What-is and How-to for Fairness in Machine Learning: A Survey, Reflection, and Perspective

ZEYU TANG, Carnegie Mellon University, United States
JIJI ZHANG, The Chinese University of Hong Kong, Hong Kong
KUN ZHANG, Carnegie Mellon University, United States

We review and reflect on fairness notions proposed in machine learning literature and make an attempt to draw connections to arguments in moral and political philosophy, especially theories of justice. We survey dynamic fairness inquiries and further consider the long-term impact induced by current prediction and decision. We present a flowchart that encompasses implicit assumptions and expected outcomes of different fairness inquiries on the data-generating process, the predicted outcome, and the induced impact, respectively. We demonstrate the importance of matching the mission (what kind of fairness to enforce) and the means (which appropriate fairness spectrum to analyze) to fulfill the intended purpose.

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**; **Machine learning;**

Additional Key Words and Phrases: Algorithmic fairness, causality, bias mitigation, dynamic process, fair machine learning

ACM Reference format:
Zeyu Tang, Jiji Zhang, and Kun Zhang. 2023. What-is and How-to for Fairness in Machine Learning: A Survey, Reflection, and Perspective. *ACM Comput. Surv.* 55, 13s, Article 299 (July 2023), 37 pages.
https://doi.org/10.1145/3597199

1 INTRODUCTION

With the widespread utilization of machine learning models in our daily life, researchers have been thinking about the potential social consequences of the prediction/decision made by algorithms. To date, there is ample evidence that machine learning models have resulted in discrimination against certain groups of individuals under many circumstances, for instance, the discrimination in ad delivery when searching for names that can be predictive of the race of an individual [174]; the gender discrimination in job-related ads push [47]; stereotypes associated with gender in word embeddings [22]; the bias against certain ethnic groups in the assessment of recidivism

Kun Zhang also with Mohamed bin Zayed University of Artificial Intelligence, United Arab Emirates.
The work was supported in part by the NSF-Convergence Accelerator Track-D Award No. 2134901, by the National Institutes of Health (NIH) under Contract No. R01HL159805, by grants from Apple Inc., KDDI Research, Quris AI, and IBT, and by generous gifts from Amazon, Microsoft Research, and Salesforce. J.Z.'s research was supported in part by the RGC of Hong Kong (Grant No. GRF13602720).
Authors' addresses: Z. Tang and K. Zhang, Department of Philosophy, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213 USA; emails: {zeyutang, kunz1}@cmu.edu; J. Zhang, Department of Philosophy, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong; email: jijizhang@cuhk.edu.hk.

0360-0300/2023/07-ART299 $15.00
https://doi.org/10.1145/3597199

risk [7, 19]; the violation of anti-discrimination law (e.g., Title VII of the 1964 Civil Rights Act) in data mining [13].

In the effort of enforcing fairness in machine learning, various notions as well as techniques to regulate discrimination under different scenarios have been proposed in the literature. There are multiple different perspectives of fairness analysis. In terms of the type of relation between variables that is encoded in the fairness criterion, there are associative notions of fairness that are defined in terms of correlation or dependence between variables, e.g., *Demographic Parity* [51], *Equalized Odds* [74], and *Predictive Parity* [32, 48, 195]; there are also causal notions of fairness that are defined in terms of causal relation between variables, e.g., *Counterfactual Fairness* [111], *No Unresolved Discrimination* [104], and *Path-specific Counterfactual Fairness* [29, 192]. In terms of the scope of application, there are group-level fairness notions, e.g., *Equalized Odds* [74], *Fairness on Average Causal Effect* [103], *Equality of Effort* [84]; there are also individual-level fairness notions, e.g., *Individual Fairness* [51], *Counterfactual Fairness* [111], *Individual Fairness on Hindsight* [73]. In terms of the technique to eliminate or suppress discrimination, there are *pre-processing* approaches [25, 42, 51, 121, 197, 198, 210], *in-processing* approaches [12, 50, 89, 96, 124, 143, 152, 168, 187, 195, 196], and *post-processing* approaches [52, 63, 74]. In terms of the time span within which fairness is considered, other than the analysis merely with respect to a snapshot of reality, the literature also includes fairness analyses in dynamic settings [45, 75, 77, 80, 115, 176, 189, 207, 209].

There are explications on available choices to quantify discrimination and enforce fairness in recent survey papers [26, 33, 38, 120, 122, 123, 126, 132, 139, 144, 153, 184, 208] as well as an investigation into public attitudes toward different notions [157]. However, the philosophical and methodological contents of the underlying fairness considerations are often not clearly articulated. In this article, beyond the aforementioned canonical ways of categorizing fairness notions, we review and reflect on previous characterizations of algorithmic fairness in both static and dynamic settings. In particular, we consider fairness inquiries with different semantic emphases and present a corresponding flowchart to navigate through various fairness spectra. We believe disentanglement of discriminations based on the intended fairness semantics is vital toward a precise and reasonable quantification of different types of discrimination, so that we can consider the suitable fairness spectrum to better accomplish the goal. With an extensive discussion into nuances between different intrinsic goals to achieve, we provide a clear picture to make sure that there is no mismatch between the mission (precisely which type of discrimination we really hope to deal with) and the means (which spectrum of fairness we should consider).

The article proceeds as follows: in Section 2, we take a quick look into how fairness and justice are approached in philosophical discussions; in Section 3, we introduce the notation conventions and provide a brief introduction to causal reasoning; in Sections 4 and 5, we review commonly used algorithmic fairness notions and present fair machine learning studies in the dynamic setting; in Section 6, we present different spectra of algorithmic fairness inquiries; in Section 7, we clarify the role of causality in fairness analysis; in Section 8, we propose an algorithmic fairness flowchart, from which we can see a clearer picture regarding how we should approach the pursuit of different types of fairness; we summarize with concluding remarks and future works in Section 9.

2 WHAT WE TALK ABOUT WHEN WE TALK ABOUT FAIRNESS

As we have seen in Section 1, machine learning literature has proposed a deluge of algorithmic fairness definitions, each of which comes with explicit or implicit assumptions on the discrimination of interest and the corresponding mathematical formulation that captures it. Justice, as a very closely related topic under a different name than "(algorithmic) fairness," has been of significant

interest to moral and political philosophers.[1] It is therefore not surprising to see fairness notions proposed from the machine learning community echo certain justice considerations in ethical theories. Several recent works have pointed out the necessity of reflecting on such connections [1, 14, 20, 38, 46, 60, 71, 79, 98, 110].

In this section, we first present an empirical scenario as a motivating example. Then, by listing various fairness-related questions that one might be interested in asking, we lay out different aspects of justice that are formalized and considered in moral and political philosophy. Here, we do not intend to give an overview of theories of justice (see, e.g., Miller [129]). Instead, we humbly borrow the wisdom of the rich literature of theories of justice to present a big picture regarding what we are talking about when we talk about algorithmic fairness.

In particular, we examine conceptual dimensions, scopes, and overarching theoretical frameworks. We demonstrate how one can benefit from a rather rich literature of theories of justice and reflect on the current literature of algorithmic fairness. The demonstration serves as a starting point, from which one can think about intuitions, (implicit) assumptions, and expectations involved in technical treatments in a principled way. The example also serves as a preamble to our detailed discussions on spectra of algorithmic fairness inquires in Section 6, on subtleties of utilization of causality in fairness analysis in Section 7, and on achieving algorithmic fairness of different spectra in Section 8.

2.1 Music School Admission: A Motivating Example

Let us consider a music school admission example. Each year, the music school committee considers the admission of applicants to the violin performance program based on their personal information, educational background, instrumental performance, and so on. The committee also has access to the transcripts of previously admitted students together with their aforementioned information when they applied to the program.

When talking about "fairness" in this empirical scenario, based on individuals' intuitive understanding or expectation of fairness, different people may ask different questions, as shown in a non-exhaustive list below:

Question 1 (*Ideal* or *Nonideal* methodologies): When we evaluate fairness of the admission, do we need to first construct an *ideal* world where the admission is fair and to which we then compare our current reality? Or do we cope with injustices in the current world and try to move to something better, e.g., a less biased admission in the future?

Question 2 (*Corrective* or *Distributive* objectives): Are we discussing fairness of the admission for the purpose of *correcting* potentially discriminatory historical decisions, e.g., by admitting a student that was wrongfully denied previously? Or are we focusing on *distributing* admission opportunities among current applicants?

Question 3 (*Procedural* or *Substantive* emphases): Are we considering fairness in terms of how the committee produces the admission decisions, i.e., the decision-making *procedure*? Or do we only care about what the final decision outcomes look like, i.e., who are admitted to the music school this year?

Question 4 (*Comparative* or *Non-Comparative* considerations): Does the fairness consideration involve *comparisons* among individuals, e.g., to compare the decisions received by two applicants who appear to be roughly equally qualified? Or are we considering applications separately, i.e., the decision received by one applicant does not affect other applications?

[1]It has been recognized that "justice" and "fairness" are not the same thing (see, e.g., Goldman and Cropanzano [72]). Therefore, instead of using "justice" and "fairness" interchangeably, throughout this section, we follow the terminology used by the referenced work to avoid conceptual misunderstandings.

Question 5 (The scope of fairness inquiries): Is the fairness consideration limited to the relationship between the music school and applicants? Or are we concerned with a broader *scope* on which the admission decision might have an influence, e.g., the future development of students and their impact on the entire community?

In the rest of this section, we will use this example to demonstrate the connections between intuitive understandings of fairness and discussions of justice in ethical theories. We will revisit this running example in Section 6, where we provide additional inquires from a technical treatment point of view and reflect on different spectra of algorithmic fairness inquiries.

2.2 Conceptual Dimensions, Scopes, and Overarching Theories

The idea of justice remains a spotlight of attention in moral, legal, and political philosophy. As we have seen in the motivating example presented in Section 2.1, there are various fairness inquiries one might be interested in conducting, each of which reveals specific aspects of fairness or justice one would like to pursue. It is therefore desirable to look at a big picture of ways in which fairness or justice has been approached in ethical theories, so that our discussion can be principled before diving into technical treatments (which will be discussed in the later part of our article). Following Miller [129], we examine conceptual dimensions (Section 2.2.1), scopes (Section 2.2.2), and overarching theoretical frameworks (Section 2.2.3) of theories of justice.

2.2.1 Conceptual Dimensions of Justice. In this subsection, let us take a look at four essential contrasts in the conceptual apprehension of fairness or justice [129], in particular, the *ideal* and *nonideal* methodologies (e.g., Question 1), the *corrective* and *distributive* objectives (e.g., Question 2), the *procedural* and *substantive* emphases (e.g., Question 3), and the *comparative* and *noncomparative* considerations (e.g., Question 4).

Ideal and Nonideal Methodologies. There are two methodological approaches in political philosophy. The *ideal* approach advances ideal principles according to which a perfectly just (ideal) world operates. For example, the "difference principle," a principle proposed by Rawls [148, 149] that requires social and economical inequalities to be regulated so that they work to the greatest benefit of the least advantaged member of the society, counts as an *ideal* principle of justice. The *nonideal* approach, however, does not posit principles and ideals for a perfectly just society. Instead, one needs to cope with injustices in the current world and try to move to something better. For example, as proposed by Anderson [4], one can evaluate the mechanisms that cause the problem of injustice, as well as responsibilities of different agents to alter these mechanisms, to determine what ought to be done and who should be charged. Recently, Fazelpour and Lipton [60] have discussed the connection between fair machine learning and the literature on *ideal* and *nonideal* methodological approaches in political philosophy.

Corrective and Distributive Objectives. In terms of the objective of fairness inquiries, the contrast between *corrective* and *distributive* justice can date back to Aristotle (*The Nicomachean Ethics, Book V*). The *corrective* objective of justice concerns a bilateral relationship between the wrongdoer and its victim, emphasizing the remedy that restores the victim to the status before the wrongful behavior occurred. In contrast, the *distributive* objective of justice involves a multilateral relationship, and formulates justice as a principle to distribute goods of various kinds to individuals. While *corrective* justice appears more frequently in law practices, current algorithmic fairness literature largely focuses on *distributive* objectives of justice, e.g., the distribution of admission opportunities in our music school example.

Procedural and Substantive Emphases. The contrast between the *procedural* and *substantive* emphases reflects different determinants of justice, namely, the justice defined in terms of the procedure itself (e.g., the process how admission committee make the decision) and the justice defined on the substantive outcome (e.g., the final admission decisions of the music school committee). The distinction between *Disparate Impact* (with a *substantive* emphasis) and *Disparate Treatment* (with a *procedural* emphasis) has been established in law (e.g., Title VII of the 1964 Civil Rights Act) and discussed in the era of big data [13, 188]. Thanks to the development of causal analysis [81, 142, 145, 171], fairness in machine learning literature has witnessed extended and ongoing efforts on mathematically formulating and empirically regulating discriminations, both *procedural* and *substantive* ones, which we will see in more detail in Section 3.

Comparative and Non-comparative Considerations. Justice can take *comparative* and *non-comparative* forms of considerations. *Comparative* justice requires one to examine what others can claim when determining what is due to an individual, while *non-comparative* justice determines what is due to an individual merely based on his/her relevant qualities. In our music school admission example, a fairness inquiry of *comparative* consideration may examine how the admission decision received by one applicant or one demographic group, compared to those received by other applicants or demographic groups. An inquiry of *non-comparative* consideration may concern whether the decision received by an individual truly respect his/her ability to succeed in the violin program.

2.2.2 Scope of Justice. In Section 2.2.1, we have seen contrasts in conceptual apprehensions of justice. An important parallel question to ask is when, and to whom, we should apply the concepts or principles of justice.

Local and Global Views. A *local* view argues that principles of justice apply only among individuals who stand in a certain relationship to each other and that the scope is limited to those within such a relationship, e.g., relational theory of justice [148] and local justice [54, 138]. In our running example, the discussion of fairness limited within the scope of music school admission itself is a *local* view of algorithmic fairness, where the relationship only involves the music school and its applicants. However, one can consider a broader scope of fairness and ask, for example, what is the long-term impact of current admission decision on potential future developments of applicants as well as their contributions to the society.

2.2.3 Overarching Theoretical Frameworks to Discuss Justice. In this subsection, we present three theoretical frameworks in terms of which justice can be understood, namely, *Utilitarianism*, *Contractarianism*, and *Egalitarianism*. We note that we highlight these frameworks since they are the more influential ones that have been implicated one way or another in current algorithmic fairness literature. These frameworks are not intended to be mutually exclusive or exhaustive.

Utilitarian Perspective of Justice. On a high level, *Utilitarianism* aims to maximize the overall welfare, and to bring about the greatest amount of good in terms of the aggregated utilities. It has been recognized that pure *utilitarianism* is not the final answer to fairness because of several obstacles it faces [53, 148, 149]: The "currency" of justice or fairness should take the form of benefits/burdens, i.e., the means to gain happiness rather than happiness/unhappiness itself as in *Utilitarianism*; *Utilitarianism* evaluates outcome in terms of the aggregated overall utilities, instead of how utilities are distributed among individuals; the evaluation is only with respect to the consequences without any consideration about how the consequence is derived in the first place.

Contractarian Perspective of Justice. Contractarian philosophers approach justice by looking for (hypothetical) principles in forms of agreements that institutions and individuals all commit to.

David Gauthier characterizes the social contract as a bargain between rational agents and presents the principle of *Minimax Relative Concession* [67]; John Rawls presents the scenario where people know that their "conceptions of the good" are in general different, but at the same time, each individual's conception of the good is placed behind "a veil of ignorance" [148, 149]; T. M. Scanlon aims to account for "what we owe to each other" and presents the idea of justice as a general agreement where no individual, that is informed and unforced, could reasonably reject [158].

