6-8-2022

What-Is and How-To for Fairness in Machine Learning: A Survey, Reflection, and Perspective

Zeyu Tang
Department of Philosophy, Carnegie Mellon University, United States

Jiji Zhang
Department of Religion and Philosophy, Hong Kong Baptist University, Hong Kong

Kun Zhang
Carnegie Mellon University & Mohamed bin Zayed University of Artificial Intelligence

What-Is and How-To for Fairness in Machine Learning: A Survey, Reflection, and Perspective

Zeyu Tang ZEYUTANG@CMU.EDU
Department of Philosophy
Carnegie Mellon University

Jiji Zhang ZHANGJIJI@HKBU.EDU.HK
Department of Religion and Philosophy
Hong Kong Baptist University

Kun Zhang KUNZ1@CMU.EDU
Department of Philosophy, Carnegie Mellon University, and
Machine Learning Department, Mohamed bin Zayed University of Artificial Intelligence

Abstract

Algorithmic fairness has attracted increasing attention in the machine learning community. Various definitions are proposed in the literature, but the differences and connections among them are not clearly addressed. In this paper, we review and reflect on various fairness notions previously proposed in machine learning literature, and make an attempt to draw connections to arguments in moral and political philosophy, especially theories of justice. We also consider fairness inquiries from a dynamic perspective, and further consider the long-term impact that is induced by current prediction and decision. In light of the differences in the characterized fairness, we present a flowchart that encompasses implicit assumptions and expected outcomes of different types of fairness inquiries on the data generating process, on the predicted outcome, and on the induced impact, respectively. This paper demonstrates the importance of matching the mission (which kind of fairness one would like to enforce) and the means (which spectrum of fairness analysis is of interest, what is the appropriate analyzing scheme) to fulfill the intended purpose.

Keywords: Fairness, Causality, Bias Mitigation, Dynamic Process, Machine Learning

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 (Sweeney, 2013); the gender discrimination in job-related ads push (Datta et al., 2015); stereotypes associated with gender in word embeddings (Bolukbasi et al., 2016); the bias against certain ethnic groups in the assessment of recidivism risk (Angwin et al., 2016; Berk et al., 2021); the violation of anti-discrimination law (e.g., Title VII of the 1964 Civil Rights Act) in data mining (Barocas and Selbst, 2016).

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* (Dwork et al., 2012), *Equalized Odds* (Hardt et al., 2016), and *Predictive Parity* (Dieterich et al., 2016; Chouldechova, 2017; Zafar et al., 2017a); there are also causal notions of fairness that are defined in terms of causal relation between variables, e.g., *Counterfactual Fairness* (Kusner et al., 2017), *No Unresolved Discrimination* (Kilbertus et al., 2017), and *Path-specific Counterfactual Fairness* (Chiappa, 2019; Wu et al., 2019). In terms of the scope of application, there are group-level fairness notions, e.g., *Equalized Odds* (Hardt et al., 2016), *Fairness on Average Causal Effect* (Khademi et al., 2019), *Equality of Effort* (Huang et al., 2020); there are also individual-level fairness notions, e.g., *Individual Fairness* (Dwork et al., 2012), *Counterfactual Fairness* (Kusner et al., 2017), *Individual Fairness on Hindsight* (Gupta and Kamble, 2019). In terms of the techniques to eliminate or suppress discrimination, there are *pre-processing* approaches (Calders et al., 2009; Dwork et al., 2012; Zemel et al., 2013; Zhang et al., 2018; Madras et al., 2018; Creager et al., 2019; Zhao et al., 2020), *in-processing* approaches (Kamishima et al., 2011; Pérez-Suay et al., 2017; Zafar et al., 2017a,b; Donini et al., 2018; Song et al., 2019; Mary et al., 2019; Baharlouei et al., 2020; Romano et al., 2020; Wang et al., 2020; Jang et al., 2021), and *post-processing* approaches (Hardt et al., 2016; Fish et al., 2016; Dwork et al., 2018). 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 analysis in dynamical settings (Liu et al., 2018; Hashimoto et al., 2018; Heidari et al., 2019b; Zhang et al., 2019; D'Amour et al., 2020; Zhang et al., 2020; Wen et al., 2021; Heidari and Kleinberg, 2021).

There are explications on available choices to quantify discrimination and enforce fairness in recent survey papers (Romei and Ruggieri, 2014; Loftus et al., 2018; Corbett-Davies and Goel, 2018; Mitchell et al., 2018; Narayanan, 2018; Verma and Rubin, 2018; Caton and Haas, 2020; Makhlouf et al., 2020; Mehrabi et al., 2021; Zhang and Liu, 2021) as well as an investigation into public attitudes towards different notions (Saxena et al., 2019). However, the philosophical and methodological contents of the underlying fairness considerations are often not clearly articulated. In this paper, beyond the aforementioned canonical ways of categorizing fairness notions, we review and reflect on commonly used notions of static fairness in machine learning. In particular, we consider fairness inquiries with different semantic emphases and present a corresponding flowchart to navigate through various fairness spectrums. We believe disentanglement of discriminations based on the intended fairness semantics is vital towards a precise and reasonable quantification of different types of discrimination, so that we can consider suitable fairness spectrum to better accomplish the goal. With an extensive discussion into the 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 paper proceeds as follows: in Section 2 we take a quick look into how fairness and justice are approached in the philosophical discussions; in Section 3 we introduce the notation conventions and provide a brief introduction to causal reasoning; in Section 4 and Section 5 we review commonly used algorithmic fairness notions and present fair machine learning studies in the dynamic setting; in Section 6 we present different spectrums of algorithmic fairness inquiries; in Section 7 we clarify the role of causality in fairness analysis; in Section 8 we propose algorithmic fairness flowchart, from which we can see a clearer picture regarding how we should approach different types of fairness pursuit; 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 (Danks and London, 2017; Binns, 2018; Corbett-Davies and Goel, 2018; Glymour and Herington, 2019; Heidari et al., 2019a; Abebe et al., 2020; Fazelpour and Lipton, 2020; Barocas et al., 2020).

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 2021). Instead, we would like to 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 would like to demonstrate how one can benefit from a rather rich literature of theories of justice and reflect on current literature of algorithmic fairness. The demonstration serves as a starting point, from which one can think about (implicit) intuitions, assumptions, and expectations involved in technical treatments in a principled way. The example also serves as a preamble to our detailed discussions on spectrums 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 spectrums 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 would decide the admission of the 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 might 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 just (with respect to which we compare our current reality), or, do we cope with the injustices in the current world and try to move to something better (e.g., 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

1. It has been recognized that "justice" and "fairness" are not the same thing (see, e.g., Goldman and Cropanzano 2015). Therefore, instead of using "justice" and "fairness" interchangeably, throughout this section we follow the terminology used by the referenced work to avoid conceptual misunderstandings.

admitting a student that was wrongfully denied previously), or, for the purpose of *distributing* admission opportunities among currant applicants?

Question 3 (*Procedural* or *Substantive* emphases) Are we considering fairness in terms of how the committee produce the admission decisions (i.e., the decision-making *procedure*), or, only in terms of how 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)?

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 spectrums 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 aspect(s) 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 paper). Following Miller (2021), 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 THE CONCEPTUAL DIMENSIONS OF JUSTICE

In this section, we take a look at four essential contrasts in the conceptual apprehension of fairness or justice (Miller, 2021), 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" (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) proposed by John Rawls (Rawls, 1971, 2001) counts as an *ideal* principle of justice. The *nonideal* approach, on the other hand, 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 Elizabeth Anderson (Anderson, 2010), 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 (2020) have presented a discussion regarding 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 (Aristotle, 2000). 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, fairness in machine learning is largely limited to considerations that have *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 (Barocas and Selbst, 2016). Thanks to the development of causal analysis (Spirtes et al., 1993; Pearl, 2009), 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 *noncomparative* 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 (demographic group) compared to those received by other applicants (demographic groups), while an inquiry of *noncomparative* consideration may concern whether the decision received by an individual truly respect his/her ability to succeed in the violin program.

2.2.2 THE 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. One difference in scope is the following contrast.

***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 (Rawls, 1971) and local justice (Elster, 1992; Nagel, 2005). 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 THE OVERARCHING THEORETICAL FRAMEWORKS TO DISCUSS JUSTICE

In this section, we present three theoretical frameworks in terms of which justice can be understood, namely, *Utilitarianism*, *Contractarianism*, and *Egalitarianism*.

***Utilitarian* Perspective of Justice** On a high level, *Utilitarianism* aims to maximize the overall welfare, and to bring about the greatest amount (in terms of the aggregated utilities) of good for the greatest number. It has been recognized that pure *utilitarianism* is not the final answer to fairness because of several obstacles it faces (Rawls, 1971, 2001; Dworkin, 2002): 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 intuitions 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* (Gauthier, 1987); 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". 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 (Scanlon, 2000).

***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 (Mason, 2006; Knight and Stemplowska, 2011); *Luck Egalitarianism*, as one type of *Responsibility-sensitive Egalitarianism*, adds an additional restriction that the inequalities resulting from brute luck should be constrained (Arneson, 1989); the debates over the role played by luck and desert also remain a major strand in *Egalitarianism* considerations (Anderson, 1999; Miller, 2001; Cohen et al., 2009).

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 when we review (a non-exhaustive list of) definitions of fairness in machine learning, ideas of justice often echo in the intuitions behind algorithmic fairness notions proposed in the machine learning literature.

3. Technical Preliminaries

In this section, we first present the notation conventions used throughout the paper 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 (which may take values a or a') with domain of value $\mathcal{A}$, additional (observable) features(s) by X, with domain of value $\mathcal{X}$, ground truth (label) variable by Y and its predictor by $\widehat{Y}$, with domain of value $\mathcal{Y}$.

Throughout the paper, 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 garrymendering* when considering subgroups formed by structured combinations of protected features (Kearns et al., 2018, 2019), the challenge of unobserved protected features (Hashimoto et al., 2018; Chen et al., 2019) and uncertain/incomplete information of data pair of protected feature and group truth (Awasthi et al., 2021). 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 *cause* of Y if there is a change of distribution for Y when we apply different *interventions* on X (Spirtes et al., 1993; Pearl, 2009). 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) $\{f_1, f_2, \dots, f_n\}$ where each f_i corresponds to one variable $V_i \in V$ and is a mapping $U \cup V \setminus \{V_i\} \rightarrow V_i$.

