Differential Expressiveness: A Data-Centered Perspective on Algorithmic Bias

an undergraduate thesis presented
by
Eric Meng Shen
to
The Department of Computer Science

in partial fulfillment of the requirements
for the degree of
Bachelor of Arts
in the subject of
Computer Science and Mathematics

Harvard University
Cambridge, Massachusetts
April 2024

Thesis advisor: Professor Cynthia Dwork Eric Meng Shen

Differential Expressiveness: A Data-Centered Perspective on Algorithmic Bias

Abstract

The interdisciplinary study of algorithmic fairness and bias has enjoyed a meteoric rise in popularity over the past several years, motivated in no small part by the increasingly influential impact of machine learning in many aspects of daily life. One part of this field examines the foundational issue of bias being present in the training data that is provided to an algorithm, seeking to develop ways to describe and mitigate this issue.

We propose a new and broad characterization of a kind of data bias that we call differential expressiveness (DE). We formulate DE as being quality of an individual feature in a dataset, conveying a condition where the values of the feature cannot be consistently interpreted across different individuals. Contextualizing our presentation with an overview of the development of algorithmic fairness, we give two mathematical interpretations of DE and explore how the interpretations relate to one another. In addition, we discuss a variety of case studies illustrating how we can use DE to interpret data bias in real-world examples. Finally, we explore how DE complements existing frameworks in the literature for modeling data bias.

Contents

Acknowledgments

I am gratefully indebted to my thesis (and concentration!) advisor, Professor Cynthia Dwork, for her indispensable help and guidance throughout the thesis-writing process, as well as suggesting the initial idea that developed into the topic of this thesis. The depth of her knowledge, her openness to discussion, and her eagerness to share interdisciplinary perspectives have not been lost on me, and our conversations surrounding topics relevant to this thesis and otherwise have created what has truly been an intellectually enriching experience during my senior year. It has been a great privilege and absolute pleasure to work with and learn from one of the progenitors of algorithmic fairness, among several other groundbreaking fields in and outside computer science. Albeit brief, I have tremendously enjoyed the mentorship she has provided.

I owe thanks to Professor Seth Neel and Professor Stephen Chong, for their willingness to serve as thesis readers. Professor Neel's work in algorithmic fairness has been inspiring and stands as a distinguished example of efforts to engage with, evaluate, and push forward the field's conceptions of what fairness is and how we can achieve or guarantee it. I have had the good fortune of taking a course taught by Professor Stephen Chong, which, although not having concerned algorithmic fairness, enriched my recognition and respect for the complicated processes lying behind the creation and compilation of software we use everyday—comparable, in a way, to the deceiving definitional complexity that underlies our broad intuition of fairness. Professor Chong has also served as a supervisor for an Independent Study in the past, and his continual initiative to oversee my academic projects is duly appreciated.

I am also thankful to Professor Richard Weissbourd from the Harvard Graduate School of Education for sharing thoughts with me about equity and equality in education, which forms just a part of his grounded and impactful scholarship. Some of his takeaways are discussed in Part 2 of this thesis.

0

Introduction

0.1 A World of (Flawed) Data and Algorithms

In 2021, more than 26 million mortgage applications were made in the United States as reported by the Consumer Financial Protection Bureau[35]. That same year, despite the presence of COVID restrictions on educational activities, 1.5 million high school students took the standardized college admissions exam SAT,

as reported by the College Board, the organization which develops and administers the exam[24]. Finally, from October 1 of that year through September 30, 2022, data from the United States Sentencing Commission shows that some 64,142 individuals were federally sentenced across the country[99]. While describing completely different parts of life, each of the scenarios that these above numbers reflect—for instance, securing a mortgage to purchase a first home, completing a standardized test that plays a crucial role in college applications, and receiving a prison sentence—represent actions whose implications can irreversibly affect entire lives, families, and communities. The sheer magnitude of these numbers can make the cumulative impact of these decisions hard to comprehend.

It is therefore not unnatural to feel unease when considering another fact that these actions, like many others, all share in common: that they are increasingly becoming the domain of algorithms used in parts of the decision-making processes in the name of automation and efficiency. Several review papers and articles, including a report by the International Monetary Fund, have documented the increasing adoption of algorithms within finance[11,93]. Consumer banking has been no exception to this trend, where algorithms using techniques as diverse as decision trees and recurrent neural networks have been developed to assess the credit risk of loan applicants[95,54]. As we will later further discuss, risk assessment algorithms, trained using machine learning techniques, have been used to predict the "danger" that criminal defendants pose. One notable example is the *Correctional Offender Management Profiling for Alternative Sanctions*, or COMPAS, which offers a tool meant to measure the likelihood that an individual recidivates (re-commits crimi-

nal activity), which has been used as part of the sentencing process in several US states since 2001[5].