Egalitarian Perspective of Justice. On a high level, *Egalitarianism* aims to establish some sorts of equality. To a certain extent, equality could act as a default when we intuitively comprehend the idea of fairness and justice. A natural question faced by *Egalitarianism* is how to make the idea of fairness as equality more specific and reasonable in different contexts. *Responsibility-sensitive Egalitarianism* approaches this question by treating equal distribution (of opportunities) as a starting point, and allowing for departures from the equality baseline if such departures result from responsible choices of individuals [109, 125]; *Luck Egalitarianism*, as one type of *Responsibility-sensitive Egalitarianism*, adds an additional restriction that the inequalities resulting from brute luck should be constrained [8]; the debates over the role played by *luck* and *desert* also remain a major strand in *Egalitarianism* considerations [3, 37, 128].

2.3 Remark: Theories of Justice and Notions of Algorithmic Fairness

In Section 2.2, we have seen how theories of justice can shed light on various aspects one might consider when discussing "fairness" in our running example of music school admission (Section 2.1). As we shall see in Section 4 where we review a non-exhaustive list of definitions of fairness in machine learning literature, ideas of justice often echo in the intuitions behind the proposed algorithmic fairness notions.

3 TECHNICAL PRELIMINARIES

In this section, we first present the notation conventions used throughout the article in Section 3.1. Then, we present a brief introduction to causal reasoning in Section 3.2.

3.1 Notations

We use uppercase letters to refer to variables, lowercase letters to refer to specific values that variables can take, and calligraphic letters to refer to domains of value. For instance, we denote the protected feature by A, with domain of value $\mathcal{A}$, additional (observable) feature(s) by X, with domain of value $\mathcal{X}$, and ground-truth (label) variable by Y and its predictor by $\widehat{Y}$, with domain of value $\mathcal{Y}$.

Throughout the article, without loss of generality, we assume that there is only one protected feature and one ground-truth variable for the purpose of simplifying notation. Since the protected feature (e.g., race, sex, ratio of ethnic groups within community) and the ground-truth variable (e.g., recidivism, annual income) can be discrete or continuous, we do not assume discreteness of the corresponding variables.

There might be additional technical considerations for certain fairness notions to be able to apply in different practical scenarios, for instance, the phenomenon of *fairness gerrymandering* [99, 100] and the quantification of *differential fairness* [65] when considering subgroups formed by structured combinations of protected features (the theory of "intersectionality" [24, 44]), the challenge introduced by unobserved protected features during learning [27, 75] and auditing [11], and the risks and opportunities involved in the data collection of demographic information [5, 6]. However, these challenges will not impede us from discussing and reflecting on the intuitions and insights behind fairness notions.

3.2 Causal Modeling

Since we will review commonly used causal notions of fairness and discuss subtleties regarding the role played by causality in fairness analysis, we give a brief introduction to causal modeling and inference in this section.[2] Readers that are already familiar with the related topics may feel free to skip the content.

3.2.1 Definition and Representation of Causality. For two random variables X and Y, we say that X is a *direct cause* of Y if there is a change of distribution for Y when we apply different *interventions* on X while holding all other variables fixed [142, 171]. We can represent a causal model with a tuple $(U, V, \mathbf{F})$ such that:

(1) V is a set of observed variables involved in the system of interest;
(2) U is a set of exogenous variables that we cannot directly observe but contains the background information representing all other causes of V and jointly follows a distribution $P(U)$;
(3) $\mathbf{F}$ is a set of functions, also known as structural equations, $\{f_1, f_2, \ldots, f_n\}$ where each f_i corresponds to one variable $V_i \in V$ and is a mapping $U \cup V \setminus \{V_i\} \rightarrow V_i$.

The triplet $(U, V, \mathbf{F})$ is known as the **structural causal model (SCM)**. We can also capture causal relations among variables via a **directed acyclic graph (DAG)** $\mathcal{G}$, where nodes (vertices) represent variables and edges represent functional relations between variables and the corresponding direct causes, i.e., observed parents and unobserved exogenous terms.

Here, by representing causal relations via a DAG, we explicitly limit the consideration within the scope of acyclic SCMs (also known as recursive SCMs).[3] There are several reasons behind this modeling choice. To begin with, to the best of our knowledge, current algorithmic fairness literature only considers acyclic SCMs for both static and dynamic settings, and the proposed definitions are largely with respect to causal modeling based on DAGs. Therefore, when presenting the current literature, we intend to align with this default modeling choice to avoid potential misunderstandings. Furthermore, limiting the scope of discussion within acyclic SCMs involves additional technical considerations. The class of acyclic SCMs is a well-studied subclass of SCMs. Although acyclic SCMs do not have the capacity to model systems with causal cycles, e.g., equilibrium status of dynamic processes, acyclic SCMs have convenient properties including, but not limited to, the uniqueness of the induced distribution, the closedness under (perfect) interventions, the closedness under marginalizations that respect latent projection, and the obedience of various Markov conditions [59, 112, 113, 151, 171]. In general, these properties do not hold true in cyclic SCMs [23, 170].

3.2.2 Interventions and Counterfactuals. Following Pearl [142], we use the $do(\cdot)$ operator to denote an intervention, which is a manipulation of the model such that the value of a variable (or a set of variables) is set to specific values regardless of the corresponding structural equation(s), while leaving other structural equations invariant. For example, the distribution of Y under the intervention $do(X = x)$ where $X \subseteq V$, is denoted by $P(Y \mid do(X = x))$, which reads "the distribution of Y if we were to force $X = x$ in the population (regardless of the value X takes originally)."

The aforementioned intervention can also be carried out through a specific path (or a set of paths), where a path consists of nodes (variables) connected with a directed edge or a flow of

[2] Another field in the causality study is causal discovery where the primary goal is to recover the causal relations among variables from the data [31, 161, 171, 172, 200–202]. Causal discovery is not directly related to characterization of fairness in machine learning and therefore is not reviewed in this article.

[3] The causal graphs discussed in this article are limited to DAGs, and causal models represented by cyclic graphs are beyond the scope of this article.

directed edges. For example, let a path π from X to Y be a direct path $X \to Y$ (or an indirect path $X \to \cdots \to Y$), then the distribution of Y under the path-specific intervention $do(X = x|_\pi)$ along the path π, is denoted by $P(Y \mid do(X = x|_\pi))$, which reads "the distribution of Y if we were to force $X = x$ only along the path π (the value change of X is transmitted only along that path) and leave the value of X unchanged along other paths that are not π."[4]

The full knowledge about the structural equations $\mathbf{F}$ is a rather strong assumption, but it also allows us to infer counterfactual quantities. For example, let $O, X \subseteq V$ with an observation $O = o$, the counterfactual distribution of Y if X had taken value x is denoted by $P(Y_{X \leftarrow x}(U) \mid O = o)$, which reads "the distribution of Y had X been set to x given that we actually observe $O = o$." The inference of the counterfactual quantity $P(Y_{X \leftarrow x}(U) \mid O = o)$ involves a three-step procedure (as explained in more detail in Pearl [142]):

(1) **Abduction:** for a given prior on U, compute the posterior distribution of U given the observation $O = o$;
(2) **Action:** substitute the structural equation that determines the value of X with the intervention $X = x$ and get modified set of structural equations $\mathbf{F}_{\text{modify}}$;
(3) **Prediction:** compute the distribution of Y using $\mathbf{F}_{\text{modify}}$ and the posterior $P(U \mid O = o)$.

The counterfactual quantities can also be defined in a path-specific manner. For example, suppose that the intervention on X is only transmitted through the path π, then the path-specific counterfactual distribution of Y if X had taken value x only along the path π is denoted by $P(Y_{X \leftarrow x|\pi}(U) \mid O = o)$, which reads "the distribution of Y had X been set to x only along the specific path π given that we actually observed $O = o$."

The identifiability of various causal quantities has been extensively studied in the literature [10, 85, 162–164, 178].

4 INSTANTANEOUS NOTIONS OF ALGORITHMIC FAIRNESS

In Section 2, we present normative considerations of fairness and justice, in this section, we present technical details of a non-exhaustive list of instantaneous fairness notions proposed in the literature. Here, by "instantaneous," we are referring to the fact that the fairness inquiry is with respect to a given snapshot of reality. This characteristic is also called "static" fairness in the literature [45]. Considering the fact that instantaneous fairness notions are also considered in dynamic settings in the literature, to avoid confusion, we use the term "instantaneous" to indicate that the fairness notion is not explicitly time-dependent, and we reserve the term "static" to distinguish from "dynamic" when discussing different settings where fairness inquiries take place (we review fairness considerations in the dynamic setting in Section 5). When presenting various previously proposed fairness notions, we unify the notations for consistency while keeping their meanings intact.

4.1 Demographic Parity

Demographic Parity, also known as *Statistical Parity*, is one of the earliest fairness notions proposed in the literature [25, 51, 61, 197]. In the context of binary classification ($\mathcal{Y} = \{0, 1\}$), *Demographic Parity* requires that the ratio of positive decisions among different groups is equal:

$$\forall a, a' \in \mathcal{A} : \; P(\widehat{Y} = 1 \mid A = a) = P(\widehat{Y} = 1 \mid A = a'). \tag{1}$$

[4]With a slight abuse of the notation, if there is no confusion, then we may also use π to denote a set of paths of interest. The path-specific intervention with respect to the set of paths involves transmitting the value change of X along all paths in the set.

In general contexts, *Demographic Parity* is characterized via the independence between the prediction $\widehat{Y}$ and the protected feature A.

Definition 4.1 (Demographic Parity). We say that a predictor $\widehat{Y}$ is fair in terms of *Demographic Parity* with respect to the protected feature A, if $\widehat{Y}$ is independent from A, i.e., $\widehat{Y} \perp\!\!\!\perp A$.

While it is intuitive to characterize fairness through the aforementioned independence, the notion has significant drawbacks [51]. For instance, when there is unobjectionable dependence between the ground truth Y and the protected feature A, i.e., $Y \not\perp\!\!\!\perp A$, by definition the perfect predictor is also dependent on A ($\widehat{Y} \not\perp\!\!\!\perp A$ since $\widehat{Y} = Y$). It is not intuitive why we should rule out the perfect predictor (although this might not be achievable in reality) for the sake of satisfying the *Demographic Parity* fairness requirement on the prediction even if we allow $Y \not\perp\!\!\!\perp A$ in the data.

4.2 Equalized Odds

In light of the limitation of *Demographic Parity*, Hardt et al. [74] propose the *Equalized Odds* notion of fairness. In the context of binary classification, *Equalized Odds* requires that the **True Positive Rate (TPR)** and **False Positive Rate (FPR)** of each group match the population TPR and FPR, respectively:

$$\forall a \in \mathcal{A}, y \in \{0,1\}: \; P(\widehat{Y} = 1 \mid A = a, Y = y) = P(\widehat{Y} = 1 \mid Y = y). \tag{2}$$

In general contexts, this notion is characterized by stating the conditional independence between the prediction $\widehat{Y}$ and the protected feature A given the ground truth of the target Y.

Definition 4.2 (Equalized Odds). We say that a predictor $\widehat{Y}$ is fair in terms of *Equalized Odds* with respect to the protected feature A and the outcome Y, if $\widehat{Y}$ is conditionally independent from A given Y, i.e., $\widehat{Y} \perp\!\!\!\perp A \mid Y$.

The intuition behind this group-level fairness notion is that, once we know the true value of the target (in the hypothetical ideal world), the additional information of the value of the protected feature should not further alter our prediction results.

4.3 Predictive Parity

First proposed by Dieterich et al. [48], *Predictive Parity* is another group-level fairness notion, which is also referred to as calibration, *Test Fairness* [32], and *No Disparate Mistreatment* [195]. In the context of binary classification, *Predictive Parity* requires that among those whose predicted value is positive (negative), their probability of actually having a positive (negative) label should be the same regardless of the value of the protected feature:

$$\forall a \in \mathcal{A}, \hat{y} \in \{0,1\}: \; P(Y = 1 \mid A = a, \widehat{Y} = \hat{y}) = P(Y = 1 \mid \widehat{Y} = \hat{y}). \tag{3}$$

Similar to *Equalized Odds*, *Predictive Parity* can also be characterized through the conditional independence relation among $(A, Y, \widehat{Y})$.

Definition 4.3 (Predictive Parity). We say that a predictor $\widehat{Y}$ is fair in terms of *Predictive Parity* with respect to the protected feature A and the outcome Y, if Y is conditionally independent from A given $\widehat{Y}$, i.e., $Y \perp\!\!\!\perp A \mid \widehat{Y}$.

Although they look similar, *Demographic Parity*, *Predictive Parity*, and *Equalized Odds* are actually incompatible with each other. It is shown independently by Kleinberg et al. [107] and Chouldechova [32] that any two of the three conditions cannot be attained at the same time except in very special cases. For example, one can achieve *Demographic Parity* and *Equalized Odds* at the same

time when A and Y are independent, or $\widehat{Y}$ is a trivial predictor (always constant or completely random).

In fact, the aforementioned incompatibility results also provide additional insights regarding the widely observed trade-offs between fairness and utility [2, 12, 18, 32, 39, 61, 94, 107, 124, 127, 153, 191]. According to the information bottleneck principle [180, 181], we would like the predicted outcome $\widehat{Y}$ to be an information bottleneck through which we capture as much information as possible between the target variable Y and features including, but not limited to, the protected feature. The information bottleneck principle aligns with the conditional independence relation required by *Predictive Parity* notion of fairness (Definition 4.3). This indicates that the unconstrained optimization can have *Predictive Parity* fairness as a byproduct. In other words, there is no conflict in principle, and therefore, no trade-off, between *Predictive Parity* fairness and unconstrained optimizations. This phenomenon is also referred to as an "implicit fairness criterion of unconstrained learning" [116]. This also indicates that notions that are incompatible with *Predictive Parity* will necessarily involve trade-offs between fairness and accuracy compared to unconstrained optimizations.

4.4 Associative Notions of Individual-level Fairness

Apart from associative notions of group-level fairness (e.g., *Demographic Parity*, *Equalized Odds*, *Predictive Parity*), individual-level fairness notions that are based on associative relations among variables are also proposed in the literature.[5] A canonical example of the individual-level associative notion is *Individual Fairness* proposed by Dwork et al. [51]. The intuition behind *Individual Fairness* is that we want similar predicted outcome for similar individuals according to specific similarity metrics. Apart from various choices of the metric, there are different formulations in the literature to mathematically capture the aforementioned intuition. For completeness, we present both the *Lipschitz Mapping*-based formulation [51] and the $(\epsilon - \delta)$ language-based formulation [69] of *Individual Fairness*.

Definition 4.4 ((Lipschitz Mapping) Individual Fairness). We say that a mapping $h : \mathcal{A} \times \mathcal{X} \to \mathcal{Y}$ satisfies *Individual Fairness* if it is L-Lipschitz with respect to appropriate metrics on the domain, i.e., $\mathcal{A} \times \mathcal{X}$, and the codomain, i.e., $\mathcal{Y}$:

$$\forall (a, x), (a', x') \in \mathcal{A} \times \mathcal{X} : \; d_{\mathcal{Y}}\Big(h(a, x), h(a', x')\Big) \leq L \cdot d_{\mathcal{A}\times\mathcal{X}}\Big((a, x), (a', x')\Big).$$

Definition 4.5 ($(\epsilon - \delta)$ Individual Fairness). Let us consider $\epsilon \geq 0$, $\delta \geq 0$, and a mapping $h : \mathcal{A} \times \mathcal{X} \to \mathcal{Y}$. We say that h satisfies *Individual Fairness* if

$$\forall (a, x), (a', x') \in \mathcal{A} \times \mathcal{X} : \; d_{\mathcal{A}\times\mathcal{X}}\Big((a, x), (a', x')\Big) \leq \epsilon \implies d_{\mathcal{Y}}\Big(h(a, x), h(a', x')\Big) \leq \delta.$$

While *Individual Fairness* is general enough to be applicable in various practical scenarios, the specification of the similarity metric is not often straightforward. Recent literature has explored ways to achieve individual fairness of different flavors [17, 66, 70, 76, 86, 91, 93, 102, 105, 134, 154, 160, 193, 194]. The connection between group-level and individual-level fairness notions beyond their seemingly apparent conflicts also draws attentions [21, 64, 169].