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 (Spirtes et al., 1993, 1995; Chickering, 2002; Shimizu et al., 2006; Zhang and Hyvärinen, 2009; Zhang et al., 2011, 2017a). Causal discovery is not directly related to characterization of fairness in machine learning and therefore is not reviewed in this paper.

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).[3]

3.2.2 INTERVENTIONS AND COUNTERFACTUALS

Following Pearl (2009), 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 one specific value 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 $P\big(Y \mid do(X = x)\big)$, 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 directed edges. For example, let a path π from X to Y be a direct path $X \rightarrow Y$ (or an indirect path $X \rightarrow \cdots \rightarrow Y$), then the distribution of Y under the path-specific intervention $do(X = x|_\pi)$ along the path π, is denoted by $P\big(Y \mid do(X = x|_\pi)\big)$, 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 let the value of X unchanged along other paths that are not π".

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 $P\big(Y_{X \leftarrow x}(U) \mid O = o\big)$, 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 $P\big(Y_{X \leftarrow x}(U) \mid O = o\big)$ involves a three-step procedure (as explained in more detail in Pearl 2009):

(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 $P\big(Y_{X \leftarrow x|\pi}(U) \mid O = o\big)$, 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 (see, e.g., Tian and Pearl 2002; Avin et al. 2005; Huang and Valtorta 2006; Shpitser and Pearl 2006, 2007, 2008).

4. 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.

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

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 (see, e.g., D'Amour et al. 2020). 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 (Calders et al., 2009; Dwork et al., 2012; Zemel et al., 2013; Feldman et al., 2015). In the context of binary classification ($\mathcal{Y} = \{0, 1\}$), *Demographic Parity* requires that the ratio of positive decisions among different groups equals to each other:

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

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

Definition 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 (Dwork et al., 2012). 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. (2016) propose *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 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. (2016), *Predictive Parity* is another group-level fairness notion, which is also referred to as calibration, *Test Fairness* (Chouldechova, 2017) and *No Disparate Mistreatment* (Zafar et al., 2017a). 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 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 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. (2017) and Chouldechova (2017) that any two out of three conditions can not 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 tradeoffs between fairness and utility (Kamiran and Calders, 2012; Romei and Ruggieri, 2014; Feldman et al., 2015; Chouldechova, 2017; Berk et al., 2017; Corbett-Davies et al., 2017; Kleinberg et al., 2017; Menon and Williamson, 2018; Agarwal et al., 2018; Mary et al., 2019; Wick et al., 2019; Baharlouei et al., 2020). According to the information bottleneck principle (Tishby et al., 2000; Tishby and Zaslavsky, 2015), 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 3). This indicates that the unconstrained optimization can have *Predictive Parity* fairness as a byproduct, i.e., there is no conflict in principle (and therefore, no tradeoff) between *Predictive Parity* fairness and unconstrained optimizations. This phenomenon is also referred to as an "implicit fairness criterion of unconstrained learning" (Liu et al., 2019). This also indicates that notions that are incompatible with *Predictive Parity* will necessarily involve tradeoffs between fairness and accuracy compared to unconstrained optimizations.

4.4 No Direct/Indirect Discrimination

Previous fairness notions (*Demographic Parity*, *Equalized Odds*, and *Predictive Parity*) 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.

In the legislation literature, the discrimination is commonly divided into two categories: direct discrimination (e.g., rejecting a well-qualified loan applicator 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 although they are not the protected feature itself, i.e., redlining attributes R. 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 (Pearl, 2009; Zhang et al., 2016, 2017b; Zhang and Wu, 2017; Zhang et al., 2017c; Nabi and Shpitser, 2018; Zhang and Bareinboim, 2018; Nabi et al., 2019a,b):

Definition 4 ***No Direct Discrimination*** *Let us denote π_d as the path set that contains only the direct path from the protected feature A to the predictor $\widehat{Y}$, i.e., $A \to \widehat{Y}$. 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 5 ***No Indirect Discrimination*** *Let us denote π_i as 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. (2017) to further distinguish different types of attributes that are decendants 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 influence 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 6 ***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 7 ***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" (Kamiran et al., 2013), "readlining attribute" (Zhang et al., 2017c), and "admissible variables" (Salimi et al., 2019), the notion of "resolving variable" and "proxy variable" are just decendants of A with different user-specified characteristics. Compared to *No Indirect Discrimination*, although *No Proxy Discrimination* is also capturing indirect discrimination through proxy variables, i.e., redlining attributes, the intervention based on the proxy variable is conceptually easier to parse compared to the intervention on the protected feature itself – especially considering the fact that the protected feature, e.g., gender or race, is a deeply rooted personal property and it is impossible to perform a randomized trial (VanderWeele and Robinson, 2014).

4.5 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. (2017), compared to previous ones, is more fine-grained since it captures individual-level notion of fairness.

The canonical individual-level fairness notion is *Individual Fairness* proposed by Dwork et al. (2012). The intuition behind *Individual Fairness* is that we want similar predicted outcome for similar individuals (in terms of the user-specified similarity metric). 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 have explored ways to achieve individual fairness of different flavors (Friedler et al., 2016; Kearns et al., 2017; Gillen et al., 2018; Heidari et al., 2018; Sharifi-Malvajerdi et al., 2019; Mukherjee et al., 2020). The similarity between group-level and individual-level fairness notions beyond their seemingly apparent conflicts also draw attentions (Speicher et al., 2018; Binns, 2020).

Counterfactual Fairness, on the other hand, approaches the individual-level fairness problem from a different angle. In particular, the intuition behind *Counterfactual Fairness* is that a decision is fair towards 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 8 ***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:*

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

4.6 Path-specific Counterfactual Fairness

The *Path-specific Counterfactual Fairness* notion (Chiappa, 2019; Wu et al., 2019), as indicated by the name, shares the similar intuition with *Counterfactual Fairness* and captures the difference in decision between the actual world and a counterfactual world.[4] 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 (that compare the factual world with the counterfactual world) are characterized only through unfair paths.

Definition 9 ***Path-specific Counterfactual Fairness (PC 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*):*

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

4. Wu et al. (2019) uses the abbreviated term *PC Fairness* to denote a unified formula for various causal notions of fairness.

For different configurations of the observation $O = o$ and the path set of interest π, *PC Fairness* can capture different types of causal effect, 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 value of $O = o$) reduces to *Counterfactual Fairness* (Wu et al., 2019).

4.7 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 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 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 how the ideally fair world should look like. For example, Definition 1 proposes an independence relation as the ideal principle, and Definition 8 advocates for zero counterfactual causal effect. The *Nonideal* methodology has attracted attentions in recent algorithmic fairness literature (see, e.g., Fazelpour and Lipton 2020) 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 as well as other notions that are defined on the amount of effort one need to make in order to get preferable results (see, e.g., Heidari et al. 2019b; Huang et al. 2020; von Kügelgen et al. 2022).

In terms of the scope of consideration, the current algorithmic fairness literature primarily focus 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 spectrums, 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 towards *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-lasting changes happening in practical scenarios, it has been widely recognized that fairness audits in purely static setting may not serve the purpose of understanding the long-term impact of machine learning algorithms (Ensign et al., 2017, 2018; Hashimoto et al., 2018; Heidari and Krause, 2018; Hu and Chen, 2018; Kim and Loury, 2018; Liu et al., 2018; Bechavod et al., 2019; Elzayn et al., 2019; Heidari et al., 2019b; Hu et al., 2019; Kannan et al., 2019; Milli et al., 2019; Mouzannar et al., 2019; Zhang et al., 2019; Creager et al., 2020; D'Amour et al., 2020; Kleinberg and Raghavan, 2020; Liu et al., 2020; Zhang et al., 2020; Heidari and Kleinberg, 2021; Raab and Liu, 2021; Wen et al., 2021; Zhang and Liu, 2021; von Kügelgen et al., 2022).

In this section, we review existing literature on fairness studies in the dynamic setting. The practical application provides a specific context for fairness considerations, depending on which one would expect context-dependent interpretations of technical findings. Therefore, we start from presenting application-specific fairness inquiries that are motivated by empirical scenarios (Section 5.1 - Section 5.4); then, after immediate remarks on application-specific inquiries (Section 5.5), we turn to studies that frame dynamic fairness audits in more general settings and present approaches proposed in the literature that vary across choices of analysis frameworks (Section 5.6 - Section 5.8); in Section 5.9, we provide a remark on various types of dynamics modeled in previous literature.

Considering the fact that take-away messages are closely related to modeling choices, when presenting the previously reported results in 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 Application-specific Study: Fairness in Labor Market

The increasing utilization of algorithm-based decision making in the hiring process has provoked discussions about the impact on candidates of different backgrounds (Kleinberg et al., 2018; Gillis and Spiess, 2019; Li et al., 2020). From the dynamic fairness audit perspective, Hu and Chen (2018) focus on labor market dynamics and consider the role of institutes in the relation between group-level outcome of agents and hiring policies. They construct a dynamic reputational model of the labor market to capture the reinforcing nature of disparate outcomes. Resulting from groups' different access to resources, outcome disparities result in variant choices of investment, which have further impact on future outcomes. Even if every agent is rational, the (initial or historical) disparate group outcomes can persist since rational behaviors of agents can "lock" a system in the unfavorable (unfair) equilibrium (Arrow, 1971; Phelps, 1972). Hu and Chen (2018) propose an fairness intervention that first jolts the system out of the unfair steady state, and then pushes the system on a path to a preferable equilibrium that is guaranteed to be self-sustaining (stable) and group-equitable.

