 is significantly better qualified than applicant C_2. But suppose also that this is a customer-facing job in a highly racist society, and only C_1 is a member of an underrepresented minority. How ought the algorithm evaluate the candidates? It cannot simply ignore our customers' racism, as that societal background condition will significantly impact not only C_1's ability to do the job, but also chances of turnover, performance appraisals that influence C_1's later work opportunities, and even the data for future iterations of the hiring algorithm. But if the algorithm ranks C_2 higher because they are more likely to succeed at the job, then we (via the algorithm) are thereby tacitly acquiescing to that racism, rather than attempting to counteract it.

As this example shows, we cannot consider an algorithmic judgment in isolation, but rather must consider the relevant timescales, sequential nature of decisions, downstream impacts on various systems, reasonably foreseeable societal changes, and much more. The challenge is not that the proper fairness metric is unknown; metrics for single judgments are deeply insufficient for this challenge. Every one of these elements is dynamic in nature and cannot be reduced to (a precise probability distribution over) a sequence of states. Instead, our moral appraisals must apply to (collections of) trajectories, including clarity about, and specification of, a variety of resulting tradeoffs.

5.c Diversity of perspectives on trajectories

The previous discussions of temporal dynamics and robustness demonstrate a strong dependence of moral evaluations on prior modeling decisions. Which aspects of the social environment should we include into our deliberations, and how? These abstractions and idealizations have both epistemic impact on the accuracy of our appraisals, and ethical impact on our value assessments of different trajectories. Our moral appraisals of trajectories will depend partly on what we choose to abstract or idealize. One might hope that these choices could be relatively "value-free." This hope seems to underpin the traditional division of labor between social scientists and regulators (who are assumed to make value-free modeling decisions) versus philosophers and policymakers (who provide value-laden appraisals, given a model). However, these representational choices are themselves value-laden and, so, we cannot simply divide the task into these two components. That is, these representational choices are yet another dimension for moral appraisal of trajectories.

[19]This connects to worries that allowing nonideal conditions to exert too much influence on our plans and policies can result in excessive timidity in our justice-seeking efforts (Stemplowska and Swift 2012).

https://doi.org/10.1017/can.2021.24 Published online by Cambridge University Press

As many philosophers of science have argued, the legitimacy of choices made in modeling contexts (at least partially) depends on the intended aims of the model (e.g., Potochnik 2017; Weisberg 2012). That is, our representational choices should be sensitive to the decision-makers' (assumed) ethical and epistemic goals and responsibilities. Hence, different aspects of the social environment will be relevant for decision-makers depending on their aims (e.g., doing no further harm versus remedying upstream historical injustices). Ethical ambitions can thus lead to representational choices with significant epistemic requirements, producing yet another ethical-epistemic tradeoff. Suppose, for example, that a local (as opposed to centralized) organization wants its hiring to reduce biases by identifying and compensating for the impact of upstream luck (e.g., family background, educational opportunities).[20] On top of the challenge of distinguishing between upstream luck and effort, the ethical ambition increases the scope of representational choices and, so, the associated epistemic demands. In particular, we must represent not only past-oriented elements needed for corrections, but also future-oriented information required for coordination with other relevant decision-makers (Elster 1992). These epistemic demands might, in turn, influence decision-makers' aims and ambitions.

More generally, as discussed in section 4, assessments of dynamics of interventions depend on potentially value-laden causal and statistical analyses. For example, value judgments influence selection of the environmental features that are perceived to be causally relevant (Icard, Kominsky, and Knobe 2017) as well as the relevant causal model itself (Statham 2020). And when the value judgments occur early in the process, they can restrict our attention in ethically and epistemically troubling ways. As a counter to this premature narrowing of possibilities, a number of researchers have advocated for inclusion of a diversity of perspectives, whether through participatory design and collaborative causal theory formation (Martin et al. 2020) or in theorizing about justice and democracy in political philosophy (Muldoon 2016; Gaus 2016; Anderson 2006).