4.5 No Direct/Indirect Discrimination

Fairness notions presented in Sections 4.1–4.4 are based on associative relations among variables. Going beyond these observational criteria, it is desirable if we can further capture the structure of the data generating process by making use of causal modeling.

[5] We will see causal notions of individual-level fairness in Section 4.6.

In the legislation literature, discrimination is commonly divided into two categories: direct discrimination (e.g., rejecting a well-qualified loan applicant only because of the demographic identity) and indirect discrimination (e.g., refusing service to areas with certain Zip code). The motivation behind detecting indirect discrimination is that: among the non-protected attributes X, there is a set of attributes whose usage may still remain (potentially) unjustified, i.e., redlining attributes R, although they are not the protected feature itself. In the language of causal reasoning, given a causal graph, we can start from the node for the protected feature and trace along the paths all the way to the node of interest by following the arrowheads in the graph. Therefore, we can characterize direct and indirect discrimination as different path-specific causal effects with respect to the protected feature [135–137, 142, 199, 203–206].

Definition 4.6 (No Direct Discrimination). Let us denote as π_d the path set that contains only the direct path from the protected feature A to the predictor $\widehat{Y}$, i.e., $A \rightarrow \widehat{Y}$. We say that a predictor $\widehat{Y}$ is fair in terms of *No Direct Discrimination* with respect to the protected feature A and the path set π_d, if for any $a, a' \in \mathcal{A}$ and $\hat{y} \in \mathcal{Y}$ the π_d-specific causal effect of the change in A from a to a' on $\widehat{Y} = \hat{y}$ satisfies

$$P\Big(\widehat{Y} = \hat{y} \mid do(A = a'|_{\pi_d})\Big) - P\Big(\widehat{Y} = \hat{y} \mid do(A = a)\Big) = 0. \tag{4}$$

Definition 4.7 (No Indirect Discrimination). Let us denote as π_i the path set that contains all causal paths from the protected feature A to the predictor $\widehat{Y}$, which go though redlining attributes R, i.e., each path within the set π_i includes at least one node from R. We say that a predictor $\widehat{Y}$ is fair in terms of *No Indirect Discrimination* with respect to the protected feature A and the path set π_i, if for any $a, a' \in \mathcal{A}$ and $\hat{y} \in \mathcal{Y}$ the π_i-specific causal effect of the change in A from a to a' on $\widehat{Y} = \hat{y}$ satisfies

$$P\Big(\widehat{Y} = \hat{y} \mid do(A = a'|_{\pi_i})\Big) - P\Big(\widehat{Y} = \hat{y} \mid do(A = a)\Big) = 0. \tag{5}$$

Motivated by the idea of capturing discrimination through different types of causal effects of the protected feature on the predictor, similar notions are also proposed by Kilbertus et al. [104] to further distinguish different types of attributes that are descendants of the protected feature. In particular, for attributes that are influenced by the protected feature A in a manner that we deem as non-discriminatory, i.e., resolving variables, the path-specific causal effects of A on $\widehat{Y}$ through these attributes are "resolved." For attributes that are influenced by A in an unjustifiable way, i.e., proxy variables, the path-specific causal effects of A on $\widehat{Y}$ through these attributes are "unresolved."

Definition 4.8 (No Unresolved Discrimination). We say that a predictor $\widehat{Y}$ is fair in terms of *No Unresolved Discrimination*, if each path from A to $\widehat{Y}$ is blocked by a resolving variable in the corresponding causal graph.

Definition 4.9 (No Proxy Discrimination). We say that a predictor $\widehat{Y}$ is fair in terms of *No Proxy Discrimination* with respect to a proxy R, if for any $r, r' \in \mathcal{R}$ and $\hat{y} \in \mathcal{Y}$:

$$P\Big(\widehat{Y} = \hat{y} \mid do(R = r)\Big) - P\Big(\widehat{Y} = \hat{y} \mid do(R = r')\Big) = 0. \tag{6}$$

Similar to related notions like "explanatory feature" [95], "redlining attribute" [206], and "admissible variables" [156], the notion of "resolving variable" and "proxy variable" are just descendants of A with different user-specified characteristics. Compared to *No Indirect Discrimination*, although *No Proxy Discrimination* is also capturing indirect discrimination through proxy variables, the intervention based on the proxy variable is conceptually easier to parse compared to the intervention on the protected feature itself. The protected feature, e.g., gender or race, is a deeply rooted personal property and it is impossible to perform a randomized trial [183].

4.6 Counterfactual Fairness

So far, the causal notions of fairness (*No Direct/Indirect Discrimination*, *No Unresolved Discrimination*, *No Proxy Discrimination*) are quantifying the discrimination on the group level. *Counterfactual Fairness* proposed by Kusner et al. [111], compared to previous ones, is more fine-grained, since it captures individual-level notion of fairness.

In Section 4.4, we have seen the characterization of individual-level fairness by making use of associative relations among variables. *Counterfactual Fairness*, however, approaches the individual-level fairness problem from a different angle by making use of causal relations among variables. In particular, the intuition behind *Counterfactual Fairness* is that a decision is fair toward an individual if the decision remains the same in the actual world (the current reality) and a properly defined counterfactual world (the hypothetical world where this individual had a different demographic property).

Definition 4.10 (Counterfactual Fairness). Given a causal model $(U, V, \mathbf{F})$ where V consists of all features $V := \{A, X\}$, we say that a predictor $\widehat{Y}$ is fair in terms of *Counterfactual Fairness* with respect to the protected feature A, if for any $a, a' \in \mathcal{A}, x \in \mathcal{X}, \hat{y} \in \mathcal{Y}$ the following holds true:

$$P\Big(\widehat{Y}_{A\leftarrow a}(U) = \hat{y} \mid A = a, X = x\Big) - P\Big(\widehat{Y}_{A\leftarrow a'}(U) = \hat{y} \mid A = a, X = x\Big) = 0. \tag{7}$$

4.7 Path-specific Counterfactual Fairness

The *Path-specific Counterfactual Fairness* notion [29, 192] shares the similar intuition with *Counterfactual Fairness* and captures the difference in decision between the actual world and a counterfactual world.[6] Different from *Counterfactual Fairness*, more fine-grained causal effects are utilized by *Path-specific Counterfactual Fairness*—path-specific counterfactual effects, i.e., the counterfactual causal effects are characterized only through unfair paths.

Definition 4.11 (Path-specific Counterfactual Fairness). Given a causal model $(U, V, \mathbf{F})$ and a factual observation $O = o$, where V consists of all features $V := \{A, X\}$ and $O \subseteq \{A, X, Y\}$, we say that a predictor $\widehat{Y}$ is fair in terms of *Path-specific Counterfactual Fairness (PC Fairness)* with respect to the protected feature A and the path set π, if for any $a, a' \in \mathcal{A}, \hat{y} \in \mathcal{Y}$ the π-specific counterfactual causal effect of the change in A from a to a' on $\widehat{Y} = \hat{y}$ satisfies (let $\bar{\pi}$ denote the set containing all other paths in the graph that are not elements of π):

$$P\Big(\widehat{Y}_{A\leftarrow a'|\pi, A\leftarrow a|\bar{\pi}}(U) = \hat{y} \mid O = o\Big) - P\Big(\widehat{Y}_{A\leftarrow a}(U) = \hat{y} \mid O = o\Big) = 0. \tag{8}$$

For different configurations of the observation $O = o$ and the path set of interest π, *PC Fairness* can capture different types of causal effects, which results in various flavors of fairness notions. For example, if π consists of all paths in the graph and $O = \{A, X\}$, this configuration of *PC Fairness* (for every $O = o$) reduces to *Counterfactual Fairness* [192].

4.8 Remark: Connect Theories of Justice and Notions of Algorithmic Fairness

In Section 2, we have seen that ethical arguments about fairness or justice can vary across conceptual dimensions, scopes, and overarching theoretical frameworks. Although it is less extensively elaborated in the algorithmic fairness literature, the difference in proposed fairness notions reveals the nuances behind different understandings about what and how algorithmic fairness should be captured.

In terms of the overarching theoretical frameworks, on a high level, the commonly used algorithmic fairness notions rest upon specific types of equality, which align with the idea advocated

[6] Wu et al. [192] uses the abbreviated term *PC Fairness* to denote a unified formula for various causal notions of fairness.

in *Egalitarianism*; at the same time, the practice of performance optimization (with fairness considerations) aligns with *Utilitarian* considerations.

In terms of conceptual dimensions, recent algorithmic fairness notions largely follow the *Ideal* methodology where an ideal principle is advocated regarding what the ideally fair world should look like. For example, Definition 4.1 proposes an independence relation as the ideal principle, and Definition 4.10 advocates a zero counterfactual causal effect. The *Nonideal* methodology has attracted attentions in recent algorithmic fairness literature (see, e.g., Fazelpour and Lipton [60]) but is relatively less explored compared to the *Ideal* counterpart. The distinction between *Procedural* and *Substantive* considerations is well-represented by the distinction between causal and associative notions of algorithmic fairness. The form of *Comparative* consideration (i.e., to draw comparisons between individuals) echoes in various individual fairness notions (Section 4.4) as well as other notions that are defined on the amount of effort one needs to make to get preferable results [80, 84, 185].

In terms of the scope of consideration, the current algorithmic fairness literature primarily focuses on *Local* fairness in the sense that the fairness inquiry is limited to the specific scenario at hand. In Section 6, when we present fairness spectra, and in Section 8, when we present the flowchart for algorithmic fairness, we argue that when fairness inquiries are performed in a closed-loop format, one can potentially further improve fairness to a larger scale, i.e., going beyond *Local* fairness and toward *Global* fairness.

5 FAIRNESS IN DYNAMIC SETTINGS

In Section 4, we have seen various fairness notions defined in an instantaneous manner, i.e., with respect to a fixed snapshot of reality. Considering the ever-present changes happening in practical scenarios, it has been widely recognized that fairness audits in a purely static setting may not serve the purpose of understanding the long-term impact of machine learning algorithms [16, 43, 45, 55–57, 75, 77, 78, 80, 82, 83, 97, 106, 108, 115, 117, 130, 133, 147, 176, 185, 189, 207–209]. In this section, we review existing literature on fairness studies in the dynamic setting.

There are application-specific studies in the dynamic fairness literature: the opportunity allocation in labor market [82], a pipeline consisting of college admission followed by hiring [97], the opportunity allocation in credit application [115], and the resource allocation in predictive policing [57].

Apart from application-specific studies, the literature has adopted various analyzing frameworks to approach dynamic fairness audits, for instance, the utilization of the Pólya urn model in incident discovery [57, 82] and intergenerational mobility analysis [77], fairness inquiries conducted through one-step analyses [97, 115, 133, 207], the leverage of **reinforcement learning (RL)** [173] techniques, e.g., **multi-armed bandits (MABs)** [35, 70, 90, 91, 114, 119, 141, 175, 186] and **Markov decision processes (MDPs)** [68, 88, 165, 189, 209, 211], fairness inquiries conducted in online settings (where algorithms improve as new samples arrive sequentially) [15, 16, 55, 78], the challenge introduced by domain shifts [118, 150, 159, 167], the game-theoretic equilibrium analyses [36, 117, 133], the efforts toward providing interpretations of dynamic and long-term fairness via causal modeling [43] and simulation studies [45].

The practical application provides a specific context for fairness considerations, depending on which one would expect context-dependent interpretations of technical findings. This indicates the importance of the modeling choice in the dynamic setting. In the following subsections, we present common choices of analyzing frameworks, namely, the Pólya urn model (Section 5.1), the one-step feedback model (Section 5.2), and the reinforcement learning framework (Section 5.3). Then in Section 5.4, we provide a remark on types of dynamics modeled in the literature.

Considering the fact that take-away messages are closely related to modeling choices, when presenting the previously reported results in the literature, we lay out assumptions and modeling choices before summarizing findings and only resort to detailed technical representations when necessary. Since modeling choices and notations vary across different approaches, in each subsection, we follow the original notation scheme used by authors of the referenced work.

5.1 Choice of Analyzing Framework: Pólya Urn Model

In the (generalized) Pólya urn model, there are two colors of balls, let us say red and black, in the urn. At each time step, one ball is drawn from the urn, then its color is noted and the ball is replaced. There is a replacement matrix of the following form:

$$\begin{array}{lcc} & \text{Red addition} & \text{Black addition} \\ \text{Sample red} & a & b \\ \text{Sample black} & c & d \end{array}, \tag{9}$$

which governs how urn content is updated. For example, if the urn follows the replacement dynamics as detailed in Equation (9), every time we sample a red (black) ball, we replace it and further add a (c) more red balls and b (d) more black balls into the urn.

Ensign et al. [57] use the Pólya urn to model the recording and (re-)occurrence of crime incidents in certain neighborhoods. In particular, they consider an urn that contains two colors of balls (red and black) that correspond to two neighborhoods (A and B). At each time step, the police patrol in neighborhood A (B) corresponds to drawing a red (black) ball from the urn, and observing a crime in the neighborhood corresponds to placing a ball of the same color into the urn. The initially drawn balls will always be replaced before the next time step. The ratio between counts for red and black balls represents the observed crime statistics, and the long-term distribution of color proportions reflects the modeled long-term belief about crime prevalence in neighborhoods.

A similar instantiation of the Pólya urn model is also (implicitly) utilized in intergenerational mobility analysis [77]. In their model, the population consists of two groups, the advantaged group (A) and the disadvantaged group (D). The group identity of an individual is not fixed across the temporal axis. In each time step, the society can only offer opportunities to an α (fixed) fraction of the population, and the problem at hand is how to allocate this limited amount of opportunities in the society. Individual with different socioeconomic status (advantaged/disadvantaged) has different probability of succeeding if provided with an opportunity. Any individual in the disadvantaged group D who succeeds after being offered the opportunity will relocate into the advantaged group A. Then, every individual is replaced with its next generation of the same socioeconomic status, and the aforementioned process continues. In their standard model, the "replacement" of individuals in the new generation is essentially controlled by hyperparameters in the replacement matrix, i.e., the standard Pólya urn model by setting $a = d = 1$ and $b = c = 0$ in Equation (9). If individual's offspring does not perfectly inherit its socioeconomic status, then the generalized Pólya urn model will be utilized.

5.2 Choice of Analyzing Framework: One-step Feedback Model

Different from analyzing dynamic fairness along multiple time steps, previous works also consider one-step feedback models [97, 115, 133, 207].

Kannan et al. [97] focus on a pipeline consisting of college admission and hiring. They propose a two-stage model with the hiring result at the end of the pipeline as the single one-step feedback. Liu et al. [115] utilize a one-step feedback model to study how static fairness notions interact with well-being of agents on the temporal axis.

Mouzannar et al. [133] focus on the *Demographic Parity* (Definition 4.1) form of affirmative action (fairness intervention) and model at the same time (1) a selection process where the utility-maximizing institution performs binary classification according to the qualification of agents from different groups, and (2) the evolution of group qualifications after imposing the selection with affirmative actions. In their one-step feedback model, the institution uses a deterministic threshold policy on the one-dimensional summary attribute of the agent at the time step t, and this selection process influences the group-level qualification profiles at the time step $t + 1$.

Zhang et al. [207] focus on the relation between the enforced fairness and group representations, as well as the impact of decision on underlying feature distributions. They model group representations via a one-step update function, which governs how the expected number of customers in a group at the time step $t + 1$ is determined by quantities at the time step t: the expected number of customers from that group, current customer retention rate, and the expected new customers arrivals from that group.