5.1.1 Model Setup (Hu and Chen, 2018)

A society consists of workers (with binary group identities $\mu \in \{B, W\}$) who enter the labor market sequentially with time step $t = 0, 1, 2, \ldots$. Prior to entering market, workers invest in human capital by choosing an education investment level, which acts as a observed by noisy signal of the hidden qualification characteristics that is unobservable to the labor market. Hu and Chen (2018) assume inherent equality in ability between groups, and assign differences in observed investment choices or job outcomes to unequal societal standings of workers from different groups. The labor market consists of two segments: a Temporary Labor Market (TLM) and a Permanent Labor Market (PLM). Workers first enter TLM (with TLM institutes making the hiring decisions); those who got hired exert effort on the TLM job, whose good outcome positively affect reputation of individuals and their group (group reputation is aggregated individual reputations with time delay); then, workers enter PLM (with PLM institutes making the position assigning decisions), earning wages according to the assigned position with different skill levels. The following quantities remain constant: the total number of workers within labor markets, the ratio between workers residing in TLM and PLM, and the proportion in population workers with different group identities. The entering, relocating, and leaving of the labor market are governed by Poisson processes with different hyperparameters.

A group's reputation at time step t, denoted as π_t^μ, captures the proportion of individuals in group μ who produce "good" outcomes in the labor market over a certain time interval $[t-\tau, t]$. A group's reputation will benefit workers in that group in the sense that the cost of investment for a worker is more favorable compared to its counterpart with equal ability and equal wage expectation but in a group with lower reputation.

For Temporary Labor Market, the TLM hiring strategy is a mapping $\mathcal{H}_T : \mathbb{R}_{\geq 0} \times \mu \to \{0, 1\}$. The decision received by workers depends on his/her observed 1-D investment level before entering the market and the group identity. The TLM hiring strategy must satisfy *Demographic Parity* notion of fairness (see Definition 1). For Permanent Labor Market, the PLM assigning strategy is a mapping $\mathcal{H}_P : [0, 1] \to \{0, 1\}$. The PLM decision is only a function of individual reputation, which depends on worker's history of observable outcomes. The PLM assigning strategy is purely *reputational* (Tirole, 1996; Winfree and McCluskey, 2005; Levin et al., 2009) and does not have any explicit fairness constraints.

5.1.2 RESULTS (HU AND CHEN, 2018)

In the model setup of Hu and Chen (2018), the fairness intervention only happens at the entry point of the label market, i.e., the *Demographic Parity* constraint enforced on the TLM hiring strategy. Hu and Chen (2018) show that under the specified model setup with the fairness intervention, i.e., *Demographic Parity* hiring in TLM, groups with disparate initial social standing will gradually approach a equitable reputation level according to the time-lag τ. In other word, there exists a unique stable symmetric steady-state equilibrium and a convergence time step T such that $\pi_t^B = \pi_t^W$ holds for any $t > T$.

While the fairness intervention is only imposed at the entry point of the labor market, Hu and Chen (2018) provide explanations behind the existence of such equitable equilibrium. To begin with, as social standing of a group improves, the cost condition (of investment for the expected wage) also improves, and as a result, workers of future generation can benefit from the improvement and are more likely to be qualified (since the cost is cheaper). Furthermore, when TLM hiring strategy imposes the *Demographic Parity* constraint, i.e., the opportunity allocation is independent from the group identity, the strategy coincides with the inherent equality in abilities of different workers (a model assumption as detailed in Section 5.1.1). The *Demographic Parity* TLM hiring strategy ensures that both the inherent ability distribution and the group representation in the labor market are maintained. This fairness intervention at the labor market entry point further guarantees the convergence of group outcomes in PLM, therefore the convergence of group reputation is guaranteed. Hu and Chen (2018) further note that the presented approach is just one way of imposing fairness constraint to derive a group-symmetric outcome under the specified labor market pipeline, and there might be other possibilities to produce such an equilibrium, e.g., in a richer and more complex model that is more closely aligned with the market in reality.

5.2 Application-specific Study: Fairness in Admission-Hiring Pipeline

Kannan et al. (2019) study a two-stage model where a student sequentially (1) get admitted to a college, and then if admitted to the college (2) get hired by an institute. Economic literature has considered the possibility of counteracting discrimination through affirmative actions in two-stage models, where students invest in human capital (at some cost) during the first stage and get into the labor market in the second stage Foster and Vohra (1992); Coate and Loury (1993). It has been shown that there exists a self-confirming equilibrium, i.e., only one group invest in themselves and the job opportunities are subsequently allocated to that group. Although affirmative action interventions have been proposed to jolt the system out of such discriminatory equilibrium Foster and Vohra (1992); Coate and Loury (1993), Kannan et al. (2019) argue that the effect of such interventions happens very slowly, and that next-generation students need to recognize the opportunities offered by affirmative actions and make costly investment accordingly. Instead of studying the "upstream effects" of affirmative actions, i.e., to elaborate the reason behind the different distributions of qualification for different groups of students, Kannan et al. (2019) propose to study the "downstream effects" of policies and examine the short-term effects of affirmative actions that happen quickly enough before the underlying qualification distribution changes.

Kannan et al. (2019) assume that the institute (employer) at the end of the admission-hiring pipeline makes decision purely based on the estimated posterior expectation of students' qualification ("type" in their paper) after observing the available information (e.g., college admission result, grades in school, and group identity). They focus on the role played by the college in the admission-hiring pipeline and study what kinds of fairness goals can be achieved with respect to the final hiring outcome of the pipeline. In particular, they consider possible interventions that can be carried out though the effort from the college by, for instance, designing admission rule and grading policies, and controlling what information they share with the hiring institute about graduated students.

The fairness goals explicitly considered by Kannan et al. (2019) in the context of the two-stage admission-hiring pipeline include:

Goal (1) **Equal Opportunity**: The probability of an individual getting accepted to the college and then ultimately hired by the institute may depend on individual's qualification, but should not depend on individual's demographic group identity after conditioning on the qualification;

Goal (2) **Elimination of Downstream Bias**: The hiring decision made by rational institutes (employers) on the college graduates should not be based on demographic group identity.

5.2.1 MODEL SETUP (KANNAN ET AL., 2019)

Students come from two pre-defined groups, divided according to demographic identities. Each student has an inherent qualification t randomly drawn from a group-dependent Gaussian distribution (the prior distribution), and t is not observable to college and the labor market. Students go through a two-stage pipeline consisting of admission and hiring subsequently. During the first stage, the college admit students based on their college entrance exam performance, which is a noisy signal about students' qualification. During the second stage, those who got admitted to college can be hired by an employer, and the hiring decision is a function of students' college grades.

Kannan et al. (2019) assume that college entrance exam score is an unbiased estimator of students' unmeasured true qualification, and model the score at the entrance exam for each student as a sample from the unit-variance Gaussian distribution centered at his/her qualification. Kannan et al. (2019) further assume that the college uses deterministic admission rules, i.e., admission policies correspond to setting admission thresholds on entrance exam scores, and model the grading policy as a variable-variance Gaussian distribution centered at students' true qualification.[5] By specifying

5. As a direct consequence of assumptions on the admission rule and grading policy, a student's grade is conditionally independent from the entrance exam score, conditional on student's true qualification.

the variance, the college can control the strength of the signal that the labor market can get about students' qualification (apart from the fact that students got admitted to the college). As a modeling choice following Spence (1974), the college can therefore act as a gatekeeper, only signaling the quality and performance of students to the labor market.

The employer has access to the group-dependent prior distributions, and also knows the admission and grading policies of the college. For each hiring opportunity allocated, the employer receives a utility $t - C$, where C is a fixed cost for hiring a student. The employer, as a single-minded utility maximizer, estimates the posterior distribution of a student's qualification based on available information and hires exactly those candidates who yield positive expected utilities.

While the college cannot directly control hiring strategies used by the employer, the college can choose its admission strategy and grading policies. Kannan et al. (2019) explore the possibility of incentivizing the employer to use hiring policies such that the final outcome of the two-stage pipeline satisfies the fairness goal. In particular, Kannan et al. (2019) consider the possibility of enforcing Goal (1) and/or two additional properties to achieve Goal (2): *Irrelevance of Group Membership* (IGM), namely, for every admitted student and for every each value of grade, the indicator that student's posterior expectation is no less than C is independent of student's group identity; *strong Irrelevance of Group Membership* (sIGM), namely, the posterior distribution of student's qualification conditional on being admitted to college is identical for different groups. If both college entrance exam and grading policies follow finite-variance Gaussian distributions, IGM in multiple threshold cases (e.g., multiple employers) implies sIGM.

5.2.2 RESULTS (KANNAN ET AL., 2019)

If initially the qualification distribution is shared by different groups, any group-symmetric admission policy will serve the purpose of achieving specified fairness goals. Therefore, Kannan et al. (2019) consider nontrivial cases where the prior qualification distributions are group-dependent.

Positive Results

- In the noiseless case where college admission is a deterministic function of students' qualification, for any grading policy there exists a simple monotone admission rule that guarantees both *Equal Opportunity*, i.e, Goal (1), and IGM for multiple thresholds, i.e., Goal (2);

- Under the specification of model setup in Section 5.2.1, if the college chooses not to report grades at all to the labor market, for any grading policy there exists a simple monotone admission rule that guarantees both *Equal Opportunity*, i.e, Goal (1), and IGM for multiple thresholds, i.e., Goal (2).

Negative Results

- Under the specification of model setup in Section 5.2.1, if the college report grades informatively, there exists no nontrivial threshold admission rule that can achieve Goal (1) and Goal (2) simultaneously;

- If we limit our attention to only Goal (2), then it is possible to obtain IGM in the single threshold case, but no monotone admission rule can nontrivially obtain sIGM (in cases with single or multiple thresholds).

5.3 Application-specific Study: Fairness in Lending Practice

Liu et al. (2018) focus on lending practices and consider the possible delayed impacts on agent benefits introduced by decisions (selections) made by institutes (e.g., the bank). They propose a one-step feedback model in the lending scenario, and examine under what circumstances do fairness criteria actually promote the long-term well-being of disadvantaged groups. Liu et al. (2018) quantify the long-term impact of decision making on different groups in the population, and consider two specific types of fairness criteria, namely, *Demographic Parity* (Definition 1), and *Equality of Opportunity* (Definition 2 where only TPR is of interest). Liu et al. (2018) study following questions under the one-step feedback analysis framework: Do the aforementioned fairness criteria actually benefit the disadvantaged group? When do they show a clear benefit over unconstrained decision policy? Under what circumstances does utility maximization of the institute align with the interest of the agent?