While the SAT itself is just a standardized exam with which the College Board has not used any algorithms (at least, they have not publicly disclosed any such efforts), we can view the exam as *itself* an algorithm translating an abstract feature such as "college academic preparedness" into a numeric score with which admissions committees can work. Insofar as we can make this interpretation, we have grounds to analyze and question the faithfulness of this translation. How accurately does a SAT score reflect academic preparedness, and are there particular kinds of students that the exam may treat unfairly, e.g. if scores are systemically biased to be higher or lower for those individuals? If so, how can we quantify this behavior? Of course, similar questions apply for the other examples we have mentioned. Each of these algorithms takes as input an individual represented as a set of chosen data fields, which we call *features* (e.g. age, race, income). Unscrupulous examination of these features can make these representations inaccurate and prejudiced. In general, we can, and arguably ought to, ask these questions given any algorithmic pipeline where humans are involved as inputs.

Undoubtedly, algorithms provide obvious benefits in decision-making. They can be deployed indefinitely, operate orders of magnitude more quickly, effectively scale with computational resources, and process volumes of data impractical for humans, which increases the availability and accessibility of services. Potentially, they can be *less* biased and more consistent judges than fallible and mercurial humans. Yet at the same time, in areas of application ranging from medicine

to advertising, they can unintentionally exhibit bias and compound existing inequalities and inequities in ways that might not be obvious, especially if their implementations remain private "black boxes"[81]. Especially considering the ever-growing ubiquity and utility of algorithms in virtually every part of life (c.f. the recent proliferation and performance of generative AI software), this dilemma is an extremely timely and consequential. This situation serves as an impetus and foundation for the burgeoning study of *algorithmic fairness*. An interdisciplinary subfield of computer science often inspired by fairness discussions in law, economics, and empirical case studies of "machine bias", its aim is mathematically characterizing and studying different notions of *algorithmic bias* and fairness, while developing methods to identify and remediate this bias and/or to achieve those notions of fairness. Work in the field ranges from theoretical analyses of machine learning techniques drawing from areas like statistics and computational learning theory to empirical investigations of deployed algorithms[71].

0.2 Contributions

Within algorithmic fairness, this thesis specifically focuses on the topic of biases which originate from data itself, i.e. *data bias*. These refer to issues innate to the often-massive datasets that are used to train algorithms: if this training data already exhibits historical biases, then an algorithm whose purpose is to recognize and learn patterns and relationships in the data will inherit these biases. For instance, algorithms for vetting loan applications may be excessively harsh toward low-income or minority applicants because they were trained on data including

instances where human decision-makers discriminated against those kinds of applicants in the past. Journalists have appropriately warned that training data is "never raw" but "cooked", and can be "dirty"[6,45]. This is a foundational, background concern: if data is unfair in the first place, then it is hard to expect that algorithms trained on it will not be. Importantly, this phenomenon exists independent of methodological considerations, such as sampling error when curating datasets: it tells us that even the best data we could hope to gather might be imperfect. It is an issue that reflects the world we live in, and is present before any downstream tasks occur, such as model training and deployment.

As the title of this thesis suggests, we will first propose a novel concept for a particular kind of data bias which we name *differential expressiveness* (DE), which can be applied to describe a large class of examples. From a high-level, DE can be summarized as a phenomenon whereby *a feature cannot be interpreted consistently to evaluate different individuals*. For instance, structural differences and norms may change what a certain value of that feature means between subpopulations of interest. We offer two broad descriptions of DE. Through one lens, DE is a product of injustices that exist in the world today, whereby contrary to our assumption, different groups of individuals posses different distributions of a feature. Through another lens, DE arises as a consequence of the feature itself signifying different things to individuals in different groups.

We will later discuss, in depth, instances of DE "in the wild", but here is as a preview a representative example. Consider the setting of college admissions in the US, wherein an admissions officer might look at a high school student's Advanced

Placement (AP) exam record, exams with registration fees that are cumulative tests for college-level courses, as a way of evaluating academic achievement. However, the financial cost of actually registering for AP exams might serve as a barrier for completing the exam for low-income students, even if they have already taken the corresponding course. A wealthier student enrolled in a set of AP courses could easily register for exams in all of them, whereas a poorer student taking the same courses (and doing equally well) may only want to register for one or two. Thus the same value for the number of exams a student has taken might mean very different things, as concerns academic achievement, based on the background of the student—a scenario where the student's background constitutes necessary context, exactly what our development of DE serves to characterize.