5.3 Choice of Analyzing Framework: Reinforcement Learning

Previous works have approached dynamic fairness audits via the framework of MABs. Joseph et al. [90, 91] study dynamic fairness in stochastic and contextual bandits problems. In their *Meritocratic Fair* definition of fairness, agents of lower qualification are never favored over agents of higher qualification, despite the possible uncertainty of the algorithm.[7]

Liu et al. [119] utilize the stochastic MAB framework and adopt the "treating similar individuals similarly" [51] notion of individual fairness. Here the notion of "individual" corresponds to an arm, and two arms are pulled near-indistinguishably if they have a "similar" qualification distribution. Liu et al. [119] complement the aforementioned work by Joseph et al. [91] by incorporating a smoothness constraint and providing a protection of higher qualifications over lower qualifications in an on-average manner.

Gillen et al. [70] consider the problem of online learning in linear contextual bandits with an unknown metric-based individual fairness [51]. They assume that only weak feedback, one that flags the violation of an unknown similarity metric but without quantification, is available, and provide an algorithm in this adversarial context.

Li et al. [114] view the hiring process as a contextual bandit problem and pay special attention to the balance between "exploitation" (selecting from groups with proven hiring records) and "exploration" (selecting from under-represented groups to gather information). Li et al. [114] propose an algorithm that emphasizes exploration by evaluating individuals' statistical upside potential, and highlight the importance of incorporating exploration in decision making in the pursuit of dynamic fairness.

Patil et al. [141] consider the fairness requirement of pulling each arm at least some pre-specified fraction of times in the stochastic MAB problem. Wang et al. [186] study the fairness of exposure [166] in the online recommending system, and propose a new objective for the stochastic bandits setting to resolve the issue of winner-takes-all allocation of exposure. Tang et al. [175] consider the setting where past actions can have delayed impacts on arm rewards in the future. They take into account the runaway feedback issue [57] due to action history, and study the delayed-impact phenomenon in the stochastic MAB context.

Previous works have also approached dynamic fairness audits via the framework of MDPs. Jabbari et al. [88] take into consideration the impact of actions on states (environments) and future rewards, and enforce the fairness notion that an algorithm never prefers an action over another

[7]The term "Meritocratic Fairness" is also utilized as a fairness notion to capture (instantaneous) subgroup fairness [102], and should not be confused with the dynamic setting considered by Joseph et al. [90, 91].

if the long-term (discounted) accumulated reward of the latter is higher (*Meritocratic Fair* [91]). Siddique et al. [165] integrate the **generalized Gini social welfare function (GGF)** [190] with **multi-objective Markov decision processes (MOMDPs)**, where rewards take the form of vector instead of scalar, to impose the specific notion of fairness. Zimmer et al. [211] consider the problem of deriving fair policies in cooperative **multi-agent reinforcement learning (MARL)**. Zhang et al. [209] consider *Demographic Parity* and *Equal Opportunity* notions of fairness with respect to the dynamics of group-level qualification, in the **partially observed Markov decision process (POMDP)** setup. They demonstrate the fact that static fairness notions can result in both improvement and deterioration of fairness depending on the specific characteristics of the POMDP. Wen et al. [189] model the feedback effect of decisions as the dynamics of MDPs, and audit fairness with respect to group-conditioned outcomes of agents in terms of *Demographic Parity* and *Equal Opportunity*. Ge et al. [68] consider long-term group-level fairness of exposure [166] with non-fixed group labels in the context of recommending systems, and formulate the recommendation problem as a **constrained Markov decision process (CMDP)**.

5.4 Remark: Differences in Modeled Dynamics

Apart from common choices of analyzing frameworks presented in Sections 5.1–5.3, previous dynamic fairness literature also considers different types of user dynamics, for instance, the retention dynamics of the customer [207], the amplification dynamics of representation disparity [57, 75], the imitation and replicator dynamics of agents [80, 147], the strategic behavior of agents [49, 58, 83, 108, 130], the algorithmic recourse for agents [92, 182, 185], the rational investments of agents [80, 82, 117], the intergenerational mobility [77]. Considering that a comprehensive literature review of algorithmic fairness inquiries in dynamic settings is beyond the scope of our article, we proceed with reflections on algorithmic fairness in the rest of the article (Sections 6–8).[8]

6 DIFFERENT SPECTRA OF FAIRNESS INQUIRIES

In Sections 4 and 5, we have surveyed fairness inquiries in both static and dynamic settings. In this section, we reflect on different spectra of fairness inquiries. We start by revisiting our running example of music school admission, and focus on the intuition behind each question on the inquiry checklist for fairness in this example. The reflection is not limited to any particular notion of fairness in the literature. Instead, we take a step back and think about the exact type of fairness each question is trying to get at by considering, for instance, the unstated assumption, the intended discussing context, and so on. The categorization of previously proposed notions of fairness, as well as technical details of potential modifications to the notion, will be discussed later in Section 8.

6.1 Revisiting Music School Admission Example

In Section 2.1, we considered an empirical scenario of music school admission and presented a list of fairness inquiries one might be interested in. When evaluating whether or not the admission is fair in general, there are additional technical inquiries, i.e, the "algorithmic" part of fairness considerations:

Question 6: With respect to the data that the committee takes as a reference (which contains the admission choices of committees in previous years), is the data free from historical discrimination?

Question 7: If we are willing to believe that the previous admission decisions do not contain any historical discrimination, based on the information at hand, then does the committee

[8]Interested readers please refer to a recent survey on fairness in learning-based sequential decision algorithms [208].

evaluate the qualification of applicants without bias (how the committee of this year evaluates the applicants)?

Question 8: For those applicants who did not manage to get admitted this year, is there any difference in their future developments compared to those who got admitted? Is there any further impact on the representation of their ethnic groups in the violinist community?

As we can see from these fairness inquiries, there are different underlying assumptions behind each question (e.g., the assumption that the previous admission results are free of historical discrimination), which determine the context and object of interest (e.g., the possible discrimination in admission results of previous years or this year specifically). The nuances between various fairness inquiries actually reflect the necessity of disentangling different types of fairness concern and clarifying the tasks that are called for correspondingly.

6.2 Algorithmic Fairness Spectra

In light of the existence of various types of discrimination, the distinction between *Without Disparate Impact* (also referred to as *Outcome Fairness*) and *Without Disparate Treatment* (also referred to as *Procedural Fairness*) has already been proposed in Title VII of the 1964 Civil Rights Act. While the procedural/outcome fairness division (or similarly, the *Procedural* and *Substantive* emphases presented in Section 2) indicates the intuition behind how different kinds of discrimination could occur, we believe that it is still preferable to have an overarching categorization of algorithmic fairness inquiries, namely, *Fairness w.r.t. Data Generating Process*, *Fairness w.r.t. Predicted Outcome*, and *Fairness w.r.t. Induced Impact*. By explicitly presenting the unstated or implicit assumptions, we further clarify the nuances between various types of fairness inquires, so that we can have a better understanding of the relative emphasis we should attribute to different algorithmic fairness spectra.

6.2.1 Fairness w.r.t. Data Generating Process. The primary focus for this type of fairness inquiry is on the underlying data generating process. Multiple factors may contribute to bias in the data [46, 126, 132]: the imperfection of previous human decisions, the lingering effect of historical discriminations, the (potentially) morally neutral statistical bias/error in the sampling and measurement, and so on.

We say the data (i.e., the population) is "clean" as a consequence of data generating process satisfying the fairness notion of interest (e.g., a choice of the practitioner, or a prevailing conception of fairness). The primary goal is therefore to quantify the discrimination with respect to the data itself, without considering downstream tasks like the prediction or decision making. In the previous music school example, Question 6 is a fairness inquiry with respect to the data generating process, inquiring the existence of discrimination in the data that results from the imperfection of previous committee decisions.

6.2.2 Fairness w.r.t. Predicted Outcome. It is a common practice to evaluate the performance of machine learning algorithms by comparing the prediction with the ground truth in the data, which might be quite problematic if the data is already biased. In light of this fact, whenever we utilize the data to train a "fair" prediction algorithm, we actually take one thing for granted (or at least implicitly assumed)—the data itself is "clean" (according to the bias definition of interest). As already pointed out in the literature [101], there is, in general, no one-size-fits-all solution in terms of what fairness notion we should use. Therefore, we do not specify the exact definition of "fair" or "clean," and the aforementioned rationale applies to the fairness notion of interest in practical scenarios.

For *Fairness w.r.t. Predicted Outcome*, we are not encouraging the practice of blindly assuming that the data at hand is unbiased. Instead, we should always keep in mind that when we discuss fairness with respect to the prediction, there is an implicit assumption of "clean" data, which itself is subject to evaluations from the spectrum of *Fairness w.r.t. Data Generating Process*. By making the assumption that the data at hand is "clean," we can lift the burden from the downstream tasks, and emphasize the utilization of information such that fairness with respect to the predicted outcome is guaranteed.

Admittedly, there are different kinds of downstream tasks and not all of them can be solved by developing a predictive model. Nevertheless, this category of fairness inquiry applies to predictive models as well as prediction-based decision-making systems. After all, human decision making also rests upon predictions to some extent [132]. We use the name "Fairness w.r.t. Predicted Outcome" to further indicate the fact that the primary goal is to quantify the discrimination with respect to the prediction of the ground truth. This does not exclude the possibility of considering downstream tasks like single-time or sequential decision making. In the music school example, Question 7 is a fairness inquiry with respect to the predicted outcome, focusing on the decision-making process of the committee under the assumption that the data (for both previous students and the applicants this year) itself is unbiased.

6.2.3 Fairness w.r.t. Induced Impact. The fairness inquiry with respect to the induced impact is different from quantifying discrimination in the data or the predicted outcome. Fairness inquiries in this spectrum focus on parties other than the prediction or decision making, for instance, how individuals could react (e.g., the interplay between the user and the system), how affirmative actions might help achieve fairness (e.g., the policy favor or investment to help the worse-off groups), and so on. Essentially, the primary goal is to consider the possibility of characterizing fairness through the efforts of external entities besides prediction and decision makers. As we will see in Section 8, fairness inquiries can involve external entities, for example, user dynamics, data dynamics, and so on. In the music school example, Question 8 is a fairness inquiry with respect to the induced impact (of deploying a decision-making system).

6.3 Remark: The Necessity of Considering Different Fairness Spectra

In this section, we have seen different spectra of fairness inquiries. Our goal is to provide a road map so that one can zoom in and see which part the current literature fit in and zoom out to see what else we can do with a clear target in mind. Here, we present additional discussions in the form of questions and answers.

6.3.1 Why Distinguish Between Data and Prediction Fairness? To begin with, as we shall see in more detail in Section 8, notions for *Fairness w.r.t. Data Generating Process* are defined without reference to a predictor. Auditing *Fairness w.r.t. Data Generating Process* is irrelevant to what predictor one uses, because the audit itself is with respect to (a sequence of) snapshots of reality. This indicates that fairness endeavor with respect to data and that with respect to predicted outcome may well differ in terms of both technical definitions and objects of interest (e.g., Y vs. $\widehat{Y}$).

Besides, even if the data is "clean," the not-so-careful utilization of the data for prediction may still introduce new discriminations. It is not necessarily the case that the prediction bias results only from data bias. For instance, the unfairness can be introduced in prediction even if the label is fair [9].

Furthermore, there are attainability and optimality analyses with respect to the *Fairness w.r.t. Predicted Outcome* notions themselves. The attainability of prediction fairness, namely, the existence of a predictor that can score zero violation of fairness in the large sample limit, is an asymptotic

property of the fairness notion [177]. Such attainability is not automatically guaranteed with clean data. It characterizes a completely different kind of violation of fairness compared to the empirical error bound of discrimination in finite-sample cases. In practice, although one can always audit violation of *Fairness w.r.t. Predicted Outcome* via an empirical quantification, because of the finite sample size one cannot expect the empirical fairness violation to be exactly zero. The absolute magnitude of the empirical fairness violation is often not informative enough, since it is not clear how small an empirical fairness violation is small enough such that the predicted outcome can be deemed as "fair," i.e., the fairness notion of interest will be attained with zero violation in the large sample limit. Therefore, it is desirable to develop prediction schemes that come with theoretical guarantees with respect to the method itself so that the fairness notion is proved to be attainable in the large sample limit. Then, we can further conduct optimality analysis among the models that have the attainability guarantee.

Last, as we have seen in Section 5, ample evidence has suggested that "fair" predictions can have adverse impact on the fairness of data because of the driving force of involved dynamics. There is no guarantee that the instantaneous rectification in the prediction/decision can somehow magically eliminate data bias.

Technically speaking, enforcing prediction fairness with or without (implicit or explicit) assumptions of clean data does not affect the algorithmic design or implementation. Although Y and $\widehat{Y}$ are in essence both random variables, clearly distinguishing between fairness considerations for each one of them not only offers conceptual clarity but also provides a clearer picture regarding what kind of fairness inquiry one is actually conducting.

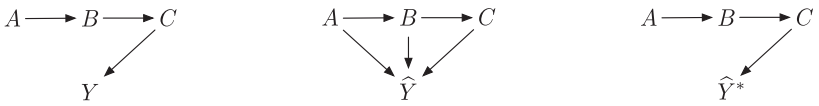
6.3.2 Is Fairness w.r.t. Induced Impact Redundant? While the difference between *Fairness w.r.t. Data Generating Process* and *Fairness w.r.t. Predicted Outcome* is relatively obvious, the distinction between *Fairness w.r.t. Induced Impact* compared to the other two is more subtle. We should not put *Fairness w.r.t. Induced Impact* under the umbrella of either one of the other two categories.

To begin with, *Fairness w.r.t. Induced Impact* itself does not necessarily assume that the data is unbiased (as does *Fairness w.r.t. Data Generating Process*) or the utilization of information is not problematic (as does *Fairness w.r.t. Predicted Outcome*). Therefore, if there is no guarantee regarding *Fairness w.r.t. Data Generating Process* or *Fairness w.r.t. Predicted Outcome*, the fairness violation may involve multiple parties including, but not limited to, the historical discrimination inherited from data, the reckless utilization of information in the prediction/decision-making process, and the interplay between the user and the system.

Furthermore, *Fairness w.r.t. Data Generating Process* and *Fairness w.r.t. Predicted Outcome* focus on either the data itself or the utilization of data, both of which are on the prediction/decision-making side; *Fairness w.r.t. Induced Impact*, however, emphasizes the side of user autonomy and/or data dynamics as well as other possible external entities. In our music school example, the difference in future developments may involve multiple parties, for instance, the committee (the decision maker), the background of the applicant (the user), the policy favor or educational investments for certain ethnic groups (the external entities), and the corresponding bias mitigation cannot be accomplished only through the effort of the music school committee.

7 SUBTLETY: THE ROLE OF CAUSALITY IN FAIRNESS ANALYSIS

In Section 3, we presented multiple instantaneous fairness notions in the literature, many of which leverage the power of causal reasoning. Before discussing the exact location where the notions might fit in the fairness spectra presented in Section 6, we believe it is necessary and important to reflect on subtleties regarding the role of causality in fairness analysis. The consideration of the subtleties motivates our (potential) modifications (in Section 8) on previous fairness notions



(a) The data generating process for the ground truth Y

(b) The data generating process for the classifier/regressor $\widehat{Y}$

(c) The data generating process for the inference output $\widehat{Y}^*$

Fig. 1. Comparison between the causal graphs that represent different data generating processes for the ground truth Y, the prediction result via regression $\widehat{Y}$, and the prediction result via inference $\widehat{Y}^*$.

before applying them to fairness inquires from certain spectrum. In particular, we argue that we should always perform sanity checks to make sure that we are quantifying discrimination in the way that matches underlying assumptions and intended types of fairness inquiries.