5.3.1 MODEL SETUP (LIU ET AL., 2018)

There are two groups $j \in \{0, 1\}$ in the population with features aggregated into a summary statistic (over support $\mathcal{X}$), e.g., a credit score, whose distributions π_j are not identical across groups. The credit score has positive correlation with the expected outcome, and a higher score value corresponds to a higher probability of a positive decision (loan approval). The credit score of an agent is observable to the institute, and the institute can choose (not necessarily same) thresholds for each group to decide if loans are approved.

Considering the fact that one of the two groups is disadvantaged, Liu et al. (2018) investigate the potential of improving the score distribution of the disadvantaged group by designing lending policies (via choices of threshold on the credit score of an agent). In particular, the institute will maximize its utility subject to either (1) no constraints at all, or (2) *Demographic Parity* constraint (by requiring equality of selection rates across groups), or (3) *Equality of Opportunity* constraint (by requiring equality of true positive rates across groups). Motivated by the designing idea that institute can have an understanding of the impact over time horizon and therefore can evaluate and redesign the system after one-step feedback, Liu et al. (2018) focus on the impact of a selection policy over one-step interventions and note that one can also repeatedly apply such process over multiple steps.

5.3.2 RESULTS (LIU ET AL., 2018)

Liu et al. (2018) assume the availability of a function $\Delta : \mathcal{X} \to \mathbb{R}$ that characterizes the expected change in the credit score for an agent with a given score in the beginning, after the one-step feedback. They treat $\Delta\mu_j$, the expected mean difference in group $j \in \{0, 1\}$, as the quantity of interest, and examine the long-term impact on the group in terms of the expected changes in the mean: (1) long-term improvement ($\Delta\mu_j > 0$), (2) long-term stagnation ($\Delta\mu_j = 0$), and (3) long-term decline ($\Delta\mu_j < 0$).

Liu et al. (2018) construct *outcome curves* which describe the relation between the expected mean difference after the one-step feedback and the natural parameter regimes (e.g., selection rate of the policy). They find out that both fairness notions (*Demographic Parity* and *Equality of Opportunity*) can lead to all possible outcomes, i.e., long-term improvement, staganation, and decline, in the natural parameter regimes. Furthermore, there are settings under which *Demographic Parity* causes decline while *Equality of Opportunity* causes improvement; there are also settings under which *Equality of Opportunity* causes decline while *Demographic Parity* causes improvement. There is no general rule of thumb as to which fairness criteria always yields better utility outcomes in all possible settings. Liu et al. (2018) emphasize the impossibility of foreseeing the impact of a fairness criterion without a careful model of the delayed outcome. Liu et al. (2018) also point out that the results are consistent with the observation that fairness in machine learning is context-sensitive.

5.4 Application-specific Study: Fairness in Predictive Policing

Ensign et al. (2018) focus on resource allocation in predictive policing, in particular, one of the most popular forms of predictive policing which attempts to determine how to deploy patrol officers given historical crime data, for instance, `PredPol` (Mohler et al., 2015). Under their definition, predictive policing is a practice of decision making with respect to the allocation of patrol officers to different areas (for the purpose of detecting and stopping crimes), based on the historical crime incident data (discovered by police or reported by public) for a collection of regions.

Predictive policing has been increasing deployed to determine not only how to allocate patrol officers, but also who and where to target surveillance, and even future victim predictions (Berk and Bleich, 2013; Perry, 2013; Berk et al., 2021). It has been empirically observed that with the increasing policing efforts based on the discovered crime incidents, the prediction made by `PredPol` algorithm introduces a substantial divergence from the true crime rate of the region, and keeps sending the patrol officers back to the same neighborhoods (Lum and Isaac, 2016). Ensign et al. (2018) note that similar algorithms, for instance, recidivism prediction, hiring algorithms, college admissions, and loan application, also reveal the aforementioned phenomenon (in different contexts), and the outcome of prediction largely determines the feedback algorithm receives.

For the purpose of demonstrating the feedback loop in the practical scenario, Ensign et al. (2018) follow the work of Lum and Isaac (2016) and limit the scope of discussion into patrol resource allocation in predictive policing (in the rest of this section, we will use "patrol allocation" and "predictive policing" interchangeably). Ensign et al. (2018) use the model of Pólya urn to demonstrate why the aforementioned feedback occurs in predictive policing, and provide remedies (in a black-box manner) to improve the behavior of currently available predictive policing algorithms.

5.4.1 MODEL SETUP (ENSIGN ET AL., 2018)

For the purpose of modeling how (bounded-)rational players interact with each other in a two-player game in the predictive policing context, Ensign et al. (2018) instantiate a generalized Pólya urn model (Erev and Roth, 1998; Pemantle, 2007), where the recording of crime incidents is modeled through drawing and replacing balls in the urn.[6] The precinct has one police officer who patrol two regions. Every day this officer is sent by the predictive policing algorithm to one of two regions where there might be crime incidents observed by the officer (such incidents are called "discovered" incidents); in addition, public can also (faithfully) report crime incidents to the officer if present in the region (such incidents are called "reported" incidents).

The goal specified by Ensign et al. (2018) is to distribute the policing resource in proportion to the number of crime incidents in each area, i.e., all neighborhoods are "perfectly" policed. Ensign et al. (2018) note that a weighted sum of "discovered" and "reported" incidents gives the total count of crimes, among which "discovered" incidents are directly implicated in the feedback loop (since police resources are allocated based on the predictive policing algorithm) while "reported" incidents are not. Ensign et al. (2018) further assume that an officer discovers a crime (public report a crime)

6. The mathemetical formulation of the Pólya urn model can be found in Section 5.6.

with probability that equals to the ground-truth crime rate, and that crime counts enter the data set only if the region is patrolled (i.e., if no police officer patrols the region, there is no crime discovered, and also no crime reported).

5.4.2 RESULTS (ENSIGN ET AL., 2018)

By describing different scenario with different urns that are parameterized differently, Ensign et al. (2018) demonstrate different long-term impacts of the feedback loop in predictive policing:

- If the crime rate is uniform across regions, the long-term policing resource allocation only depends on the prior belief about the crime rate of regions (which might not correctly reflect the true crime rates). In other words, the (probably wrong) prior belief coupled with the lack of feedback about the unobserved region prevents the system from learning that crime rates in two regions are in fact identical.

- If the crime rate is not uniform across regions, when we only consider "discovered" incidents, in general the runaway feedback loop effect will occur: the algorithm will predict one region with a much higher crime rate than another even if the actual crime rates are similar.

- If the crime rate is not uniform across regions, when we consider both "discovered" and "reported" incidents, the only scenario where feedback loop does not drive the outcome away from the truth is when we effectively ignore feedback (e.g., by down-weighting the "discovered" incident counts).

In light of these results, Ensign et al. (2018) propose a black-box method to counteract the runaway feedback loop effects in predictive policing by filtering input to the system (e.g., reweighting "discovered" and "reported" incidents).

Elzayn et al. (2019), also working on the resource allocation problem in predictive policing, go one step further by proposing an effective algorithm that is not limited to the post-processing manner (as in Ensign et al. 2018). They show that their algorithm (under simplification assumptions) can converge to an optimal fair allocation of resources even if the stationary ground-truth crime rates of regions are unknown, therefore does not suffer from the runaway feedback loop effect.

5.5 Remark: Application-specific Studies

In Section 5.1 - Section 5.4, we present application-specific studies in the dynamic fairness literature: opportunity allocation in labor market (Hu and Chen, 2018), a pipeline consisting of college admission followed by hiring (Kannan et al., 2019), opportunity allocation in credit application (Liu et al., 2018), and resource allocation in predictive policing (Ensign et al., 2018).

Apart from application-specific studies, literature has adopted various analyzing frameworks to approach dynamic fairness audits: the utilization of Pólya urn model in incident discovery (Hu and Chen, 2018; Ensign et al., 2018) and intergenerational mobility analysis (Heidari and Kleinberg, 2021), fairness inquiries conducted through one-step analyses (Liu et al., 2018; Kannan et al., 2019; Mouzannar et al., 2019; Zhang et al., 2019), the leverage of Reinforcement Learning (Sutton and Barto, 2018) techniques, e.g., Multi-Armed Bandits (MABs) (Joseph et al., 2016b,a; Liu et al., 2017; Gillen et al., 2018; Li et al., 2020; Claure et al., 2020; Patil et al., 2021; Wang et al., 2021; Tang et al., 2021) and Markov Decision Processes (MDPs) (Jabbari et al., 2017; Siddique et al., 2020; Zhang et al., 2020; Wen et al., 2021; Zimmer et al., 2021; Ge et al., 2021), fairness inquiries conducted in online settings (where algorithms improve as new samples arrive sequentially) (Heidari and Krause, 2018; Bechavod et al., 2019; Elzayn et al., 2019; Bechavod et al., 2020), the challenge introduced by domain shifts (Schumann et al., 2019; Singh et al., 2021; Rezaei et al., 2021; Liu et al., 2021), the game-theoretic equilibrium analyses (Coate and Loury, 1993; Mouzannar et al., 2019; Liu et al., 2020), the efforts towards providing interpretations of dynamic and long-term fairness via causal modeling (Creager et al., 2020) and simulation studies (D'Amour et al., 2020).

In the following subsections, we present common choices of analyzing frameworks, namely, the Pólya urn model (Section 5.6), the one-step feedback model (Section 5.7), and the reinforcement learning (RL) framework (Section 5.8).

5.6 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 (each color correspond to a region in the area). 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:

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

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. (2018) 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 reflect the modeled long-term belief about crime prevalence in neighborhoods.

A similar instantiation of the Pólya urn model is also (implicitly) utilized in the intergenerational mobility analysis conducted by Heidari and Kleinberg (2021). In their model, population consist of two groups, the advantaged group (A) and the disadvantaged group (D), whose group identity is not fixed across the temporal axis. In each time step, 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 of 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, the generalized Pólya urn model will be utilized.

5.7 Choice of Analyzing Framework: One-step Feedback Model

Different from analyzing dynamic fairness along multiple time steps, previous works also consider one-step (two-stage) feedback models (Liu et al., 2018; Kannan et al., 2019; Mouzannar et al., 2019; Zhang et al., 2019).

As detailed in Section 5.2, Kannan et al. (2019) focus on a pipeline consisting of college admission and hiring, and propose a two-stage model with a single feedback (the hiring result at the end of the pipeline). As detailed in Section 5.3, Liu et al. (2018) utilize a one-step feedback model to study how static fairness notions interact with well-beings of agents on the temporal axis.