If trajectories are objects of moral appraisal in themselves, then we have additional ways to benefit from perspectival diversity. Specifically, if we maintain a diversity of perspectives (including diverse potential aims) about the changing sociotechnical environment as well as the desired target states, then we can better understand the potential robustness of our choices, and even which choices are possible in the first place. Here, even simplified models of the environment can yield qualitative insights that enrich our understanding of how our interventions impact the broader social system. While these insights might not be fine-grained enough to guarantee the achievement of a prespecified outcome, they nonetheless offer valuable additional ways for monitoring the unfolding of justice-seeking trajectories in real time or at regular intervals.

6. Conclusion

Recent critical work on the limitations of standard approaches in fair ML highlight the many ways in which evaluation and justification of purported justice-seeking interventions cannot proceed from a static and local lens. Such evaluations need to consider the situated dynamics of such interventions and, in doing so, contend with the variety of tradeoffs that considerations of dynamics bring into view. Clearly, moral and political thinking has much to contribute to these issues. Doing so, however, first requires a recognition of how prominent philosophical analyses of justice and fairness tend to neglect critical challenges that arise because of complex dynamics and knowledge gaps. The tendency to relegate these inherently normative issues to the "applied domain"—as tasks delegated to developers or social scientists after the completion of moral

[20]For recent luck-egalitarian ideas in fair ML, see Heidari et al. (2019a) and Binns (2018).

https://doi.org/10.1017/can.2021.24 Published online by Cambridge University Press

appraisal—not only falsely restricts the domain of moral and political thinking; it might also reinforce ethical blind spots of misguided technical solutions that can result in lasting harm in practice.

Acknowledgements. We would like to thank an anonymous reviewer for the *Canadian Journal of Philosophy* for helpful comments. Alex Voorhoeve and Kate Vredenburgh provided detailed feedback on an earlier version of this article. We are grateful for their insightful suggestions. We also wish to thank the editors and the other contributors to this special issue. For most of the writing of this paper, David Danks was on the faculty at Carnegie Mellon University.

Sina Fazelpour is an assistant professor of philosophy and computer science at Northeastern University.

Zachary Lipton is an assistant professor of operations research and machine learning at Carnegie Mellon University, where he runs the Approximately Correct Machine Intelligence lab.

David Danks is a professor of data science and philosophy at the University of California, San Diego.

References

Aizenberg, Evgeni, and Jeroen van den Hoven. 2020. "Designing for Human Rights in AI." *Big Data & Society* 7 (2): 2053951720949566.

Anderson, Elizabeth. 1991. "John Stuart Mill and Experiments in Living." *Ethics* 102 (1): 4–26.

Anderson, Elizabeth. 2006. "The Epistemology of Democracy." *Episteme: A Journal of Social Epistemology* 3 (1): 8–22.

Anderson, Elizabeth. 2010. *The Imperative of Integration*. Princeton, NJ: Princeton University Press.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. "Machine Bias." 2016. *ProPublica*, May 23. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Barocas, Solon, and Andrew D. Selbst. 2016. "Big Data's Disparate Impact." *California Law Review* 104 (3): 671–732.

Binns, Reuben. 2018. "Fairness in Machine Learning: Lessons from Political Philosophy." In *Proceedings of Machine Learning Research* 81: 149–59.

Chouldechova, Alexandra. 2016. "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments." https://arxiv.org/abs/1703.00056.

Chouldechova, Alexandra, and Max G'Sell. 2017. "Fairer and More Accurate, but for Whom?" https://arxiv.org/abs/1707.00046.

Chouldechova, Alexandra, and Aaron Roth. 2020. "A Snapshot of the Frontiers of Fairness in Machine Learning." *Communications of the ACM* 63 (5): 82–89.

Corbett-Davies, Sam, and Sharad Goel. 2018. "The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning." https://arxiv.org/abs/1808.00023.

Dai, Jessica, Sina Fazelpour, and Zachary C. Lipton. 2020. "Fairness under Partial Compliance." https://arxiv.org/abs/2011.03654.

D'Amour, Alexander, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D. Sculley, and Yoni Halpern. 2020. "Fairness Is Not Static: Deeper Understanding of Long Term Fairness via Simulation Studies." In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*: 525–34.

Danks, David, and Alex John London. 2017. "Algorithmic Bias in Autonomous Systems." In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*: 4691–97.