7.1 Causal Modeling on the Object of Interest

It is widely recognized in the fairness literature that we can leverage the power of causal reasoning to help us better understand how discrimination propagates through the data generating process [29, 104, 111, 120, 137, 155, 192, 199, 206]. While the assumption of the availability of additional information about the data generating process, e.g., a causal graph, is in general acceptable, we find it questionable to directly assume that the prediction variable $\widehat{Y}$ shares the exactly same causal graph with the ground-truth variable Y.[9]

Let us consider a simple example of the performance of basketball players where there are four variables: the gender of the player (A), the height of the player (B), the player's position (C), the total points scored by the player in this season (Y). Suppose that the data generating process with respect to ground truth (the current reality), i.e., the relation among the measured variables A, B, C, and Y, can be described by Figure 1(a): the gender is a cause of the height; the height determines the position of the player on court; the position determines the total points that the player scores in the season.[10] The task is to come up with the prediction ($\widehat{Y}$) for the total points of this season (Y) based on the information available (the gender A, the height B, and the position C): $\widehat{Y} = f(A, B, C)$ where $f : \mathcal{A} \times \mathcal{B} \times \mathcal{C} \to \mathcal{Y}$ is a classification/regression algorithm. In this case, the prediction result itself can be viewed as a random variable. If we were to draw a graph that represents how $\widehat{Y}$ is generated from (A, B, C), then we will have a data generating process as shown in Figure 1(b). The reason for the extra arrows in Figure 1(b) compared to Figure 1(a) is that the classification/regression algorithm, regardless of the loss function and the optimization techniques, treats available variables merely as input features, which does not really respect the original data generating process in Figure 1(a). However, if, for example, we use a generative model and perform a probabilistic inference task on the outcome where we follow the underlying data generating process, then the inference result $\widehat{Y}^*$ can share the causal graph with that of the ground-truth variable Y (with a change of variable from Y to $\widehat{Y}^*$). The data generating process for the prediction result via inference $\widehat{Y}^*$ (Figure 1(c)) is only different (in terms of the causal graph) from its counterpart for the ground-truth variable Y (Figure 1(a)) up to a substitution of the outcome variable. Usually, we still need stronger assumptions regarding the underlying data generating process to perform the inference tasks, e.g., the availability of an SCM instead of only the causal graph.

[9]For the tasks like prediction, the output $\widehat{Y}$ is usually generated by a classification or regression algorithm in the literature.
[10]This is a simplified model with a limited number of variables involved for illustrative purposes.

As we can see in the previous example, when performing causal reasoning in fairness analysis, we should always be aware of the object of interest, i.e., the variable whose data generating process is subject to fairness consideration. When we directly assume that the causal graph can be shared by the ground truth and the prediction, there could be a mismatch between the causal model (based on which the discrimination is quantified) and the object (whose data generating process is, in fact, *not* described by this model). If there is a mismatch between the causal model and the object of interest, then the result of discrimination quantification could be unpredictable and therefore is hardly justifiable.

7.2 Causal Modeling with the Intended Interpretation

As we have seen in Sections 4.5–4.7, it is a common practice in the causal fairness literature to first combine a causal graph with assumptions on functional forms of the SCM, and then perform fairness audit on the existence of certain causal effects. However, when we are presented a causal model in fairness analysis, there are multiple interpretations that one could potentially apply to the causal model:

Interpretation 1: The causal model is *recovered* from the data at hand through causal discovery (under some conditions);

Interpretation 2: The causal model is based on assumptions or background knowledge, according to which we *believe* the data at hand is generated;

Interpretation 3: The causal model reflects our *expectation* that it should hold true in the hypothetical ideal world where there is no discrimination.

Interpretations 1 and 2 are of a similar flavor, characterizing the causal relations among (measured) variables in the current reality. The corresponding data generating process only reflects the status quo, and the causal model itself does not provide any information regarding the existence of discrimination. The existence of certain causal influence (e.g., in the form of a causal path) can be deemed as morally neutral or morally objectionable depending on the context of discussion as well as the algorithmic fairness definition. For instance, for path-based algorithmic fairness definition (e.g., Definitions 4.8 and 4.9), the path $A \to B \to C$ in causal graph presented in Figure 1(a) may be morally neutral in the basketball player example (gender influences the height of the player, which in turn determines the court position of the player). This path may be morally objectionable in a different context, for example, when A represents individual's nationality, B represents the favorite color, and C represents the reckless driving habit. Although the aforementioned two scenarios share the causal modeling in terms of the causal graph (Figure 1(a)), the existence of discrimination with respect to the causal path $A \to B \to C$ depends on the context and the definition of discrimination, both of which are not specified by the causal modeling itself under Interpretation 1 or Interpretation 2.

Interpretation 3, however, interprets the model as the one that corresponds to the ideal fair world, which may not be the case in the current reality. In the basketball player example, for instance, the practitioner may determine (for some reason) that the causal influence from B (the height of the player) to C (the position on court) is morally objectionable and therefore unfair. The practitioner argues that in the hypothetical world there should not be a path from B to C. Then, if the practitioner would like to see what the distribution of Y would be in the hypothetical ideal world, then it is no longer reasonable to refer to an SCM represented by Figure 1(a), since there is a path $B \to C$.

Among these various possible interpretations, it is not always self-explanatory from the fairness notions themselves which interpretation really corresponds to the causal model presented to us, if there is no further clarifications. In practice, we should not only keep in mind the intuition behind

the fairness notions but also make sure that the interpretation of the causal model we are using truly matches the type of the intended task.

Therefore, categorizing a fairness notion in terms of the type of relation among variables it is defined with (e.g., the division between the associative and causal notions of fairness) may not be informative enough for us to guarantee fairness. The neglect of the subtleties can easily disguise the existence of discrimination. Actually, as we shall see in Section 8, the role of causality in fairness analysis is better represented by the insights it introduces into the problem, under the condition that we carefully perform the aforementioned sanity checks for the intended task.

7.3 To Work *against* or Work *with* the Data Generating Process?

In this subsection, let us turn our focus to methodologies in causal fairness analysis. On a high level, there are two different types of methodologies in terms of how one would like to treat data generating processes (with respect to data itself, or, how prediction is derived): to work *against* or to work *with* the underlying data generating process.

To work *against* the underlying data generating process, one identifies the unwanted causal paths [104] or causal effects [29, 111, 192] as the instantiation of discrimination, and would like to make sure that the prediction is not contaminated by the specified discrimination. Current literature has witnessed various causal fairness notions that adopt the working-against methodology in instantaneous fairness analysis. For example, Definitions 4.6 and 4.7 advocate constraining direct or indirect (interventional) causal effects from the protected feature to the predicted outcome [104, 137, 206]; Definition 4.10 proposes eliminating counterfactual causal effects from the protected feature to the prediction [111]; Definition 4.11 characterizes more fine-grained versions of counterfactual causal effects and defines fairness through the nonexistence of such causal effects [29, 192].

We reflect on the methodology of working *against* the data generating process. To begin with, as presented in Sections 7.1–7.2, the causal modeling of the data generating process involves subtleties with respect to the object of interest and the intended interpretation. If one neglects the subtleties when modeling the data generating process, then the causal analysis for fairness is hardly justifiable.

Besides, even if the aforementioned sanity checks are carefully performed and the causal modeling matches the data generating process of interest, it is not always the case that such data generating process is easily manipulable. If we are enforcing causal fairness with respect to a predictor or decision-making system, then under certain technical conditions the fairness constraints can be implemented, because this specific data generating process is within the control of the algorithm or practitioner. However, if we determine that the underlying data generating process for our current reality contains certain discrimination, then such process is not always within the control of a decision-making system. For instance, social changes do not happen in an abrupt manner, and the fair solution is not simply removing an edge in the causal graph or performing certain interventions once and for all. Furthermore, the expected change in the underlying data generating process often happens on the system level, e.g., the systematic oppression in terms of opportunity hoarding [179], instead of the model level, e.g., a model for automated decision making in the loan application [115].

Last, from a dynamic and long-term perspective, the enforcement of causal fairness in the working-against manner may have unintended downstream outcomes. For instance, Nilforoshan et al. [140] consider *Counterfactual Predictive Parity* [40], *Counterfactual Equalized Odds* [131], and *Conditional Principal Fairness* [87] notions of causal fairness, and perform a one-step feedback analysis (a choice of analyzing framework reviewed in Section 5.2) in a simulated college admission

scenario. Nilforoshan et al. [140] conduct a *Utilitarian* (Section 2.2.3) analysis and demonstrate the trade-offs between causal fairness notions and the downstream social welfare.

To work *with* the data generating process, one recognizes the limited control over the underlying data generating process and focuses on the interplay between decision-making and data dynamics. In light of this, it has been advocated in the recent literature to adopt the methodology of working *with* the data generating process and explore the possibility of inducing a fairer future by analyzing the decision-distribution interplay [176].

8 ENFORCING FAIRNESS IN DIFFERENT SPECTRA

In Sections 6 and 7, we have seen different spectra of algorithmic fairness inquires and the subtleties of applying causal reasoning in fairness inquires, respectively. In this section, we discuss ways to perform fairness audits and achieve algorithmic fairness in different spectra.

In Section 8.1, we propose a flowchart corresponding to our fairness inquiry categorization. Then, in Sections 8.2–8.4, we revisit commonly used fairness notions (reviewed in Section 4), with potential necessary modifications, illustrating how they fit in the fairness spectra (presented in Section 6) so that the intuitive idea of fairness can better match the technical definition (which exact type of fairness we really would like to enforce). In particular, for *Fairness w.r.t. Data Generating Process*, the goal is to **detect** the discrimination embedded in the data; for *Fairness w.r.t. Predicted Outcome*, the goal is to **regulate** the way algorithms utilize information in the data (under the assumption that the data is "clean"); for *Fairness w.r.t. Induced Impact*, the goal is to **compensate** the potential remaining inequalities from the effort of external entities, e.g., the user and/or data dynamics, so that fairness can be further improved. In Section 8.5, we provide a remark on the potential of performing the **correction** of discrimination-contaminated data through a closed-loop analysis across fairness spectra.

8.1 Algorithmic Fairness Flowchart: Answering "How-to" Questions

In Figure 2, we present a road map to navigate through different fairness spectra. Starting from the very beginning, the input for *Fairness w.r.t. Data Generating Process* type of inquiries is the data at hand. Depending on our answer to the question regarding whether or not the data itself is free from any historical discrimination, the data could be readily available for downstream tasks (if we answer "yes") or subject to bias quantification with potential correction (if we answer "no").

Fairness w.r.t. Predicted Outcome, however, assumes that the data itself is "clean," i.e., the data passes the *Fairness w.r.t. Data Generating Process* audits, and puts emphasis on the utilization of information to perform "fair" prediction/decision making. Here "clean" and "fair" are always with respect to the fairness notions of interest, which largely remain choices of the practitioner.

For *Fairness w.r.t. Induced Impact*, the input consists of the "clean" data and the "fair" prediction, and we consider the possibility of further improving fairness by taking into account the contribution from external entities other than the data and the prediction or decision maker (the blue dotted flow in Figure 2). After going through the analysis through different types of fairness emphases, if conditions permit, then new data could be collected. This in turn would be our updated version of the data at hand, which enables us to further check the effectiveness of the elimination of discrimination by a new round of fairness audit (the red dashed flow in Figure 2) and conduct a closed-loop fairness analysis.

8.2 Fairness w.r.t. Data Generating Process

A fairness inquiry from this category focuses on the generating process of the data itself and emphasizes the detection of discrimination within the data without considering downstream tasks. In order to justify the way of discrimination quantification, we need to exploit the relation among

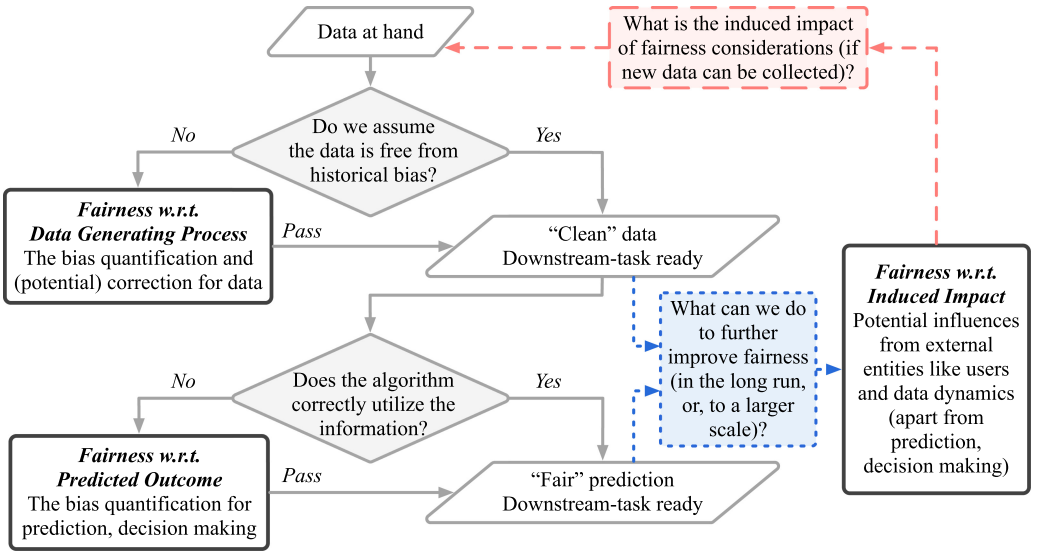


Fig. 2. Flowchart illustrating the road map to navigate through different spectra of fairness inquiries. Starting from the data at hand, based on the answer to the questions, we can sequentially audit fairness with respect to the underlying data generating process, the predicted outcome itself, and the induced impact in the future, respectively. If conditions permit, then the newly collected data could be the starting point for a new round of fairness analysis. With a clear picture in mind that is able to accommodate different types of fairness inquiries, we can conduct algorithmic fairness analysis in a closed-loop manner, making the fairness analysis more principled and to-the-point.

measured variables in terms of the underlying data generating process, which makes causal modeling a perfect tool to achieve the goal. In this subsection, we present our modifications on previously proposed causal notions of fairness, such that the modified notions are suitable for the purpose of auditing fairness with respect to the data generating process.

Multiple causal notions of fairness have been proposed in the literature [29, 103, 104, 111, 137, 156, 192, 199, 204, 206]. However, in light of the frequently neglected subtleties that we discussed in Section 7, we might need to modify causal notions to remedy the mismatch between the intended task and the object or interpretation of interest, so that the intuition behind the notion can be properly expressed. Here for the purpose of illustration, we present the modified versions of *No Direct/Indirect Discrimination* (Definitions 4.6 and 4.7), *Counterfactual Fairness* (Definition 4.10), and *Path-specific Counterfactual Fairness* (Definition 4.11) that we reviewed in Section 4.