Mouzannar et al. (2019) focus on the *Demographic Parity* (Definition 1) form of affirmative action (fairness intervention) and model at the same time (1) a selection process where the utility-maximizing institute 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 institute 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. (2019) 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 exogenous arrivals (new customers) from that group.

5.8 Choice of Analyzing Framework: Reinforcement Learning

Previous works have approached dynamic fairness audits via the framework of Multi-Armed Bandits (MABs). Joseph et al. (2016b,a) 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 of the possible uncertainty of the algorithm.[7] Liu et al. (2017) utilize the stochastic MAB framework and adopt the "treating similar individuals similarly" (Dwork et al., 2012) 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. (2017) complement the aforementioned work by Joseph et al. (2016b) by incorporating a smoothness constraint and providing a protection of higher qualifications over lower qualifications in an on-average manner. Gillen et al. (2018) consider the problem of online learning in linear contextual bandits with an unknown metric-based individual fairness Dwork et al. (2012). They assume that only weak feedback, 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. (2020) view the hiring process as a contextual bandit problem and pay special attention to the balance between "exploitation" (selection from groups with proven hiring records) and "exploration" (selecting from under-represented groups to gather information). Li et al. (2020) propose an algorithm that emphasizes on 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. (2021) consider the fairness requirement of pulling each arm at least some pre-specified

7. The term "Meritocratic Fairness" is also utilized as a fairness notion to capture (instantaneous) subgroup fairness (Kearns et al., 2017), and should not be confused with the dynamic setting considered in Joseph et al. (2016b,a).

fraction of times in the stochastic MAB problem. Wang et al. (2021) study the fairness of exposure (Singh and Joachims, 2018) 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. (2021) consider the setting where past actions can have delayed impacts on arm rewards in the future. They take into account the runaway feedback issue (Ensign et al., 2018) 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 Markov Decision Processes (MDPs). Jabbari et al. (2017) 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 if the long-term (discounted) accumulated reward of the latter is higher (*Meritocratic Fair* as in Joseph et al. 2016b). Siddique et al. (2020) integrate the *generalized Gini social welfare function* (GGF) (Weymark, 1981) with multi-objective Markov decision process (MOMDP), where rewards take the form of vector instead of scalar, to impose the specific notion of fairness. Zimmer et al. (2021) consider the problem of deriving fair policies in cooperative multi-agent reinforcement learning (MARL). Zhang et al. (2020) 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. (2021) 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 notions *Demographic Parity* and *Equal Opportunity* of fairness. Ge et al. (2021) consider long-term group-level fairness of exposure (Singh and Joachims, 2018) with non-fixed group labels in the context of recommending systems, and formulate the recommendation problem as a constrained Markov decision process (CMDP).

5.9 Remark: Differences in Modeled Dynamics

Apart from common choices of analyzing frameworks presented in Section 5.6 - Section 5.8, previous dynamic fairness literature also considers different types of user dynamics, for instance, the retention dynamics of the customer (Zhang et al., 2019), the amplification dynamics of representation disparity (Hashimoto et al., 2018; Ensign et al., 2018), the imitation and replicator dynamics of agents (Heidari et al., 2019b; Raab and Liu, 2021), the strategic behavior of agents (Dong et al., 2018; Hu et al., 2019; Milli et al., 2019; Kleinberg and Raghavan, 2020; Estornell et al., 2021), the algorithmic recourse for agents (Ustun et al., 2019; Joshi et al., 2019; von Kügelgen et al., 2022), the rational investments of agents (Hu and Chen, 2018; Heidari et al., 2019b; Liu et al., 2020), the intergeneration mobility (Heidari and Kleinberg, 2021).

Considering that a comprehensive literature review of algorithmic fairness inquiries in dynamical settings is beyond the scope of our paper, we proceed with reflections on algorithmic fairness in the rest of the paper (Section 6 - Section 8).[8]

8. Interested readers please refer to a recent survey on fairness in learning-based sequential decision algorithms (Zhang and Liu, 2021).

6. Different Spectrums of Fairness Inquiries

In Section 4 and Section 5 we have surveyed fairness inquiries in both static and dynamic settings. In this section, we reflect on different spectrums of fairness inquiries. We start by revisiting our running example of music school admission, and focus on the intuition behind each question on the audit 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 audit question is trying to get at by considering, for instance, the unstated assumption, the intended discussing context, etc. 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 Revisit 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 inquiries out of technical considerations (the "algorithmic" part of fairness considerations):

Question 6 With respect to the data that the committee takes as a reference (which reflects 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 reflect any historical discrimination, based on the information at hand, does the committee evaluate the qualification of applicants without bias (how 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 Spectrums

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 get involved, 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

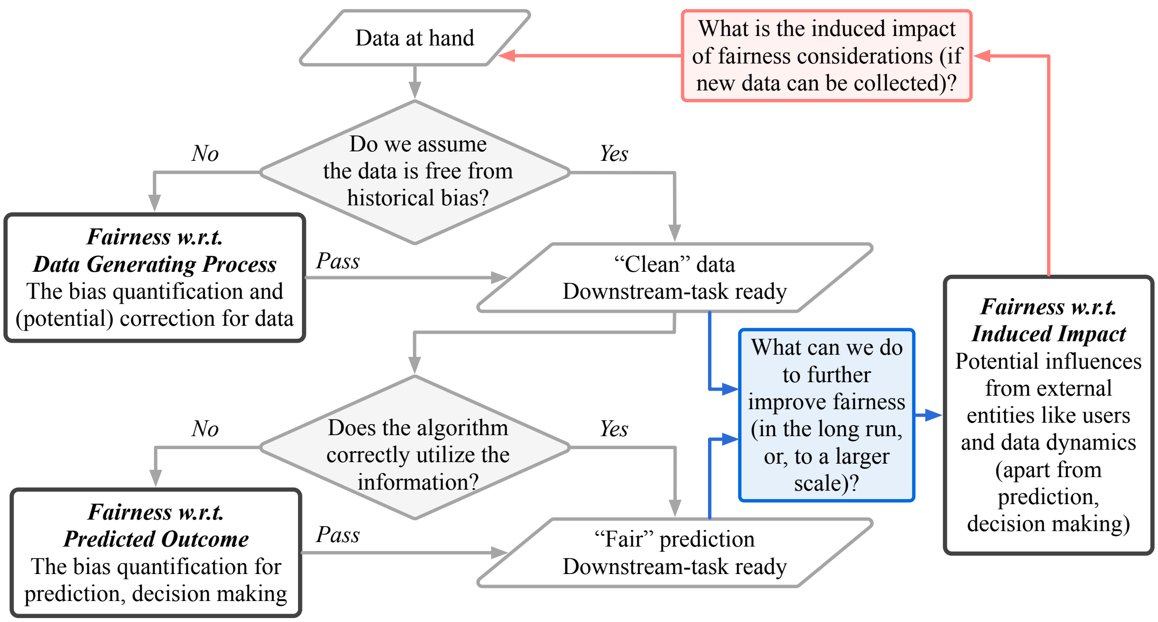


Figure 1: The flowchart that illustrates the road map to navigate through different spectrums of fairness inquiries. Starting from the data at hand, based on the answer to the questions, we can sequentially audit fairness with respect to the underlying data generating process, the predicted outcome itself, and the induced impact in the future, respectively. If conditions permit, 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.

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 spectrums.

6.2.1 Fairness w.r.t. Data Generating Process

As indicated by the name, the primary focus for this type of fairness inquiry is on the underlying data generating process. Multiple factors may contribute to the bias in the data (Danks and London, 2017; Mitchell et al., 2018; Mehrabi et al., 2021): the imperfection of previous human decisions, the lingering effect of historical discriminations, the (potentially) morally neutral statistical bias/error in the sampling and measurement, etc.

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 audit 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 (see, e.g., Kearns and Roth 2019), 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 the fairness audit with respect to data generating process (the first spectrum). By making the assumption that the data at hand is “clean”, we can lift the burden from the downstream tasks, and emphasize on 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 audit applies to predictive models as well as prediction-based decision-making systems. After all, human decision making also rests upon predictions to some extent (Mitchell et al., 2018). 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, which 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 audit with respect to the predicted outcome, auditing 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 audit 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), etc. 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, etc. In the music school example, Question 8 is a fairness audit with respect to the induced impact (of deploying a decision-making system).

6.3 Remark: The Necessity of Considering Different Fairness Spectrums

In this section, we have seen different spectrums 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 the technical definitions and the object of interest (e.g., Y vs. $\widehat{Y}$).

Technically speaking, enforcing prediction fairness 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 distinguish between fairness considerations for each one of them not only offers conceptual clarify, but also provides a clearer picture regarding what kind of fairness inquiry one is actually conducting.

Furthermore, there is no guarantee that fair data can result in fair prediction, or vise versa. On one hand, even if the data is "clean", the not-so-careful utilization of the data for prediction may still introduce new discriminations, for instance, the theoretical unattainability of fairness with a particular prediction scheme (Tang and Zhang, 2022), and the introduced unfairness in prediction even if the label is fair (Ashurst et al., 2022). It is not necessarily the case that the prediction bias results only from data bias (when data is clean, the prediction can still be unfair). On the other hand, as we have seen in Section 5, ample evidence has suggested that "fair" predictions can have adverse impact on 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.

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*, on the other hand, emphasizes on the side of user autonomy and/or data dynamics as well as (possibly) other 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 music school committee.

7. Subtlety: The Role of Causality in Fairness Analysis

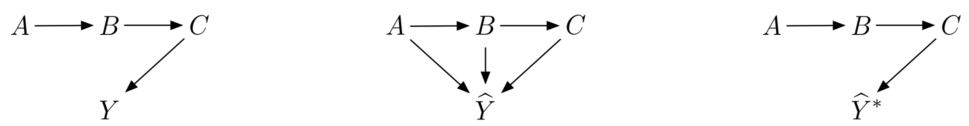
In Section 3 we presented multiple static fairness notions in the literature, many of which leverage the power of causal reasoning. Before discussing the exact location where the notions might fit in the fairness spectrums presented in Section 6, we believe it is necessary and important to reflect on 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 before applying them to fairness inquires from certain spectrums. 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 (Kilbertus et al., 2017; Kusner et al., 2017; Russell et al., 2017; Zhang et al., 2017c; Nabi and Shpitser, 2018; Zhang and Bareinboim, 2018; Loftus et al., 2018; Chiappa, 2019; Wu et al., 2019). 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 2a: the gender is a cause of the height; the height determines the

9. For the tasks like prediction, the output $\widehat{Y}$ is usually generated by a classification or regression algorithm in the literature.