We also explore how the kind of bias represented by DE ties into extant frameworks of bias and fairness. So far, we have implicitly discussed two broad primary determinants of data bias. On one hand, the legacy of historical and societal discrimination makes the data we observe in the real world today problematic. In addition, however, when building algorithms practitioners often have to choose *proxy* features to represent the underlying attributes they wish to measure, such as using standardized exam scores as proxies for academic preparedness. Importantly, these determinants can be studied independent of whatever specific algorithm is trained on this data (making our discussion a more fundamental one), as they pertain to the data itself.

Our characterization of data bias will explore how we can link together these determinants. This treatment falls in the line of previous efforts to formally de-

scribe sources of harm in the machine learning pipeline. As a primer, we mention three representative examples here. Suresh and Guttag (2021) provide a framework for where sources of harm can appear within the "life cycle" of machine learning, from dataset curation to model deployment and finetuning[98]. Mhasawade et al. (2021) make a distinction between the world "as it should and could be" and the world "as it is", and attributing possible harmful perturbations present in the transition from the former to the latter as a product of societal biases[74]. Friedler et al. (2016) focus specifically on the idea of proxy features being a source for bias, differentiating the "construct space" of underlying features that an algorithm truly wants to measure with the "observed space" of features which can actually be measured[37]. This thesis extends the contributions of Friedler et al. to incorporate how DE fits into their framework.

The aim of this thesis is not to provide a comprehensive account of the field of algorithmic fairness, or fully unify existing definitions. (Indeed, we will find that several definitions of fairness are mutually incompatible.) Nor does it concentrate on developing solutions to detecting and amending data bias (although we will mention some literature concerning this). Instead, this thesis broadly examines the problem of individuals being unfairly represented through data to algorithms—a problem that is present before we even consider what the algorithm itself does—and proposes a new characterization, grounded by a diverse set of examples, to let us better understand it. The aim is to provide a new critical perspective on the data that practitioners work with everyday when creating algorithms.

This thesis is organized into the following parts. In Part 1 we give a selective

expository overview of the field of algorithmic fairness, discussing its interdisciplinary roots in distributive justice and law while also presenting some current notions of fairness and how they can conflict with each other. In Part 2 we tackle the issue of data bias. We formally present the idea of differential expressiveness, referencing various areas of application where biased algorithms have been or may be trained on differentially expressive data. We also discuss a mathematical framework for understanding differential expressiveness, adapted from Friedler et al., as the product of societal and historical inequality and the use of flawed proxy features. We will then briefly consider ideas for how differential expressiveness can be pragmatically addressed and how to confront the ubiquitous issue of data bias on the whole.

1

A Story of (Algorithmic) Fairness

In this chapter we give a selective overview of the development of the field of algorithmic fairness and how it stands today. The focus of this survey is not providing a comprehensive categorization of *all* the notions of fairness that have been introduced in the literature. Several survey papers have been written in recent years which attempt that goal. As some examples, Mehrabi et al. present a

broad overview of various notions of bias in the literature and, along with Alves et al., categorize various fairness definitions, of which we will mention a select few[71,3]. Wang et al. in their survey also overview causality-based fairness definitions[102]. Berk et al. focus on fairness definitions in the context of their application in criminal justice and sentencing, and summarize ways in which competing fairness notions may be mutually incompatible[12].

The goal of this review is to provide a more narrative account that specifically focuses on work that will be relevant in the examination of data bias in the following chapter. An increased focus is put on foundational motivations for the field of algorithmic fairness coming from statistics, law, economics, and political philosophy. The primary reason for this slant towards ideas outside of computer science proper is that they will prove rather relevant in our discussion of data bias. We give a "story" of the origins and state of algorithmic fairness: we will see that despite its core ideas having a long history, its modern incarnation is a particularly nascent and diverse field with remarkably close connections to contemporary problems, hosting a multitude of sometimes-incompatible concepts whose future developments are worth tracking.