Definition 8.1 (No Direct Discrimination (Modified)). Given the causal graph that describes the data generating process of the current reality, let us denote as π_d the path set that contains only the direct path from the protected feature A to the outcome Y, i.e., $A \to Y$. We say that the outcome Y is fair in terms of *No Direct Discrimination* with respect to the protected feature A and the path set π_d, if for any $a, a' \in \mathcal{A}$ and $y \in \mathcal{Y}$ the π_d-specific causal effect of the change in A from a to a' on $Y = y$ satisfies

$$P\Big(Y = y \mid do(A = a'|_{\pi_d})\Big) - P\Big(Y = y \mid do(A = a)\Big) = 0. \tag{10}$$

Definition 8.2 (No Indirect Discrimination (Modified)). Given the causal graph that describes the data generating process of the current reality, let us denote as π_i the path set that contains all causal paths from the protected feature A to the outcome Y, which go though redlining attributes

R, i.e., each path within the set π_i includes at least one node from R. We say that the outcome Y is fair in terms of *No Indirect Discrimination* with respect to the protected feature A and the path set π_i, if for any $a, a' \in \mathcal{A}$ and $y \in \mathcal{Y}$ the π_i-specific causal effect of the change in A from a to a' on $Y = y$ satisfies

$$P\Big(Y = y \mid do(A = a'|_{\pi_i})\Big) - P\Big(Y = y \mid do(A = a)\Big) = 0. \tag{11}$$

Definition 8.3 (Counterfactual Fairness (Modified)). Given a causal model $(U, V, \mathbf{F})$ that describes the data generating process of the current reality, where V consists of all features $V \coloneqq \{A, X\}$, we say that the outcome Y is fair in terms of *Counterfactual Fairness* with respect to the protected feature A, if for any $a, a' \in \mathcal{A}, x \in \mathcal{X}, y \in \mathcal{Y}$ the following holds true:

$$P\Big(Y_{A \leftarrow a}(U) = y \mid A = a, X = x\Big) = P\Big(Y_{A \leftarrow a'}(U) = y \mid A = a, X = x\Big). \tag{12}$$

Definition 8.4 (Path-specific Counterfactual Fairness (Modified)). Given a causal model $(U, V, \mathbf{F})$ that describes the data generating process of the current reality and a factual observation $O = o$, where V consists of all features $V \coloneqq \{A, X\}$ and $O \subseteq \{A, X, Y\}$, we say that the outcome Y is fair in terms of *Path-specific Counterfactual Fairness (PC Fairness)* with respect to the protected feature A and the path set π, if for any $a, a' \in \mathcal{A}, y \in \mathcal{Y}$ the π-specific counterfactual causal effect of the change in A from a to a' on $Y = y$ satisfies (let $\bar{\pi}$ denote the set containing all other paths in the graph that are not elements of π)

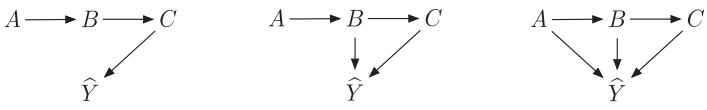
$$P\Big(Y_{A \leftarrow a'|\pi, A \leftarrow a|\bar{\pi}}(U) = y \mid O = o\Big) - P\Big(Y_{A \leftarrow a}(U) = y \mid O = o\Big) = 0. \tag{13}$$

Compared to the original notions (Definitions 4.6, 4.7, 4.10, 4.11), the modified causal notions (Definitions 8.1, 8.2, 8.3, 8.4) are quantifying discrimination with respect to the outcome variable Y instead of the prediction $\widehat{Y}$, using the data generating process behind Y in the current reality. This seemingly trivial modification is more than just exchanges of variables. In practical applications, when we assume the availability, either via an educated guess or from the expert knowledge, of a causal graph that characterizes underlying properties of the data, we are referring to the data generating process with respect to the outcome variable Y, instead of the predictor $\widehat{Y}$ [30, 137].

Furthermore, even if we can draw the causal graph for predictions as illustrated in Figure 1(b) (for prediction via classification/regression) and Figure 1(c) (for prediction via inference), we will still need to make sure that we pair up the object of interest and the technical detail of the corresponding analyzing scheme (e.g., path-based criterion, or causal effect estimation that involves additional information/assumption on the functional class).

Let us revisit the basketball player performance example in Section 7. Suppose that a practitioner would like to audit fairness with respect to the prediction and at the same time understand the source of discrimination, and the practitioner thinks that a causal notion of fairness could be very handy. Suppose, for example, the practitioner picks *Counterfactual Fairness* (Definition 4.10, which is the original notion proposed by Kusner et al. [111]), since this causal notion is with respect to $\widehat{Y}$. There are multiple strategies a practitioner might choose to audit fairness, and for each one of them it is possible to have a mismatch between the mission (which kind of fairness we really would like to capture) and the means (how exactly the fairness audit is carried out):

Strategy 1: The practitioner makes an educated guess regarding how attributes could relate to each other in the data set and draws the causal graph Figure 1(a). Considering the task is to audit fairness on $\widehat{Y}$, the practitioner directly exchanges the variable Y in the graph to $\widehat{Y}$ and draws the graph shown in Figure 3(a). The practitioner then proceeds to the fairness audit via *Counterfactual Fairness* (Definition 4.10) without knowing the detail regarding how $\widehat{Y}$ is computed (which is the regression output).



(a) Strategy 1 (b) Strategy 3 (c) Strategy 4

Fig. 3. Comparison between graphs that the practitioner draws in different strategies.

Strategy 2: The practitioner utilizes the exactly same strategy to audit fairness as in Strategy 1, without knowing the detail regarding how $\widehat{Y}$ is computed (which is, in fact, output of a inference model shown as in Figure 1(c)).

Strategy 3: The practitioner first pictures an idealized fair world where both the height B and the position C are causes of the total points Y scored by the player. Then the practitioner realizes that the task is to audit fairness on $\widehat{Y}$ and draws the graph shown in Figure 3(b). The practitioner proceeds to the fairness audit via *Counterfactual Fairness* (Definition 4.10) without knowing the detail regarding how $\widehat{Y}$ is computed (which is the regression output).

Strategy 4: The practitioner notices that $\widehat{Y}$ is the output of a regression algorithm and draws the causal graph that corresponds to the data generating process of $\widehat{Y}$ as shown in Figure 3(c). The practitioner then proceeds to the fairness audit via *Counterfactual Fairness* (Definition 4.10) with respect to $\widehat{Y}$.

Let us take a closer look at these different strategies. For Strategy 1, there is a mismatch between the object of interest ($\widehat{Y}$) and the corresponding data generating process (it should be the graph shown in Figure 1(b), instead of Figure 3(a)).

For Strategy 2, there seems to be no mismatch between the object of interest ($\widehat{Y}$) and the corresponding data generating process in terms of the causal graph, since Figure 3(a) happens to be identical to Figure 1(c) (except for the asterisk symbol in Figure 1(c)). Although the causal graphs agree with each other, the details of causal modeling (e.g., functional classes in the SCM) may differ across the algorithm builder (who generates $\widehat{Y}$) and the practitioner (who audits fairness on $\widehat{Y}$), which may still incur a mismatch between the object of interest and the corresponding data generating process.

For Strategy 3, there is a mismatch of the causal modeling both in terms of the intended interpretation (using the graph that reflects the hypothetical ideal world) and the object of interest (substituting Y with $\widehat{Y}$ without justification).

For Strategy 4, there seems to be no mismatch, since Figure 3(c) is identical to Figure 1(b). However, while there is no significant difference in terms of technical treatments when estimating causal effects on Y and $\widehat{Y}$ (if we were to draw a causal graph for the regression output), only the data generating process behind Y reflects what happens in the real world. After all, one of the strongest motivations behind the usage of a causal notion is the insight into the data generating process behind the outcome Y in the current reality, but this purpose does not seem to be well-served if we consider the data generating process behind the prediction $\widehat{Y}$.

As we can see from different possible strategies in this example, there are many subtleties involved in enforcing/auditing causal notions of fairness. Neglecting these subtleties may result in mismatches between the mission and the means. Unfortunately, the precautions against these negligence are often not well packed into the causal notions of fairness themselves in the current literature. To some extent, the causal notions of fairness with respect to $\widehat{Y}$ (unintentionally) invites the negligence of subtleties discussed in Section 7.

In fact, it is not uncommon to see (variants of) the aforementioned Strategy 1 utilized in current literature [104, 111, 192, 206]. Therefore, our modification on causal notions of fairness is necessary and important to make sure that the notions are correctly used for the suitable task—to **detect** discrimination within the current data and audit *Fairness w.r.t. Data Generating Process*.

Admittedly, the detection of the existence of discrimination in the data does not easily translate into possible ways to perform correction. Nevertheless, a sensible and justifiable scheme that fully characterizes our intuitions behind fairness considerations would encourage further explorations to better accomplish the task, and therefore, is always desirable. We provide the discussion regarding the potential to correct the data via a closed-loop analysis in Section 8.5.

8.3 Fairness w.r.t. Predicted Outcome

While various fairness notions proposed in the literature are with respect to the prediction $\widehat{Y}$, as discussed in Section 8.2 not all of them are suitable for the intended fairness audit at hand. Different from *Fairness w.r.t. Data Generating Process*, where the goal is to detect the discrimination within data, *Fairness w.r.t. Predicted Outcome* assumes that the data at hand is free from discrimination (in the sense that the data passes the fairness audit from the *Fairness w.r.t. Data Generating Process* category) and **regulates** the utilization of information when performing predictions. In practical scenarios, the prediction is often performed by a classification or regression algorithm, which would only treat available features as input, regardless of the data generating process underlying the real world. Therefore, as a rule of thumb, for *Fairness w.r.t. Predicted Outcome*, associative notions of fairness, e.g., *Individual Fairness* [51], *Demographic Parity* [25], *Equalized Odds* [74], are most suitable for the intended fairness audits in this category.

In the algorithmic fairness literature, the phenomenon of the "trade-off between fairness and accuracy" for the prediction has been widely observed and discussed [2, 12, 18, 32, 39, 61, 94, 107, 124, 127, 146, 153, 191]. However, as is discussed in Section 6, only when we assume/know that the data does not contain discrimination can we really justify the practice of enforcing fairness and accuracy at the same time for the prediction result. After all, if Y contains discrimination, enforcing the prediction $\widehat{Y}$ to be close to Y (even if with fairness regularization) is not desirable. Therefore, for *Fairness w.r.t. Predicted Outcome*, we would like to explicitly assume that the data itself is clean so that we can focus on the utilization of information.

8.4 Fairness w.r.t. Induced Impact

In Section 6, we discussed the difference between *Fairness w.r.t. Induced Impact* and other fairness spectra, i.e., *Fairness w.r.t. Data Generating Process* and *Fairness w.r.t. Predicted Outcome*. In this section, we argue that we can explore the possibility of further improving fairness through the effort of external entities.

As we have discussed in Section 6, we cannot put the *Fairness w.r.t. Induced Impact* inquires under the umbrella of *Fairness w.r.t. Data Generating Process* or *Fairness w.r.t. Predicted Outcome* categories. In light of the practical interpretation of *Fairness w.r.t. Induced Impact* audits, we can go beyond the prediction/decision-making itself and explore the possibility of leveraging the effort of external entities to further improve fairness.

Furthermore, if we observe a shared issue among various prediction/decision-making cases, e.g., the recourse cost for certain group is always higher than others for both loan application and school admission, then this may indicate the disadvantage suffered by the group at a larger scale. This disadvantage may be better **compensated** by (global) policy supports (e.g., investments in education for certain community to improve the overall socioeconomic status in the long run) compared to (localized) separated efforts from prediction/decision-making in different scenarios.

Here, by "global" and "localized," we are referring to the scope of effectiveness (e.g., the *Local* and *Global* views presented in Section 2.2.2): a policy support can potentially be effective in multiple prediction/decision-making scenarios, while prediction/decision-making itself is usually limited to the specific task at hand, i.e., the scenario for which the algorithm is implemented, like loan application or school admission.

Some might argue that the *Fairness w.r.t. Induced Impact* task sounds like *Fairness w.r.t. Data Generating Process*, since we are characterizing historical discrimination in some sense. While *Fairness w.r.t. Data Generating Process* specializes in detecting discrimination (with potential correction) within the data, the scope is limited to measured variables in the data set at hand. The long-term influence on latent attributes, e.g., the unobserved socio-economic status of individuals, are often not readily available for us when we audit fairness with respect to the current reality. One might need to model the decision-distribution interplay and consider the behind-the-scenes situation changes on the unobserved latent causal factors that directly carry out the influence from the current decision to the future data distribution [176].

Some might also argue that the *Fairness w.r.t. Induced Impact* can be enforced in the same way as *Fairness w.r.t. Predicted Outcome* by regulating the utilization of information in the prediction. While it is a reasonable proposal, the focus of the *Fairness w.r.t. Induced Impact* category often involves multiple parties including, but not limited to, the prediction/decision-making, the user dynamics, the external incentives (like affirmative actions). The interplay between these stakeholders cannot be simplified into the analysis on the prediction/decision-making itself and we need to model dynamics for each party separately [80, 115, 176, 209]. The complexity of the practical implication of predicted outcome at a larger spatial or temporal scale also indicates the necessity of interpreting fairness robustness and fairness transferability in terms of not only the predicted outcome itself but also the induced impact [28, 34, 41, 62, 118].

8.5 Remark: Closed-loop Algorithmic Fairness Analysis

As we have seen in Section 5, current dynamic fairness studies already indicate the importance of considering induced impact of predictions/decisions. We argue that the benefit of considering different spectra of fairness inquires can be extended to go beyond merely auditing the existence of bias but also correcting bias in the data.

The road map we presented earlier (Figure 2) is intended to enable a closed-loop fairness analysis by navigating through different spectra of algorithmic fairness inquiries. We do not intend to claim that one can only consider the current fairness endeavor under the condition that the previous step in the flowchart is already satisfied. Instead, we provide a guiding framework so that fairness analysis can follow a principled navigation. For example, a prominent goal of algorithmic fairness inquiries is to make sure the historical bias is eliminated in the future. To achieve this goal, it is not fruitful to consider prediction fairness in a static setting and hope that the prediction will somehow magically eliminate the bias embedded in data itself. Since the underlying data generating process is the object of interest (*Fairness w.r.t. Data Generating Process*), and the prediction/decision making itself does not offer a direct answer regarding how we can manipulate the underlying data generating process, we should instead follow the flowchart (Figure 2) and explore the possibility of inducing a fair data generating process in the future by conducting a closed-loop fairness analysis and considering *Fairness w.r.t. Predicted Outcome* and *Fairness w.r.t. Induced Impact* at the same time.

9 CONCLUSION

In this article, we provide a survey of, a reflection on, and a new perspective for fairness in machine learning. In particular, we propose a framework that consists of fairness considerations from

different perspectives, namely, data generating process, predicted outcome, and induced impact, and provide a road map, along with sanity checks, to navigate through different fairness spectra.

For fairness with respect to data generating process, considering the often neglected subtleties regarding the role played by causality in fairness analysis, we propose necessary modifications to previous causal notions of fairness and discuss the goal of detecting the discrimination within the data. For fairness with respect to predicted outcome, we highlight the importance of clarifying assumptions on the data, as well as the often-overlooked attainability of fairness notions. For fairness with respect to induced impact, we aim to explore the possibility of further improving fairness through the effort of external entities beyond prediction/decision-making.

Future research directions naturally span different spectra of fairness we laid out. For fairness with respect to data generating process, it is desirable to develop methods to evaluate and guarantee the effectiveness of the pursuit of fairness with respect to the underlying data generating process. This is especially important for the potential correction, i.e., going beyond detection, of the discriminations within the data in the long-term, dynamic setting. For fairness with respect to predicted outcome, a thorough understanding of the fairness notion of interest (e.g., the one that is, or will be, deployed in the real world) calls for analysis with respect to attainability and optimality, which, if carefully characterized, is very informative and helpful both in terms of theoretical rigorousness and practical significance (e.g., the development of better learning strategies that come with theoretical guarantees). For fairness with respect to induced impact, the potential unification of the findings from fairness audits conducted in separated but highly-related scenarios (e.g., school admission, loan application, occupational outlook, etc.) would be very helpful to identify potential ways to systematically promote fairness from a wider scope.

The flowchart we propose (Figure 2) also highlights the potential to quantify the effectiveness of fairness endeavor of the current iteration through another round of fairness audits (e.g., the red dashed flow in Figure 2). With meaningful interpretations of the result, the findings from multiple fairness spectra across different rounds of fairness audits would be a very informative guidance (for prediction/decision-making systems, as well as policy designers and lawmakers) to achieve fairness in an organized and principled way, which is of great theoretical and practical significance.

REFERENCES

[1] Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G. Robinson. 2020. Roles for computing in social change. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 252–260.

[2] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. In *Proceedings of the International Conference on Machine Learning*. 60–69.

[3] Elizabeth Anderson. 1999. What is the point of equality? *Ethics* 109, 2 (1999), 287–337.