(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 $\widehat{Y}^*$

Figure 2: The 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 $\widehat{Y}^*$.

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), we will have a data generating process as shown in Figure 2b. The reason of the extra arrows in Figure 2b compared to Figure 2a 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 2a. However, if for example we use a generative model and perform an probabilistic inference task on the outcome where we follow the underlying data generating process, 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 $\widehat{Y}^*$).[11] The data generating process for the prediction result via inference $\widehat{Y}^*$ (Figure 2c) is only different (in terms of the causal graph) from its counterpart for the ground truth variable Y (Figure 2a) up to a substitution of the outcome variable.

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, the result of discrimination quantification could be unpredictable and therefore is hardly justifiable.

7.2 Causal Modeling with the Intended Semantics

When we mention causal modeling in the fairness analysis, there are several potential candidates in terms of which causal model we are referring to. For instance, a causal model that describes "the data generating process for the ground truth" could be interpreted with multiple different semantics: the causal model recovered from the data at hand through causal discovery (i.e., the model that corresponds to the current reality), the external knowledge from experts or just an assumption on the data generating process (i.e., the model that we assume with professional knowledge to be able

10. This is a simplified model with a very limited number of variables involved for the purpose of illustration.
11. Usually, we need stronger assumptions regarding the underlying data generating process in order to perform the inference tasks, e.g., the availability of a structural equation model (SCM) instead of only the causal graph.

to describe the current reality), the relations among variables in the hypothetical ideal world where there is no discrimination (i.e., the model that corresponds to the ideal reality which may not be the case for the current reality), and so on.

Among these various possible interpretations, it is not always self-explainable from the fairness notions themselves regarding which interpretation really corresponds to the causal model presented to us, if without further clarifications. Therefore in practice, we should not only keep in mind the intuition behind the fairness notions, but also make sure that the semantics of the causal model we are using truly matches the type of the intended task.

In the basketball player example, if we were to quantify the discrimination hidden within the current data, we need to refer to Figure 2a and at the same time make sure that it reflects relations among variables within the current data at hand. If we decide that there is historical discrimination in the data and we want to see how the distribution of Y would have been like were there no discrimination, we need to refer to a graph that corresponds to a hypothetical ideal world (which may not necessarily be the same as Figure 2a).

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.

8. Enforce Fairness in Different Spectrums

In Section 6 and Section 7 we have seen different spectrums 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 spectrums. In Section 8.1, we propose a flowchart corresponding to our fairness audit categorization. Then, in Section 8.2 - Section 8.4, we revisit commonly used fairness notions (reviewed in Section 4), illustrating how they fit in the fairness spectrums (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 (and possibly **correct** the discrimination-contaminated 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.

8.1 The Algorithmic Fairness Flowchart: Answering "How-To" Questions

In Figure 1 we present a road map to navigate through fairness from different spectrums. Start from the very beginning, the input for *Fairness w.r.t. Data Generating Process* type of audits 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, on the other hand, 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 flow in Figure 1). After going through the analysis through different types of fairness emphases, if conditions permit the 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 flow in Figure 1) and conduct a closed-loop fairness analysis.

In the rest of this section, we discuss the fairness notions (with potential necessary modifications) we should use to achieve distinct fairness goals.

8.2 Fairness w.r.t. Data Generating Process

As indicated by the name, fairness audit 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 measured variables in terms of the underlying data generating process, which makes causal modeling a perfect tool to achieve the goal. In this section, 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 (Zhang et al., 2016; Kilbertus et al., 2017; Kusner et al., 2017; Zhang et al., 2017c; Nabi and Shpitser, 2018; Zhang and Bareinboim, 2018; Khademi et al., 2019; Chiappa, 2019; Wu et al., 2019; Salimi et al., 2019). 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 semantic 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* (Definition 4 and Definition 5), *Counterfactual Fairness* (Definition 8), and *Path-specific Counterfactual Fairness* (Definition 9) that we reviewed in Section 4:

Definition 10 ***No Direct Discrimination (modified)*** *Given the causal graph that describes the data generating process of the current reality, let us denote π_d as 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 11 ***No Indirect Discrimination (modified)*** *Given the causal graph that describes the data generating process of the current reality, let us denote π_i as 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 12 ***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 $V := \{A, X\}$, 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 13 ***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 $V := \{A, X\}$ 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):*

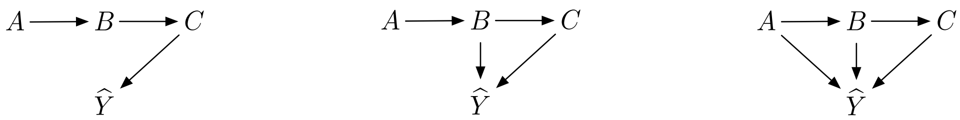
$$P\big(Y_{A\leftarrow a'|\pi, A\leftarrow a|\bar{\pi}}(U) = y \mid O = o\big) - P\big(Y_{A\leftarrow a}(U) = y \mid O = o\big) = 0. \tag{13}$$

Compared to the original notions (Definitions 4, 5, 8, 9), the modified causal notions (Definitions 10, 11, 12, 13) are quantifying discrimination with respect to the outcome variable Y instead of the prediction $\widehat{Y}$, using the data generating process behind Y with respect to the current reality. This seemingly trivial modification is more than just exchanges of variables. In practical applications, when we assume the availability (either via a 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 $\widehat{Y}$ (Chiappa and Isaac, 2018; Nabi and Shpitser, 2018).

Furthermore, even if we can draw the causal graph for predictions as illustrated in Figure 2b (for prediction via classification/regression) and Figure 2c (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 8, which is the original notion proposed by Kusner et al. 2017) 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 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 2a. Considering the task is to audit fairness on $\widehat{Y}$, the practitioner directly exchanges the variable Y in the graph to $\widehat{Y}$



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

Figure 3: The comparison between graphs that the practitioner draws in different strategies.

and draws the graph shown in Figure 3a. The practitioner then proceeds to the fairness audit via *Counterfactual Fairness* (Definition 8) without knowing the detail regarding how $\widehat{Y}$ is computed (which is the output of a regressor).

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 2c).

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 scored by the player Y. Then the practitioner realizes that the task is to audit fairness on $\widehat{Y}$ and draws the graph shown in Figure 3b. The practitioner proceeds to the fairness audit via *Counterfactual Fairness* (Definition 8) without knowing the detail regarding how $\widehat{Y}$ is computed (which is the output of a regressor).

Strategy (4) The practitioner notices that $\widehat{Y}$ is the output of a regression algorithm and draws the causal graph that corresponds the data generating process of $\widehat{Y}$ as shown in Figure 3c. The practitioner then proceeds to the fairness audit via *Counterfactual Fairness* (Definition 8) with respect to $\widehat{Y}$ (which is the output of a regressor).

Let us take a closer look to 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 2b, instead of the one shown in Figure 3a.

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 3a happens to be identical to Figure 2c (except for the asterisk symbol in Figure 2c). 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 semantics (using the graph which 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 3c is identical to Figure 2b. 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 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 (Kilbertus et al., 2017; Kusner et al., 2017; Zhang et al., 2017c; Chiappa, 2019; Wu et al., 2019). 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 – **detect** discrimination within the current data and enforce *Fairness w.r.t. Data Generating Process*.

Admittedly, the detection of existence of the discrimination in the data does not easily translate into potential 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 through 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 is 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 focuses on 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 processing 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* (Dwork et al., 2012), *Demographic Parity* (Calders et al., 2009), *Equalized Odds* (Hardt et al., 2016), are most suitable for the intended fairness audits in this category.

In the algorithmic fairness literature, the phenomenon of the "tradeoff between fairness and accuracy" for the prediction has been widely observed and discussed (Kamiran and Calders, 2012; Romei and Ruggieri, 2014; Feldman et al., 2015; Chouldechova, 2017; Berk et al., 2017; Corbett-Davies et al., 2017; Kleinberg et al., 2017; Menon and Williamson, 2018; Agarwal et al., 2018; Mary et al., 2019; Wick et al., 2019; Baharlouei et al., 2020). 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 audit categories, 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 (other than prediction/decision-making).

As we have discussed in Section 6, we cannot put audits from the *Fairness w.r.t. Induced Impact* 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, 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) separate 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 the measured variables in the data set at hand. Deeply rooted socioeconomic attributes are often not readily available for us when we audit fairness.

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 (Liu et al., 2018; Heidari et al., 2019b; Zhang et al., 2020). We need to disentangle not only different types of discriminations (in terms of different fairness hierarchies), but also exploit the efforts devoted by all involved parties, so that fairness can be further improved.

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 spectrums 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 1) is intended to enable a closed-loop fairness analysis by navigating through different spectrums 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 would like to 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. In order 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 1) and explore the possibility of inducing a fair data generating process in the future by conducting a closed-loop fairness analysis and analyzing *Fairness w.r.t. Predicted Outcome* and *Fairness w.r.t. Induced Impact* at the same time.

9. Conclusion

In this paper, 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 inquiries.

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, one would like to explore the possibility of further improving fairness through the effort of external entities beyond prediction/decision-making.

Future research directions naturally span across different spectrums 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 fairness pursuit with respect to the underlying data generating process, especially for the potential correction (going beyond detection) of the data to eliminate discriminations within the data in the long-term, dynamic fairness pursuit. 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 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 1) also highlights the potential to quantify the effectiveness of fairness pursuits of the current iteration through another round of fairness audits (e.g., the red flow in Figure 1). With meaningful interpretations of the result, the findings from multiple fairness spectrums 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

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

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

Elizabeth Anderson. The imperative of integration. In *The Imperative of Integration*. Princeton University Press, 2010.

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

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

Aristotle. *Aristotle: Nicomachean Ethics (Trans., Roger Crisp)*. Cambridge University Press, 2000.