Delen, Dursun. 2011. "Predicting Student Attrition with Data Mining Methods." *Journal of College Student Retention: Research, Theory & Practice* 13 (1): 17–35.

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. "Fairness through Awareness." In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, ITCS* '12: 214–26.

Elster, Jon. 1992. *Local Justice: How Institutions Allocate Scarce Goods and Necessary Burdens*. New York: Russell Sage Foundation.

Elster, Jon. 2013. *Securities Against Misrule: Juries, Assemblies, Elections*. Cambridge: Cambridge University Press.

Fazelpour, Sina, and David Danks. 2021. "Algorithmic Bias: Senses, Sources, Solutions." *Philosophy Compass* 16 (8): e12760.

Fazelpour, Sina, and Zachary C Lipton. 2020. Algorithmic Fairness from a Non-Ideal Perspective. In *AAAI/ACM Conference on AI, Ethics, and Society (*AIES 2020*)*: 57–63.

Friedman, Batya, and David G. Hendry. 2019. *Value Sensitive Design: Shaping Technology with Moral Imagination*. Cambridge, MA: MIT Press.

Gaus, Gerald. 2016. *The Tyranny of the Ideal: Justice in a Diverse Society*. Princeton, NJ: Princeton University Press.

Gerlsbeck, Felix. 2014. Elster's Benthamite Project. *European Journal of Sociology* 55 (3): 441–46.

https://doi.org/10.1017/can.2021.24 Published online by Cambridge University Press

Glymour, Bruce and Jonathan Herington. 2019. "Measuring the Biases That Matter: The Ethical and Casual Foundations for Measures of Fairness in Algorithms." In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT** 19: 269–78.

Grgić-Hlača, Nina, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller. 2018. "Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning." In *Proceedings of the AAAI Conference on Artificial Intelligence* 32 (1): 51–60.

Guo, Xinjian, Yilong Yin, Cailing Dong, Gongping Yang, and Guangtong Zhou. 2008. "On the Class Imbalance Problem." In *2008 Fourth International Conference on Natural Computation*, vol. 4: 192–201.

Hansson, Sven Ove. 2013. *The Ethics of Risk: Ethical Analysis in an Uncertain World*. New York: Palgrave Macmillan.

Hardt, Moritz, Eric Price, and Nathan Srebro. 2016. "Equality of Opportunity in Supervised Learning." In *Advances in Neural Information Processing Systems* 29: 3315–23.

Heidari, Hoda, Michele Loi, Krishna P. Gummadi, and Andreas Krause. 2019a. "A Moral Framework for Understanding Fair ML through Economic Models of Equality of Opportunity." In *Proceedings of the Conference on Fairness, Accountability, and Transparency*: 181–90.

Heidari, Hoda, Vedant Nanda, and Krishna P. Gummadi. 2019b. "On the Long-Term Impact of Algorithmic Decision Policies: Effort Unfairness and Feature Segregation through Social Learning." In *International Conference on Machine Learning*: 2692–701.

Hellman, Deborah. 2019. "Measuring Algorithmic Fairness." *Virginia Public Law and Legal Theory Research Paper*, No. 2019-39.

Herr, Elizabeth, and Larry Burt. 2005. "Predicting Student Loan Default for the University of Texas at Austin." *Journal of Student Financial Aid* 35 (2): 27–49.

Hu, Lily, and Yiling Chen. 2018. "A Short-Term Intervention for Long-Term Fairness in the Labor Market." In *Proceedings of the 2018 World Wide Web Conference*: 1389–98.

Icard, Thomas F., Jonathan F. Kominsky, and Joshua Knobe. 2017. "Normality and Actual Causal Strength." *Cognition* 161: 80–93.

Jobin, Anna, Marcello Ienca, and Effy Vayena. 2019. "The Global Landscape of AI Ethics Guidelines." *Nature Machine Intelligence* 1 (9): 389–99.

Kamiran, Faisal, and Toon Calders. 2012. "Data Preprocessing Techniques for Classification without Discrimination." *Knowledge and Information Systems* 33 (1): 1–33.