[4] Elizabeth Anderson. 2010. *The Imperative of Integration*. Princeton University Press.

[5] McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. 2021. What we can't measure, we can't understand: Challenges to demographic data procurement in the pursuit of fairness. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. 249–260.

[6] McKane Andrus and Sarah Villeneuve. 2022. Demographic-reliant algorithmic fairness: Characterizing the risks of demographic data collection in the pursuit of fairness. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. 1709–1721.

[7] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals, and it's biased against blacks. *ProPublica* (2016).

[8] Richard J. Arneson. 1989. Equality and equal opportunity for welfare. *Philos. Studies: Int. J. Philos. Anal. Trad.* 56, 1 (1989), 77–93.

[9] Carolyn Ashurst, Ryan Carey, Silvia Chiappa, and Tom Everitt. 2022. Why fair labels can yield unfair predictions: Graphical conditions for introduced unfairness. Retrieved from https://arXiv:2202.10816.

[10] Chen Avin, Ilya Shpitser, and Judea Pearl. 2005. Identifiability of path-specific effects. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*.

[11] Pranjal Awasthi, Alex Beutel, Matthäus Kleindessner, Jamie Morgenstern, and Xuezhi Wang. 2021. Evaluating fairness of machine learning models under uncertain and incomplete information. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. 206–214.
[12] Sina Baharlouei, Maher Nouiehed, Ahmad Beirami, and Meisam Razaviyayn. 2020. Rényi fair inference. In *Proceedings of the International Conference on Learning Representations*.
[13] Solon Barocas and Andrew D. Selbst. 2016. Big data's disparate impact. *California Law Review* (2016), 671–732.
[14] Solon Barocas, Andrew D. Selbst, and Manish Raghavan. 2020. The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 80–89.
[15] Yahav Bechavod, Christopher Jung, and Zhiwei Steven Wu. 2020. Metric-free individual fairness in online learning. In *Advances in Neural Information Processing Systems*, Vol. 33. 11214–11225.
[16] Yahav Bechavod, Katrina Ligett, Aaron Roth, Bo Waggoner, and Zhiwei Steven Wu. 2019. Equal opportunity in online classification with partial feedback. In *Advances in Neural Information Processing Systems*, Vol. 32.
[17] Elias Benussi, Andrea Patane, Matthew Wicker, L. Laurenti, and Marta Kwiatkowska. 2022. Individual fairness guarantees for neural networks. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence*. IJCAI.
[18] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. A convex framework for fair regression. Retrieved from https://arXiv:1706.02409.
[19] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2021. Fairness in criminal justice risk assessments: The state of the art. *Sociol. Methods Res.* 50, 1 (2021), 3–44.
[20] Reuben Binns. 2018. Fairness in machine learning: Lessons from political philosophy. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. PMLR, 149–159.
[21] Reuben Binns. 2020. On the apparent conflict between individual and group fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 514–524.
[22] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems*. 4349–4357.
[23] Stephan Bongers, Patrick Forré, Jonas Peters, and Joris M. Mooij. 2021. Foundations of structural causal models with cycles and latent variables. *Ann. Stat.* 49, 5 (2021), 2885–2915.
[24] Liam Kofi Bright, Daniel Malinsky, and Morgan Thompson. 2016. Causally interpreting intersectionality theory. *Philos. Sci.* 83, 1 (2016), 60–81.
[25] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building classifiers with independency constraints. In *Proceedings of the IEEE International Conference on Data Mining Workshops*. IEEE, 13–18.
[26] Simon Caton and Christian Haas. 2020. Fairness in machine learning: A survey. Retrieved from https://arXiv:2010.04053.
[27] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. 2019. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 339–348.
[28] Yatong Chen, Reilly Raab, Jialu Wang, and Yang Liu. 2022. Fairness transferability subject to bounded distribution shift. In *Advances in Neural Information Processing Systems*.
[29] Silvia Chiappa. 2019. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 7801–7808.
[30] Silvia Chiappa and William S. Isaac. 2018. A causal Bayesian networks viewpoint on fairness. In *Proceedings of the IFIP International Summer School on Privacy and Identity Management*. Springer, 3–20.
[31] David Maxwell Chickering. 2002. Optimal structure identification with greedy search. *J. Mach. Learn. Res.* 3 (Nov. 2002), 507–554.
[32] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* 5, 2 (2017), 153–163.
[33] Alexandra Chouldechova and Aaron Roth. 2020. A snapshot of the frontiers of fairness in machine learning. *Commun. ACM* 63, 5 (2020), 82–89.
[34] Ching-Yao Chuang and Youssef Mroueh. 2021. Fair mixup: Fairness via interpolation. In *Proceedings of the International Conference on Learning Representations*.
[35] Houston Claure, Yifang Chen, Jignesh Modi, Malte Jung, and Stefanos Nikolaidis. 2020. Multi-armed bandits with fairness constraints for distributing resources to human teammates. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*. 299–308.
[36] Stephen Coate and Glenn C. Loury. 1993. Will affirmative-action policies eliminate negative stereotypes? *Amer. Econ. Rev.* (1993), 1220–1240.
[37] G. A. Cohen, S. de Wijze, M. H. Kramer, and I. Carter. 2009. Fairness and legitimacy in justice, and: Does option luck ever preserve justice. *Hillel Steiner Anat. Justice: Themes Chall.* 16 (2009), 1.

[38] Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. Retrieved from https://arXiv:1808.00023.
[39] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 797–806.
[40] Amanda Coston, Alan Mishler, Edward H. Kennedy, and Alexandra Chouldechova. 2020. Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 582–593.
[41] Andrew Cotter, Maya Gupta, Heinrich Jiang, Nathan Srebro, Karthik Sridharan, Serena Wang, Blake Woodworth, and Seungil You. 2019. Training well-generalizing classifiers for fairness metrics and other data-dependent constraints. In *Proceedings of the International Conference on Machine Learning*. PMLR, 1397–1405.
[42] Elliot Creager, David Madras, Joern-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. 2019. Flexibly fair representation learning by disentanglement. In *Proceedings of the International Conference on Machine Learning*. 1436–1445.
[43] Elliot Creager, David Madras, Toniann Pitassi, and Richard Zemel. 2020. Causal modeling for fairness in dynamical systems. In *Proceedings of the International Conference on Machine Learning*. PMLR, 2185–2195.
[44] Kimberle Crenshaw. 1990. Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stan. L. Rev.* 43 (1990).
[45] Alexander D'Amour, 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 Conference on Fairness, Accountability, and Transparency*. 525–534.
[46] David Danks and Alex John London. 2017. Algorithmic bias in autonomous systems. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. 4691–4697.
[47] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2015. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *Proc. Priv. Enhanc. Technol.* 2015, 1 (2015), 92–112.
[48] William Dieterich, Christina Mendoza, and Tim Brennan. 2016. COMPAS risk scales: Demonstrating accuracy equity and predictive parity. Technical report, Northpointe Inc., Broward County, FL.
[49] Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. 2018. Strategic classification from revealed preferences. In *Proceedings of the ACM Conference on Economics and Computation*. 55–70.
[50] Michele Donini, Luca Oneto, Shai Ben-David, John S. Shawe-Taylor, and Massimiliano Pontil. 2018. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*. 2791–2801.
[51] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. 214–226.
[52] Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. 2018. Decoupled classifiers for group-fair and efficient machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 119–133.
[53] Ronald Dworkin. 2002. *Sovereign Virtue: The Theory and Practice of Equality*. Harvard University Press.
[54] Jon Elster. 1992. *Local Justice: How Institutions Allocate Scarce Goods and Necessary Burdens*. Russell Sage Foundation.
[55] Hadi Elzayn, Shahin Jabbari, Christopher Jung, Michael Kearns, Seth Neel, Aaron Roth, and Zachary Schutzman. 2019. Fair algorithms for learning in allocation problems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 170–179.
[56] Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. 2017. Decision making with limited feedback: Error bounds for recidivism prediction and predictive policing. In *Proceedings of the Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML'17)*.
[57] Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. 2018. Runaway feedback loops in predictive policing. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. PMLR, 160–171.
[58] Andrew Estornell, Sanmay Das, Yang Liu, and Yevgeniy Vorobeychik. 2021. Unfairness despite awareness: Group-fair classification with strategic agents. Retrieved from https://arXiv:2112.02746.
[59] Robin J. Evans. 2016. Graphs for margins of Bayesian networks. *Scand. J. Stat.* 43 (2016), 625–648.
[60] Sina Fazelpour and Zachary C. Lipton. 2020. Algorithmic fairness from a non-ideal perspective. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 57–63.
[61] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 259–268.
[62] Julien Ferry, Ulrich Aivodji, Sébastien Gambs, Marie-José Huguet, and Mohamed Siala. 2022. Improving fairness generalization through a sample-robust optimization method. *Mach. Learn.* (2022), 1–62.

[63] Benjamin Fish, Jeremy Kun, and Ádám D. Lelkes. 2016. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the SIAM International Conference on Data Mining*. SIAM, 144–152.
[64] Will Fleisher. 2021. What's fair about individual fairness?. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 480–490.
[65] James R. Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. 2020. An intersectional definition of fairness. In *Proceedings of the IEEE 36th International Conference on Data Engineering (ICDE'20)*. IEEE, 1918–1921.
[66] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im)possibility of fairness. Retrieved from https://arXiv:1609.07236.
[67] David Gauthier. 1987. *Morals by Agreement*. Clarendon Press.
[68] Yingqiang Ge, Shuchang Liu, Ruoyuan Gao, Yikun Xian, Yunqi Li, Xiangyu Zhao, Changhua Pei, Fei Sun, Junfeng Ge, Wenwu Ou, and Yongfeng Zhang. 2021. Towards long-term fairness in recommendation. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 445–453.
[69] Philips George John, Deepak Vijaykeerthy, and Diptikalyan Saha. 2020. Verifying individual fairness in machine learning models. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI'20)*, Vol. 124. PMLR, 749–758.
[70] Stephen Gillen, Christopher Jung, Michael Kearns, and Aaron Roth. 2018. Online learning with an unknown fairness metric. *Advances in Neural Information Processing Systems* 31 (2018).
[71] Bruce Glymour 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*. 269–278.
[72] Barry Goldman and Russell Cropanzano. 2015. "Justice" and "fairness" are not the same thing. *J. Organiz. Behav.* 36, 2 (2015), 313–318.
[73] Swati Gupta and Vijay Kamble. 2021. Individual fairness in hindsight. *J. Mach. Learn. Res.* 22, 144 (2021), 1–35.
[74] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*. 3315–3323.
[75] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness without demographics in repeated loss minimization. In *Proceedings of the International Conference on Machine Learning*. PMLR, 1929–1938.
[76] Hoda Heidari, Claudio Ferrari, Krishna Gummadi, and Andreas Krause. 2018. Fairness behind a veil of ignorance: A welfare analysis for automated decision making. In *Advances in Neural Information Processing Systems*, Vol. 31.
[77] Hoda Heidari and Jon Kleinberg. 2021. Allocating opportunities in a dynamic model of intergenerational mobility. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. 15–25.
[78] Hoda Heidari and Andreas Krause. 2018. Preventing disparate treatment in sequential decision making. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'18)*. 2248–2254.
[79] Hoda Heidari, Michele Loi, Krishna P. Gummadi, and Andreas Krause. 2019. A moral framework for understanding fair ML through economic models of equality of opportunity. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 181–190.
[80] Hoda Heidari, Vedant Nanda, and Krishna Gummadi. 2019. On the long-term impact of algorithmic decision policies: Effort unfairness and feature segregation through social learning. In *Proceedings of the International Conference on Machine Learning*. PMLR, 2692–2701.
[81] Miguel A. Hernán and James M. Robins. 2020. *Causal Inference: What If*. Chapman & Hall/CRC, Boca Raton, FL.
[82] Lily Hu and Yiling Chen. 2018. A short-term intervention for long-term fairness in the labor market. In *Proceedings of the World Wide Web Conference*. 1389–1398.
[83] Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. 2019. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 259–268.
[84] Wen Huang, Yongkai Wu, Lu Zhang, and Xintao Wu. 2020. Fairness through equality of effort. In *Proceedings of the Web Conference*.
[85] Yimin Huang and Marco Valtorta. 2006. Identifiability in causal Bayesian networks: A sound and complete algorithm. In *Proceedings of the 21st Conference on Artificial Intelligence (AAAI'06)*. 1149–1154.
[86] Christina Ilvento. 2020. Metric learning for individual fairness. In *Proceedings of the 1st Symposium on Foundations of Responsible Computing*.
[87] Kosuke Imai and Zhichao Jiang. 2020. Principal fairness for human and algorithmic decision-making. Retrieved from https://arXiv:2005.10400.
[88] Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. 2017. Fairness in reinforcement learning. In *Proceedings of the International Conference on Machine Learning*. PMLR, 1617–1626.
[89] Taeuk Jang, Feng Zheng, and Xiaoqian Wang. 2021. Constructing a fair classifier with generated fair data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 7908–7916.

[90] Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2016. Fair algorithms for infinite and contextual bandits. Retrieved from https://arXiv:1610.09559.
[91] Matthew Joseph, Michael Kearns, Jamie H. Morgenstern, and Aaron Roth. 2016. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*, Vol. 29.
[92] Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. 2019. Towards realistic individual recourse and actionable explanations in black-box decision making systems. Retrieved from https://arXiv:1907.09615.
[93] Christopher Jung, Michael Kearns, Seth Neel, Aaron Roth, Logan Stapleton, and Zhiwei Steven Wu. 2021. An algorithmic framework for fairness elicitation. In *Proceedings of the 2nd Symposium on Foundations of Responsible Computing*, Vol. 31. 21.
[94] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowl. Info. Syst.* 33, 1 (2012), 1–33.
[95] Faisal Kamiran, Indrė Žliobaitė, and Toon Calders. 2013. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowl. Info. Syst.* 35, 3 (2013), 613–644.
[96] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. 2011. Fairness-aware learning through regularization approach. In *Proceedings of the IEEE 11th International Conference on Data Mining Workshops*. IEEE, 643–650.
[97] Sampath Kannan, Aaron Roth, and Juba Ziani. 2019. Downstream effects of affirmative action. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 240–248.
[98] Atoosa Kasirzadeh and Andrew Smart. 2021. The use and misuse of counterfactuals in ethical machine learning. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. 228–236.
[99] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *Proceedings of the International Conference on Machine Learning*. PMLR, 2564–2572.
[100] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2019. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 100–109.
[101] Michael Kearns and Aaron Roth. 2019. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford University Press.
[102] Michael Kearns, Aaron Roth, and Zhiwei Steven Wu. 2017. Meritocratic fairness for cross-population selection. In *Proceedings of the International Conference on Machine Learning*. PMLR, 1828–1836.
[103] Aria Khademi, Sanghack Lee, David Foley, and Vasant Honavar. 2019. Fairness in algorithmic decision making: An excursion through the lens of causality. In *Proceedings of the World Wide Web Conference*. 2907–2914.
[104] Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, Vol. 30. 656–666.
[105] Michael Kim, Omer Reingold, and Guy Rothblum. 2018. Fairness through computationally-bounded awareness. In *Advances in Neural Information Processing Systems*, Vol. 31.
[106] Young-Chul Kim and Glenn C. Loury. 2018. Collective reputation and the dynamics of statistical discrimination. *Int. Econ. Rev.* 59, 1 (2018), 3–18.
[107] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent trade-offs in the fair determination of risk scores. In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference (ITCS'17)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.
[108] Jon Kleinberg and Manish Raghavan. 2020. How do classifiers induce agents to invest effort strategically? *ACM Trans. Econ. Comput.* 8, 4 (2020), 1–23.
[109] Carl Knight and Zofia Stemplowska. 2011. *Responsibility and Distributive Justice*. Oxford University Press.
[110] Youjin Kong. 2022. Are "intersectionally fair" AI algorithms really fair to women of color? A philosophical analysis. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. 485–494.
[111] Matt J. Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*. 4066–4076.
[112] Steffen L. Lauritzen. 1996. *Graphical Models*. Vol. 17. Clarendon Press.
[113] Steffen L. Lauritzen, A. Philip Dawid, Birgitte N. Larsen, and H.-G. Leimer. 1990. Independence properties of directed Markov fields. *Networks* 20 (1990), 491–505.
[114] Danielle Li, Lindsey R. Raymond, and Peter Bergman. 2020. *Hiring as Exploration*. Technical Report. National Bureau of Economic Research.
[115] Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed impact of fair machine learning. In *Proceedings of the International Conference on Machine Learning*. PMLR, 3150–3158.
[116] Lydia T. Liu, Max Simchowitz, and Moritz Hardt. 2019. The implicit fairness criterion of unconstrained learning. In *Proceedings of the International Conference on Machine Learning*. PMLR, 4051–4060.