Richard J Arneson. Equality and equal opportunity for welfare. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 56(1):77–93, 1989.

Kenneth Arrow. The theory of discrimination. Technical report, Princeton University, Department of Economics, Industrial Relations Section., 1971.

Carolyn Ashurst, Ryan Carey, Silvia Chiappa, and Tom Everitt. Why fair labels can yield unfair predictions: Graphical conditions for introduced unfairness. *arXiv preprint arXiv:2202.10816*, 2022.

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

Pranjal Awasthi, Alex Beutel, Matthäus Kleindessner, Jamie Morgenstern, and Xuezhi Wang. Evaluating fairness of machine learning models under uncertain and incomplete information. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 206–214, 2021.

Sina Baharlouei, Maher Nouiehed, Ahmad Beirami, and Meisam Razaviyayn. Rényi fair inference. In *International Conference on Learning Representations*, 2020.

Solon Barocas and Andrew D Selbst. Big data's disparate impact. *CALIFORNIA LAW REVIEW*, pages 671–732, 2016.

Solon Barocas, Andrew D Selbst, and Manish Raghavan. The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 80–89, 2020.

Yahav Bechavod, Katrina Ligett, Aaron Roth, Bo Waggoner, and Steven Z Wu. Equal opportunity in online classification with partial feedback. *Advances in Neural Information Processing Systems*, 32, 2019.

Yahav Bechavod, Christopher Jung, and Steven Z Wu. Metric-free individual fairness in online learning. *Advances in neural information processing systems*, 33:11214–11225, 2020.

Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*, 2017.

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021.

Richard A Berk and Justin Bleich. Statistical procedures for forecasting criminal behavior: A comparative assessment. *Criminology & Pub. Pol'y*, 12:513, 2013.

Reuben Binns. Fairness in machine learning: Lessons from political philosophy. In *Conference on Fairness, Accountability and Transparency*, pages 149–159. PMLR, 2018.

Reuben Binns. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 514–524, 2020.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357, 2016.

Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18. IEEE, 2009.

Simon Caton and Christian Haas. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*, 2020.

Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 339–348, 2019.

Silvia Chiappa. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7801–7808, 2019.

Silvia Chiappa and William S Isaac. A causal bayesian networks viewpoint on fairness. In *IFIP International Summer School on Privacy and Identity Management*, pages 3–20. Springer, 2018.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

Houston Claure, Yifang Chen, Jignesh Modi, Malte Jung, and Stefanos Nikolaidis. Multi-armed bandits with fairness constraints for distributing resources to human teammates. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 299–308, 2020.

Stephen Coate and Glenn C Loury. Will affirmative-action policies eliminate negative stereotypes? *The American Economic Review*, pages 1220–1240, 1993.

GA Cohen, S de Wijze, MH Kramer, and I Carter. Fairness and legitimacy in justice, and: does option luck ever preserve justice. *Hillel Steiner and the Anatomy of Justice: Themes and Challenges*, 16: 1, 2009.

Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806, 2017.

Elliot Creager, David Madras, Joern-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In *International Conference on Machine Learning*, pages 1436–1445, 2019.

Elliot Creager, David Madras, Toniann Pitassi, and Richard Zemel. Causal modeling for fairness in dynamical systems. In *International Conference on Machine Learning*, pages 2185–2195. PMLR, 2020.

Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D Sculley, and Yoni Halpern. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 525–534, 2020.

David Danks and Alex John London. Algorithmic bias in autonomous systems. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4691–4697, 2017.

Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *Proceedings on privacy enhancing technologies*, 2015(1):92–112, 2015.

William Dieterich, Christina Mendoza, and Tim Brennan. Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc*, 2016.

Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 55–70, 2018.

Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, pages 2791–2801, 2018.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on Fairness, Accountability and Transparency*, pages 119–133, 2018.

Ronald Dworkin. *Sovereign virtue: The theory and practice of equality*. Harvard university press, 2002.

Jon Elster. *Local justice: How institutions allocate scarce goods and necessary burdens*. Russell Sage Foundation, 1992.

Hadi Elzayn, Shahin Jabbari, Christopher Jung, Michael Kearns, Seth Neel, Aaron Roth, and Zachary Schutzman. Fair algorithms for learning in allocation problems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 170–179, 2019.

Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. Decision making with limited feedback: Error bounds for recidivism prediction and predictive policing. *Proceedings of FAT/ML 2017*, 2017.

Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. Runaway feedback loops in predictive policing. In *Conference on Fairness, Accountability and Transparency*, pages 160–171. PMLR, 2018.

Ido Erev and Alvin E Roth. Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American economic review*, pages 848–881, 1998.

Andrew Estornell, Sanmay Das, Yang Liu, and Yevgeniy Vorobeychik. Unfairness despite awareness: Group-fair classification with strategic agents. *arXiv preprint arXiv:2112.02746*, 2021.

Sina Fazelpour and Zachary C Lipton. Algorithmic fairness from a non-ideal perspective. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 57–63, 2020.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

Benjamin Fish, Jeremy Kun, and Ádám D Lelkes. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 144–152. SIAM, 2016.

Dean P Foster and Rakesh V Vohra. An economic argument for affirmative action. *Rationality and Society*, 4(2):176–188, 1992.

Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236*, 2016.

David Gauthier. *Morals by agreement*. clarendon Press, 1987.

Yingqiang Ge, Shuchang Liu, Ruoyuan Gao, Yikun Xian, Yunqi Li, Xiangyu Zhao, Changhua Pei, Fei Sun, Junfeng Ge, Wenwu Ou, and Yongfeng Zhang. Towards long-term fairness in recommendation. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 445–453, 2021.

Stephen Gillen, Christopher Jung, Michael Kearns, and Aaron Roth. Online learning with an unknown fairness metric. *Advances in neural information processing systems*, 31, 2018.

Talia B Gillis and Jann L Spiess. Big data and discrimination. *The University of Chicago Law Review*, 86(2):459–488, 2019.

Bruce Glymour and Jonathan Herington. 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*, pages 269–278, 2019.

Barry Goldman and Russell Cropanzano. "justice" and "fairness" are not the same thing. *Journal of Organizational Behavior*, 36(2):313–318, 2015.

Swati Gupta and Vijay Kamble. Individual fairness in hindsight. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 805–806, 2019.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938. PMLR, 2018.

Hoda Heidari and Jon Kleinberg. Allocating opportunities in a dynamic model of intergenerational mobility. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 15–25, 2021.

Hoda Heidari and Andreas Krause. Preventing disparate treatment in sequential decision making. In *IJCAI*, pages 2248–2254, 2018.

Hoda Heidari, Claudio Ferrari, Krishna P Gummadi, and Andreas Krause. Fairness behind a veil of ignorance: a welfare analysis for automated decision making. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 1273–1283, 2018.

Hoda Heidari, Michele Loi, Krishna P Gummadi, and Andreas Krause. A moral framework for understanding fair ml through economic models of equality of opportunity. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 181–190, 2019a.

Hoda Heidari, Vedant Nanda, and Krishna Gummadi. On the long-term impact of algorithmic decision policies: Effort unfairness and feature segregation through social learning. In *36th International Conference on Machine Learning*, pages 2692–2701, 2019b.

Lily Hu and Yiling Chen. A short-term intervention for long-term fairness in the labor market. In *Proceedings of the 2018 World Wide Web Conference*, pages 1389–1398, 2018.

Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 259–268, 2019.

Wen Huang, Yongkai Wu, Lu Zhang, and Xintao Wu. Fairness through equality of effort. In *Companion Proceedings of the Web Conference 2020*, pages 743–751, 2020.

Yimin Huang and Marco Valtorta. Identifiability in causal bayesian networks: A sound and complete algorithm. In *AAAI*, pages 1149–1154, 2006.

Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in reinforcement learning. In *International conference on machine learning*, pages 1617–1626. PMLR, 2017.

Taeuk Jang, Feng Zheng, and Xiaoqian Wang. Constructing a fair classifier with generated fair data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7908–7916, 2021.

Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. Fair algorithms for infinite and contextual bandits. *arXiv preprint arXiv:1610.09559*, 2016a.

Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. *Advances in neural information processing systems*, 29, 2016b.

Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. Towards realistic individual recourse and actionable explanations in black-box decision making systems. *arXiv preprint arXiv:1907.09615*, 2019.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

Faisal Kamiran, Indrė Žliobaitė, and Toon Calders. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and information systems*, 35(3):613–644, 2013.

Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 643–650. IEEE, 2011.

Sampath Kannan, Aaron Roth, and Juba Ziani. Downstream effects of affirmative action. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 240–248, 2019.

Michael Kearns and Aaron Roth. *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press, 2019.

Michael Kearns, Aaron Roth, and Zhiwei Steven Wu. Meritocratic fairness for cross-population selection. In *International Conference on Machine Learning*, pages 1828–1836. PMLR, 2017.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2564–2572. PMLR, 2018.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 100–109, 2019.

Aria Khademi, Sanghack Lee, David Foley, and Vasant Honavar. Fairness in algorithmic decision making: An excursion through the lens of causality. In *The World Wide Web Conference*, pages 2907–2914, 2019.

Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 656–666, 2017.

Young-Chul Kim and Glenn C Loury. Collective reputation and the dynamics of statistical discrimination. *International Economic Review*, 59(1):3–18, 2018.

Jon Kleinberg and Manish Raghavan. How do classifiers induce agents to invest effort strategically? *ACM Transactions on Economics and Computation (TEAC)*, 8(4):1–23, 2020.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Cass R Sunstein. Discrimination in the age of algorithms. *Journal of Legal Analysis*, 10:113–174, 2018.

Carl Knight and Zofia Stemplowska. *Responsibility and distributive justice*. Oxford University Press, 2011.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076, 2017.

Jonathan Levin et al. The dynamics of collective reputation. *The BE Journal of Theoretical Economics*, 9(1):1–25, 2009.

Danielle Li, Lindsey R Raymond, and Peter Bergman. Hiring as exploration. Technical report, National Bureau of Economic Research, 2020.

Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pages 3150–3158. PMLR, 2018.

Lydia T Liu, Max Simchowitz, and Moritz Hardt. The implicit fairness criterion of unconstrained learning. In *International Conference on Machine Learning*, pages 4051–4060. PMLR, 2019.