Kleinberg, Jon M., Sendhil Mullainathan, and Manish Raghavan. 2016. "Inherent Trade-Offs in the Fair Determination of Risk Scores." https://arxiv.org/abs/1609.05807.

Lakkaraju, Himabindu, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. "The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables." In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*: 275–84.

Leben, Derek. 2020. Normative Principles for Evaluating Fairness in Machine Learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*: 86–92.

Lipton, Zachary C., Alexandra Chouldechova, and Julian McAuley. 2018. "Does Mitigating ML's Impact Disparity Require Treatment Disparity?" In *Advances in Neural Information Processing Systems* (NIPS): 8136–46.

Liu, Lydia T., Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hard. 2018. "Delayed Impact of Fair Machine Learning." In *International Conference on Machine Learning*: 3150–58.

Martin Jr., Donald, Vinodkumar Prabhakaran, Jill Kuhlberg, Andrew Smart, and William S. Isaac. 2020. "Extending the Machine Learning Abstraction Boundary: A Complex Systems Approach to Incorporate Societal Context." https://arxiv.org/abs/2006.09663.

Marx, Charles, Flavio Calmon, and Berk Ustun. 2020. "Predictive Multiplicity in Classification." In *International Conference on Machine Learning*: 6765–74.

Meshelski, Kristina. 2016. "Procedural Justice and Affirmative Action." *Ethical Theory and Moral Practice* 19 (2): 425–443.

Mill, John Stuart. 1892. *On Liberty*. London: Longmans, Green and Company.

Milli, Smitha, John Miller, Anca D. Dragan, and Moritz Hardt. 2019. "The Social Cost of Strategic Classification." In *Proceedings of the Conference on Fairness, Accountability, and Transparency*: 230–39.

Mitchell, Shira, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. "Algorithmic Fairness: Choices, Assumptions, and Definitions." *Annual Review of Statistics and Its Application* 8 (1): 141–63.

Morton, Adam. 2012. *Bounded Thinking: Intellectual Virtues for Limited Agents*. Oxford: Oxford University Press.

Muldoon, Ryan. 2016. *Social Contract Theory for A Diverse World: Beyond Tolerance*. New York: Taylor & Francis.

Potochnik, Angela. 2017. *Idealization and the Aims of Science*. Chicago: University of Chicago Press.

Selbst, Andrew D., Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. "Fairness and Abstraction in Sociotechnical Systems." In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT** '19: 59–68.

Statham, Georgie. 2020. "Normative Commitments, Causal Structure, and Policy Disagreement." *Synthese* 197 (5): 1983–2003.

https://doi.org/10.1017/can.2021.24 Published online by Cambridge University Press

Stemplowska, Zofia, and Adam Swift. 2012. "Ideal and Nonideal Theory." In *The Oxford Handbook of Political Philosophy*, edited by David Estlund, 373–89. Oxford: Oxford University Press.

Stevenson, Megan T, and Jennifer L. Doleac. 2019. "Algorithmic Risk Assessment in the Hands of Humans." https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3489440.

Thammasiri, Dech, Dursun Delen, Phayung Meesad, and Nihat Kasap. 2014. "A Critical Assessment of Imbalanced Class Distribution Problem: The Case of Predicting Freshmen Student Attrition." *Expert Systems with Applications* 41 (2): 321–30.

Valentini, Laura. 2012. "Ideal vs. Non-Ideal Theory: A Conceptual Map." *Philosophy Compass* 7 (9): 654–64.

Weisberg, Michael. 2012. *Simulation and Similarity: Using Models to Understand the World*. Oxford: Oxford University Press.

Woodward, James. 2006. "Some Varieties of Robustness." *Journal of Economic Methodology* 13 (2): 219–40.

Zafar, Muhammad Bilal, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. "Fairness beyond Disparate Treatment and Disparate Impact: Learning Classification without Disparate Mistreatment." In *Proceedings of the 26th International Conference on World Wide Web, WWW '*17: 1171–80.

Cite this article: Fazelpour, S. Lipton, Z. C. and Danks, D. 2022. Algorithmic Fairness and the Situated Dynamics of Justice. *Canadian Journal of Philosophy* 52: 44–60, doi:10.1017/can.2021.24