[117] Lydia T. Liu, Ashia Wilson, Nika Haghtalab, Adam Tauman Kalai, Christian Borgs, and Jennifer Chayes. 2020. The disparate equilibria of algorithmic decision making when individuals invest rationally. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 381–391.
[118] Yang Liu, Yatong Chen, Zeyu Tang, and Kun Zhang. 2021. Model transferability with responsive decision subjects. Retrieved from https://arXiv:2107.05911.
[119] Yang Liu, Goran Radanovic, Christos Dimitrakakis, Debmalya Mandal, and David C. Parkes. 2017. Calibrated fairness in bandits. Retrieved from https://arXiv:1707.01875.
[120] Joshua R. Loftus, Chris Russell, Matt J. Kusner, and Ricardo Silva. 2018. Causal reasoning for algorithmic fairness. Retrieved from https://arXiv:1805.05859.
[121] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2018. Learning adversarially fair and transferable representations. Retrieved from https://arXiv:1802.06309.
[122] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. 2020. Survey on causal-based machine learning fairness notions. Retrieved from https://arXiv:2010.09553.
[123] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. 2021. On the applicability of machine learning fairness notions. *ACM SIGKDD Explor. Newsl.* 23, 1 (2021), 14–23.
[124] Jérémie Mary, Clément Calauzenes, and Noureddine El Karoui. 2019. Fairness-aware learning for continuous attributes and treatments. In *Proceedings of the International Conference on Machine Learning*. 4382–4391.
[125] Andrew Mason. 2006. *Levelling the Playing Field: The Idea of Equal Opportunity and Its Place in Egalitarian thought*. Oxford University Press, Oxford, UK.
[126] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Comput. Surveys* 54, 6 (2021), 1–35.
[127] Aditya Krishna Menon and Robert C. Williamson. 2018. The cost of fairness in binary classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 107–118.
[128] David Miller. 2001. *Principles of Social Justice*. Harvard University Press.
[129] David Miller. 2021. Justice. In *The Stanford Encyclopedia of Philosophy* (Fall 2021 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.
[130] Smitha Milli, 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*. 230–239.
[131] Alan Mishler, Edward H. Kennedy, and Alexandra Chouldechova. 2021. Fairness in risk assessment instruments: Post-processing to achieve counterfactual equalized odds. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. 386–400.
[132] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2018. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. Retrieved from https://arXiv:1811.07867.
[133] Hussein Mouzannar, Mesrob I. Ohannessian, and Nathan Srebro. 2019. From fair decision making to social equality. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 359–368.
[134] Debarghya Mukherjee, Mikhail Yurochkin, Moulinath Banerjee, and Yuekai Sun. 2020. Two simple ways to learn individual fairness metrics from data. In *Proceedings of the International Conference on Machine Learning*. PMLR, 7097–7107.
[135] Razieh Nabi, Daniel Malinsky, and Ilya Shpitser. 2019. Learning optimal fair policies. In *Proceedings of the International Conference on Machine Learning*. PMLR, 4674–4682.
[136] Razieh Nabi, Daniel Malinsky, and Ilya Shpitser. 2019. Optimal training of fair predictive models. Retrieved from https://arXiv:1910.04109.
[137] Razieh Nabi and Ilya Shpitser. 2018. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32. 1931–1940.
[138] Thomas Nagel. 2005. The problem of global justice. *Philos. Public Affairs* 33, 2 (2005), 113–147.
[139] Arvind Narayanan. 2018. Translation tutorial: 21 fairness definitions and their politics. In *Proceedings of the Conference Fairness Accountability Transparency*, Vol. 1170. 3.
[140] Hamed Nilforoshan, Johann D. Gaebler, Ravi Shroff, and Sharad Goel. 2022. Causal conceptions of fairness and their consequences. In *Proceedings of the International Conference on Machine Learning*. PMLR, 16848–16887.
[141] Vishakha Patil, Ganesh Ghalme, Vineet Nair, and Y. Narahari. 2021. Achieving fairness in the stochastic multi-armed bandit problem. *J. Mach. Learn. Res.* 22 (2021), 1–31.
[142] Judea Pearl. 2009. *Causality*. Cambridge University Press.
[143] Adrián Pérez-Suay, Valero Laparra, Gonzalo Mateo-García, Jordi Muñoz-Marí, Luis Gómez-Chova, and Gustau Camps-Valls. 2017. Fair kernel learning. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 339–355.
[144] Dana Pessach and Erez Shmueli. 2022. A review on fairness in machine learning. *ACM Comput. Surveys* 55, 3 (2022), 1–44.

[145] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press.
[146] Carlos Pinzón, Catuscia Palamidessi, Pablo Piantanida, and Frank Valencia. 2022. On the impossibility of non-trivial accuracy in presence of fairness constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 7993–8000.
[147] Reilly Raab and Yang Liu. 2021. Unintended selection: Persistent qualification rate disparities and interventions. *Advances in Neural Information Processing Systems* 34.
[148] John Rawls. 1971. *A Theory of Justice*. Harvard University Press, Cambridge.
[149] John Rawls. 2001. *Justice as Fairness: A Restatement*. Harvard University Press.
[150] Ashkan Rezaei, Anqi Liu, Omid Memarrast, and Brian D. Ziebart. 2021. Robust fairness under covariate shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 9419–9427.
[151] Thomas Richardson. 2003. Markov properties for acyclic directed mixed graphs. *Scand. J. Stat.* 30 (2003), 145–157.
[152] Yaniv Romano, Stephen Bates, and Emmanuel J. Candès. 2020. Achieving equalized odds by resampling sensitive attributes. Retrieved from https://arXiv:2006.04292.
[153] Andrea Romei and Salvatore Ruggieri. 2014. A multidisciplinary survey on discrimination analysis. *Knowl. Eng. Rev.* 29, 5 (2014), 582–638.
[154] Anian Ruoss, Mislav Balunovic, Marc Fischer, and Martin Vechev. 2020. Learning certified individually fair representations. In *Advances in Neural Information Processing Systems*, Vol. 33. 7584–7596.
[155] Chris Russell, Matt J Kusner, Joshua Loftus, and Ricardo Silva. 2017. When worlds collide: Integrating different counterfactual assumptions in fairness. In *Advances in Neural Information Processing Systems*. 6414–6423.
[156] Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. 2019. Interventional fairness: Causal database repair for algorithmic fairness. In *Proceedings of the International Conference on Management of Data*. 793–810.
[157] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C. Parkes, and Yang Liu. 2019. How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 99–106.
[158] Thomas Scanlon. 2000. *What We Owe to Each Other*. Belknap Press.
[159] Candice Schumann, Xuezhi Wang, Alex Beutel, Jilin Chen, Hai Qian, and Ed H. Chi. 2019. Transfer of machine learning fairness across domains. Retrieved from https://arXiv:1906.09688.
[160] Saeed Sharifi-Malvajerdi, Michael Kearns, and Aaron Roth. 2019. Average individual fairness: Algorithms, generalization, and experiments. In *Advances in Neural Information Processing Systems*, Vol. 32.
[161] Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. 2006. A linear non-Gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.* 7, 10 (2006).
[162] Ilya Shpitser and Judea Pearl. 2006. Identification of conditional interventional distributions. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*. 437–444.
[163] Ilya Shpitser and Judea Pearl. 2007. What counterfactuals can be tested. In *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence*. 352–359.
[164] Ilya Shpitser and Judea Pearl. 2008. Complete identification methods for the causal hierarchy. *J. Mach. Learn. Res.* 9 (2008), 1941–1979.
[165] Umer Siddique, Paul Weng, and Matthieu Zimmer. 2020. Learning fair policies in multi-objective (deep) reinforcement learning with average and discounted rewards. In *Proceedings of the International Conference on Machine Learning*. PMLR, 8905–8915.
[166] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2219–2228.
[167] Harvineet Singh, Rina Singh, Vishwali Mhasawade, and Rumi Chunara. 2021. Fairness violations and mitigation under covariate shift. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. 3–13.
[168] Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. 2019. Learning controllable fair representations. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*. 2164–2173.
[169] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P. Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. 2018. A unified approach to quantifying algorithmic unfairness: Measuring individual and group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2239–2248.
[170] Peter Spirtes. 1995. Directed cyclic graphical representations of feedback models. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*. 491–498.
[171] Peter Spirtes, Clark Glymour, and Richard Scheines. 1993. *Causation, Prediction, and Search*. Springer, New York, NY.
[172] Peter Spirtes, Christopher Meek, and Thomas Richardson. 1995. Causal inference in the presence of latent variables and selection bias. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*. 499–506.
[173] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction*. MIT Press.

[174] Latanya Sweeney. 2013. Discrimination in online ad delivery. *Queue* 11, 3 (2013), 10–29.
[175] Wei Tang, Chien-Ju Ho, and Yang Liu. 2021. Bandit learning with delayed impact of actions. In *Advances in Neural Information Processing Systems*, Vol. 34.
[176] Zeyu Tang, Yatong Chen, Yang Liu, and Kun Zhang. 2023. Tier Balancing: Towards dynamic fairness over underlying causal factors. In *Proceedings of the International Conference on Learning Representations*.
[177] Zeyu Tang and Kun Zhang. 2022. Attainability and optimality: The equalized odds fairness revisited. In *Proceedings of the Conference on Causal Learning and Reasoning*, Vol. 177. PMLR, 754–786.
[178] Jin Tian and Judea Pearl. 2002. A general identification condition for causal effects. In *Proceedings of the Innovative Applications of Artificial Intelligence Conference (AAAI/IAAI'02)*. 567–573.
[179] Charles Tilly. 1998. *Durable Inequality*. University of California Press.
[180] Naftali Tishby, Fernando C. Pereira, and William Bialek. 2000. The information bottleneck method. Retrieved from https://arxiv.org/abs/physics/0004057.
[181] Naftali Tishby and Noga Zaslavsky. 2015. Deep learning and the information bottleneck principle. In *IEEE Information Theory Workshop*. 1–5.
[182] Berk Ustun, Alexander Spangher, and Yang Liu. 2019. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 10–19.
[183] Tyler J. VanderWeele and Whitney R. Robinson. 2014. On causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiology* 25, 4 (2014), 473.
[184] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the IEEE/ACM International Workshop on Software Fairness (FairWare'18)*. IEEE, 1–7.
[185] Julius von Kügelgen, Amir-Hossein Karimi, Umang Bhatt, Isabel Valera, Adrian Weller, and Bernhard Schölkopf. 2022. On the fairness of causal algorithmic recourse. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 9584–9594.
[186] Lequn Wang, Yiwei Bai, Wen Sun, and Thorsten Joachims. 2021. Fairness of exposure in stochastic bandits. In *Proceedings of the International Conference on Machine Learning*. PMLR, 10686–10696.
[187] Serena Wang, Wenshuo Guo, Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, and Michael Jordan. 2020. Robust optimization for fairness with noisy protected groups. In *Advances in Neural Information Processing Systems*, Vol. 33. 5190–5203.
[188] Elizabeth Anne Watkins, Michael McKenna, and Jiahao Chen. 2022. The four-fifths rule is not disparate impact: A woeful tale of epistemic trespassing in algorithmic fairness. Retrieved from https://arXiv:2202.09519.
[189] Min Wen, Osbert Bastani, and Ufuk Topcu. 2021. Algorithms for fairness in sequential decision making. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*. PMLR, 1144–1152.
[190] John A. Weymark. 1981. Generalized Gini inequality indices. *Math. Soc. Sci.* 1, 4 (1981), 409–430.
[191] Michael Wick, Swetasudha Panda, and Jean-Baptiste Tristan. 2019. Unlocking fairness: A trade-off revisited. In *Advances in Neural Information Processing Systems*, Vol. 32. 8780–8789.
[192] Yongkai Wu, Lu Zhang, Xintao Wu, and Hanghang Tong. 2019. PC-fairness: A unified framework for measuring causality-based fairness. In *Advances in Neural Information Processing Systems*, Vol. 32. 3399–3409.
[193] Mikhail Yurochkin, Amanda Bower, and Yuekai Sun. 2020. Training individually fair ML models with sensitive subspace robustness. In *Proceedings of the International Conference on Learning Representations*.
[194] Mikhail Yurochkin and Yuekai Sun. 2021. SenSeI: Sensitive set invariance for enforcing individual fairness. In *Proceedings of the International Conference on Learning Representations*.
[195] Muhammad Bilal Zafar, 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*. 1171–1180.
[196] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*. 962–970.
[197] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *Proceedings of the International Conference on Machine Learning*. PMLR, 325–333.
[198] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 335–340.
[199] Junzhe Zhang and Elias Bareinboim. 2018. Fairness in decision-making—The causal explanation formula. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
[200] Kun Zhang, Biwei Huang, Jiji Zhang, Clark Glymour, and Bernhard Schölkopf. 2017. Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. 1347–1353.
[201] Kun Zhang and Aapo Hyvärinen. 2009. On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI'09)*. AUAI Press, 647–655.

[202] Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2011. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI'11)*. AUAI Press, 804–813.

[203] Lu Zhang and Xintao Wu. 2017. Anti-discrimination learning: A causal modeling-based framework. *Int. J. Data Sci. Anal.* 4, 1 (2017), 1–16.

[204] Lu Zhang, Yongkai Wu, and Xintao Wu. 2016. Situation testing-based discrimination discovery: A causal inference approach. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, Vol. 16. 2718–2724.

[205] Lu Zhang, Yongkai Wu, and Xintao Wu. 2017. Achieving non-discrimination in data release. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1335–1344.

[206] Lu Zhang, Yongkai Wu, and Xintao Wu. 2017. A causal framework for discovering and removing direct and indirect discrimination. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. 3929–3935.

[207] Xueru Zhang, Mohammadmahdi Khaliligarekani, Cem Tekin, and Mingyan Liu. 2019. Group retention when using machine learning in sequential decision making: The interplay between user dynamics and fairness. In *Advances in Neural Information Processing Systems*, Vol. 32. 15269–15278.

[208] Xueru Zhang and Mingyan Liu. 2021. Fairness in learning-based sequential decision algorithms: A survey. In *Handbook of Reinforcement Learning and Control*. Springer, 525–555.

[209] Xueru Zhang, Ruibo Tu, Yang Liu, Mingyan Liu, Hedvig Kjellstrom, Kun Zhang, and Cheng Zhang. 2020. How do fair decisions fare in long-term qualification? In *Advances in Neural Information Processing Systems*, Vol. 33. 18457–18469.

[210] Han Zhao, Amanda Coston, Tameem Adel, and Geoffrey J. Gordon. 2020. Conditional learning of fair representations. In *Proceedings of the International Conference on Learning Representations*.

[211] Matthieu Zimmer, Claire Glanois, Umer Siddique, and Paul Weng. 2021. Learning fair policies in decentralized cooperative multi-agent reinforcement learning. In *Proceedings of the International Conference on Machine Learning*. PMLR, 12967–12978.

Received 9 July 2022; revised 20 February 2023; accepted 2 May 2023