1.1 Foundations of Fairness

Let us start from first principles, upon which we come across some core ontological questions. What is an "algorithm", and what does "fairness" mean in the context of the algorithm? Much of the work in algorithmic fairness works with definitions of algorithms that are kept as broad as possible so to maximize the generality of

their theories. We adopt this convention and define an algorithm to be any process that maps input data that represents an individual to an outcome, that is to say, implements a task that acts on, or "processes", representations of individuals. The implementation of this algorithm can be arbitrary and left a "black box", where the details of the model are unknown.[i] Thus, notions of fairness apply not only to machine learning methods like neural networks or unsupervised learning, but traditional statistical techniques and even human decision-making settings. The significant condition here is that the algorithm acts on representations of individuals who are distinct beings who can be grouped together.[ii] This condition allows us to reason about fairness in a meaningful way; after all, discussions about fairness outside of computer science revolve around the core question of how individual humans are treated.

The other ontological question—what it means for an algorithm to be fair—as it turns out, has no unifying answer. Several precise and reasonable technical definitions for fairness have been proposed, some of which conflict with each other. To start from a cleaner and more fundamental slate, we postpone an overview of these definitions to the next section, and examine fairness as perceived in the world of ethics. Fortunately, our understanding of an algorithm as a process that distributes outcomes to individuals based on their qualities lends itself almost ex-

[i] Some papers do add more specific conditions on what the outputs should look like, or the type of algorithm, which we will mention later.

[ii] In fact, the broadness of these definitions means that we could possibly consider algorithmic fairness over general "entities" like companies or nations rather than just individual humans, so long as they are separate, it is meaningful to group them together, and it makes sense through the theoretical lens of fairness to reason about equality and equity. To our knowledge, however, the literature always specifically considers tasks acting on humans.

actly to the setup of distributive justice and equality of opportunity. The study of these fields provides us with nuanced and principled theories of fairness, since those theories are crucial for their goal of providing moral guidance for the distribution of benefits and burdens in societies[65]. Thus fairness should be understood for us as an evaluation of the justness of algorithms as distributive mechanisms.

As an article by Ochigame states, the relationship between mathematical conceptions of fairness and the foundations of distributive justice runs further back than one might expect. The two concepts have historically been synergistic: Aristotle formulated distributive justice theories using geometry, and early developments in the theory of probability by mathematicians like Fermat and Pascal were motivated in part by efforts to create formal treatments of fair division and arbitration. In the late nineteenth century, the rise of the insurance industry led to the development of risk assessment tools, which ignited conversations around the fairness of those tools. Critiques then could rely on more straightforward argumentation, given that these risk "algorithms" practiced overt discrimination by charging differential rates based on race. Opponents to bills banning such practices cited the fact that black individuals at the time suffered from higher mortality rates, but proponents argued from an alternate lens: that insurance policies, instead of being modeled after "fatalistic statistics", should be designed to strive toward an imagined future without disparity, where minorities' life circumstances were equal[80].[iii] Risk models were applied to credit bureaus and policing in the

[iii] It is worth mentioning that the parallel development of statistics, somewhat comparably, is problematic. Prominent statisticians including Pearson and Fisher were motivated by theories of eugenics; this historical fact has prompted calls to critically reevaluate the function and interpretation of the statistical methods that they developed and that are still used today[22].

20th century, sparking controversies over the legitimacy of "statistical objectivity" to evaluate individuals.[iv] Thus the relationship between the development of perspectives of fairness and justice, the prevalence of discriminatory practices and areas of study, and the rise of quantitative scientific methods is enshrined in historical record. To discuss the development of one of these subjects necessitates a consideration of the others.

As controversies over credit scoring and criminal sentencing began to emerge, a concurrent movement in ethics and political philosophy starting in the 1970s introduced a radical new foundation for fairness. The philosopher John Rawls, arguably the founder of this project, proposed in his works *A Theory of Justice* and *Justice as Fairness* what remains one of the most-discussed theories of distributive justice today, resting on two claims culminating from a path of reasoning he believes people would follow if they were divorced from modern society in an "original position", under a "veil of ignorance"[105]. First, there is some set of inalienable basic rights and liberties that nobody should be denied and are collectively compatible. Second, societal inequalities may be tolerated insofar as they apply to positions that are accessible under a condition of equality of opportunity, and they benefit the least well-off in society (what he calls the difference principle)[84,85]. This formulation of fairness is decidedly not utilitarian, espousing a more liberal viewpoint. Yet it is also not completely egalitarian given its descriptions of justified inequality.

[iv] As another example: in the post-World War II era, academics such as von Neumann and Morgenstern were also instrumental in expanding the mathematical toolset available to policy in practice, introducing the novel application of mathematical optimization methods to human sciences and actuarial systems, incorporating concepts from game theory and statistics[80].