Lydia T Liu, Ashia Wilson, Nika Haghtalab, Adam Tauman Kalai, Christian Borgs, and Jennifer Chayes. The disparate equilibria of algorithmic decision making when individuals invest rationally. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 381–391, 2020.

Yang Liu, Goran Radanovic, Christos Dimitrakakis, Debmalya Mandal, and David C Parkes. Calibrated fairness in bandits. *arXiv preprint arXiv:1707.01875*, 2017.

Yang Liu, Yatong Chen, Zeyu Tang, and Kun Zhang. Induced domain adaptation. *arXiv preprint arXiv:2107.05911*, 2021.

Joshua R Loftus, Chris Russell, Matt J Kusner, and Ricardo Silva. Causal reasoning for algorithmic fairness. *arXiv preprint arXiv:1805.05859*, 2018.

Kristian Lum and William Isaac. To predict and serve? *Significance*, 13(5):14–19, 2016.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309*, 2018.

Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. Survey on causal-based machine learning fairness notions. *arXiv preprint arXiv:2010.09553*, 2020.

Jérémie Mary, Clément Calauzenes, and Noureddine El Karoui. Fairness-aware learning for continuous attributes and treatments. In *International Conference on Machine Learning*, pages 4382–4391, 2019.

Andrew Mason. *Levelling the playing field: The idea of equal opportunity and its place in egalitarian thought*. OUP Oxford, 2006.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

Aditya Krishna Menon and Robert C Williamson. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*, pages 107–118, 2018.

David Miller. *Principles of social justice*. harvard university Press, 2001.

David Miller. Justice. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2021 edition, 2021.

Smitha Milli, John Miller, Anca D Dragan, and Moritz Hardt. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 230–239, 2019.

Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867*, 2018.

George O Mohler, Martin B Short, Sean Malinowski, Mark Johnson, George E Tita, Andrea L Bertozzi, and P Jeffrey Brantingham. Randomized controlled field trials of predictive policing. *Journal of the American statistical association*, 110(512):1399–1411, 2015.

Hussein Mouzannar, Mesrob I Ohannessian, and Nathan Srebro. From fair decision making to social equality. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 359–368, 2019.

Debarghya Mukherjee, Mikhail Yurochkin, Moulinath Banerjee, and Yuekai Sun. Two simple ways to learn individual fairness metrics from data. In *International Conference on Machine Learning*, pages 7097–7107. PMLR, 2020.

Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Razieh Nabi, Daniel Malinsky, and Ilya Shpitser. Learning optimal fair policies. In *International Conference on Machine Learning*, pages 4674–4682. PMLR, 2019a.

Razieh Nabi, Daniel Malinsky, and Ilya Shpitser. Optimal training of fair predictive models. *arXiv preprint arXiv:1910.04109*, 2019b.

Thomas Nagel. The problem of global justice. *Philosophy & public affairs*, 33(2):113–147, 2005.

Arvind Narayanan. Translation tutorial: 21 fairness definitions and their politics. In *Proc. Conf. Fairness Accountability Transp., New York, USA*, volume 1170, page 3, 2018.

Vishakha Patil, Ganesh Ghalme, Vineet Nair, and Y Narahari. Achieving fairness in the stochastic multi-armed bandit problem. *Journal of Machine Learning Research*, 22:1–31, 2021.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Robin Pemantle. A survey of random processes with reinforcement. *Probability surveys*, 4:1–79, 2007.

Adrián Pérez-Suay, Valero Laparra, Gonzalo Mateo-García, Jordi Muñoz-Marí, Luis Gómez-Chova, and Gustau Camps-Valls. Fair kernel learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 339–355. Springer, 2017.

Walt L Perry. *Predictive policing: The role of crime forecasting in law enforcement operations*. Rand Corporation, 2013.

Edmund S Phelps. The statistical theory of racism and sexism. *The american economic review*, 62 (4):659–661, 1972.

Reilly Raab and Yang Liu. Unintended selection: Persistent qualification rate disparities and interventions. *Advances in Neural Information Processing Systems*, 34, 2021.

John Rawls. *A theory of justice*. Cambridge: Harvard University Press, 1971.

John Rawls. *Justice as fairness: A restatement*. Harvard University Press, 2001.

Ashkan Rezaei, Anqi Liu, Omid Memarrast, and Brian D Ziebart. Robust fairness under covariate shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9419–9427, 2021.

Yaniv Romano, Stephen Bates, and Emmanuel J Candès. Achieving equalized odds by resampling sensitive attributes. *arXiv preprint arXiv:2006.04292*, 2020.

Andrea Romei and Salvatore Ruggieri. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, 29(5):582–638, 2014.

Chris Russell, Matt J Kusner, Joshua Loftus, and Ricardo Silva. When worlds collide: integrating different counterfactual assumptions in fairness. In *Advances in Neural Information Processing Systems*, pages 6414–6423, 2017.

Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. Interventional fairness: Causal database repair for algorithmic fairness. In *Proceedings of the 2019 International Conference on Management of Data*, pages 793–810, 2019.

Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C Parkes, and Yang Liu. How do fairness definitions fare? examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 99–106, 2019.

Thomas Scanlon. *What we owe to each other*. Belknap Press, 2000.

Candice Schumann, Xuezhi Wang, Alex Beutel, Jilin Chen, Hai Qian, and Ed H Chi. Transfer of machine learning fairness across domains. *arXiv preprint arXiv:1906.09688*, 2019.

Saeed Sharifi-Malvajerdi, Michael Kearns, and Aaron Roth. Average individual fairness: Algorithms, generalization and experiments. *Advances in Neural Information Processing Systems*, 32, 2019.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

Ilya Shpitser and Judea Pearl. Identification of conditional interventional distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 437–444, 2006.

Ilya Shpitser and Judea Pearl. What counterfactuals can be tested. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, pages 352–359, 2007.

Ilya Shpitser and Judea Pearl. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9:1941–1979, 2008.

Umer Siddique, Paul Weng, and Matthieu Zimmer. Learning fair policies in multi-objective (deep) reinforcement learning with average and discounted rewards. In *International Conference on Machine Learning*, pages 8905–8915. PMLR, 2020.

Ashudeep Singh and Thorsten Joachims. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2219–2228, 2018.

Harvineet Singh, Rina Singh, Vishwali Mhasawade, and Rumi Chunara. Fairness violations and mitigation under covariate shift. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 3–13, 2021.

Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. Learning controllable fair representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2164–2173, 2019.

Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. A unified approach to quantifying algorithmic unfairness: Measuring individual &group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2239–2248, 2018.

Michael Spence. *Market signaling: Informational transfer in hiring and related screening processes*. Number 143. cambridge: harvard university press, 1974.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. Springer New York, 1993.

Peter Spirtes, Christopher Meek, and Thomas Richardson. Causal inference in the presence of latent variables and selection bias. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 499–506, 1995.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Latanya Sweeney. Discrimination in online ad delivery. *Queue*, 11(3):10–29, 2013.

Wei Tang, Chien-Ju Ho, and Yang Liu. Bandit learning with delayed impact of actions. *Advances in Neural Information Processing Systems*, 34, 2021.

Zeyu Tang and Kun Zhang. Attainability and optimality: The equalized odds fairness revisited. *arXiv preprint arXiv:2202.11853*, 2022.

Jin Tian and Judea Pearl. A general identification condition for causal effects. In *Aaai/iaai*, pages 567–573, 2002.

Jean Tirole. A theory of collective reputations (with applications to the persistence of corruption and to firm quality). *The Review of Economic Studies*, 63(1):1–22, 1996.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pages 1–5. IEEE, 2015.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 10–19, 2019.

Tyler J VanderWeele and Whitney R Robinson. On causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiology (Cambridge, Mass.)*, 25(4):473, 2014.

Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*, pages 1–7. IEEE, 2018.

J. von Kügelgen, A.-H. Karimi, U. Bhatt, I. Valera, A. Weller, and B. Schölkopf. On the fairness of causal algorithmic recourse. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

Lequn Wang, Yiwei Bai, Wen Sun, and Thorsten Joachims. Fairness of exposure in stochastic bandits. In *International Conference on Machine Learning*, pages 10686–10696. PMLR, 2021.

Serena Wang, Wenshuo Guo, Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, and Michael Jordan. Robust optimization for fairness with noisy protected groups. *Advances in Neural Information Processing Systems*, 33:5190–5203, 2020.

Min Wen, Osbert Bastani, and Ufuk Topcu. Algorithms for fairness in sequential decision making. In *International Conference on Artificial Intelligence and Statistics*, pages 1144–1152. PMLR, 2021.

John A Weymark. Generalized gini inequality indices. *Mathematical Social Sciences*, 1(4):409–430, 1981.

Michael Wick, Jean-Baptiste Tristan, et al. Unlocking fairness: a trade-off revisited. In *Advances in Neural Information Processing Systems*, pages 8780–8789, 2019.

Jason A Winfree and Jill J McCluskey. Collective reputation and quality. *american Journal of agricultural Economics*, 87(1):206–213, 2005.

Yongkai Wu, Lu Zhang, Xintao Wu, and Hanghang Tong. Pc-fairness: A unified framework for measuring causality-based fairness. In *Advances in Neural Information Processing Systems*, pages 3399–3409, 2019.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180, 2017a.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970, 2017b.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

Junzhe Zhang and Elias Bareinboim. Fairness in decision-making—the causal explanation formula. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Kun Zhang and Aapo Hyvärinen. On the identifiability of the post-nonlinear causal model. In *25th Conference on Uncertainty in Artificial Intelligence (UAI 2009)*, pages 647–655. AUAI Press, 2009.

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

Kun Zhang, Biwei Huang, Jiji Zhang, Clark Glymour, and Bernhard Schölkopf. Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 1347–1353, 2017a.

Lu Zhang and Xintao Wu. Anti-discrimination learning: a causal modeling-based framework. *International Journal of Data Science and Analytics*, 4(1):1–16, 2017.

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

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

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

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

Xueru Zhang, Mohammadmahdi Khaliligarekani, Cem Tekin, et al. Group retention when using machine learning in sequential decision making: the interplay between user dynamics and fairness. *Advances in Neural Information Processing Systems*, 32:15269–15278, 2019.

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

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

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