Rawls's work began an ongoing discussion on distributive justice. One prominent response was that of Ronald Dworkin, who considered a modification of Rawls's argument now called "luck egalitarianism". Dworkin critiqued Rawls for not properly treating criteria associated with individual responsibility that he viewed as categorically distinct from uncontrollable "endowments" given at birth. His overarching understanding of moral equality was not as necessitating equal treatment (equal ends) to all people, but rather treatment of all people as equals (equal consideration). In other words, while Rawls believed all characteristics of people are irrelevant when considering how to treat them, Dworkin disagreed and took agency to be a relevant factor. While people should start with equal access to resources, he argued, inequality is tolerable so long as it is attributable to differences in voluntary acts and decisions. This still leaves space to consider compensation schemes for those with unequal endowments (e.g. ill health)[41,88].[v]

Political scientists and economics later focused on applying the theories proposed above, regarding them through an quantitative lens. H. Peyton Young, in his book *Equity: In Theory and Practice*, considers "everyday" distributive problems, including taxation, emissions markets, and organ donation. He frames his definition of equity as a grounded translation of the social justice conceptions above: to be equitable is to create a practically just distribution scheme in real-

[v]The philosopher Robert Nozick went further in elevating the significance of personal responsibility. He proposed a libertarian viewpoint reminiscent of *laissez-faire* economics and Locke, doing away with distributive notions altogether and considering a state of the world to be just as long as individuals' holdings follow rules defining the just acquisition of property, thus being wholly concerned with processes rather than ends[78]. Beyond Dworkin and Nozick, there is no shortage of review articles covering the modern discourse over distributive justice, e.g. on the Stanford Encyclopedia of Philosophy[65,2].

world scenarios, which depends on scenario-specific objectives. Pointing out that "theories of justice in the large have little to say about what it means in the small" (that is, global ideals about distribution are too abstract to be useful in situations such as economists setting tax rates), his focus is on "compartmentalized" situations where concepts like the difference principle are too abstract and pragmatics like economic incentives become relevant[109].

The economist and political scientist John Roemer has focused specifically on developing the idea of equality of opportunity (as opposed to equality of outcomes) espoused by the various political theories introduced since Rawls. One of his main contributions is involving the language of economics to quantitatively formulate the level of *effort* exerted by individuals in addition to circumstances and how that can inform real-world policy, including what it means to "level the playing field"[87,88]. He adapts Dworkin's differentiation between controllable and uncontrollable attributes to describe an individual's effort and circumstances, which respectively denote the individual's own agency and arbitrary factors. Roemer believes that individuals should not be held responsible their types but should for the things they consciously do, i.e. their effort relative to their type.

Consider a desideratum u for which we wish to equalize opportunities as a function of circumstances C, effort e, and policy φ: $u(C, e, \varphi)$. Individuals are partitioned based on their circumstances by being given one of finitely many "types" t. A choice of policy induces a distribution of effort from the continuum of individuals having that type. For instance, u may be wage-earning capacity, C the family socioeconomic status, and φ a budget of how much to spend per

student; the distribution of effort of poorer students may be systemically different than that for richer students. His model's assumptions are thus that we can group individuals into different types, and that we should think of an individual's effort relativistically: "in deciding how hard a person has tried, we compare him only to others with his circumstances". To evaluate effort, then, he reasons in terms of quantiles rather than raw values. Let $v^t(\pi, \varphi)$ denote the value of u for individuals at quantile π of the effort distribution for type t under policy φ. Roemer's equal opportunity policy aims to maximize the sum, over all percentiles, of the minimum achievement of u among all types at that percentile:

$$\varphi^{EOp} = \arg\max_{\varphi} \int_0^1 \min_t v^t(\pi, \varphi)\, d\pi.$$

In contrast, letting the number of individuals of type t be n_t, he notes that a utilitarian policy would be $\arg\max_\varphi \sum_{t=1}^{\tau} n_t \int_0^1 v^t(\pi, \varphi)\, d\pi$, and a Rawlsian policy attending to the least well-off would be $\arg\max_\varphi \min_{t,\pi} v^t(\pi, \varphi)$.

To apply this formulation to practice and argue that defining circumstances and effort levels is practically tractable, Roemer has attempted to estimate $v^t(\pi, \varphi)$ from historical data. For instance, in the aforementioned setting of equalizing wage-earning capacity, based on economic data in the United States he perhaps unsurprisingly finds that φ^{EOp} would involve disproportionately higher levels of spending per capita on students of lower socioeconomic status, by a factor of about five. Furthermore, however, he finds that such a policy does not resolve racial inequality in wage earnings, and that an equal opportunity policy that also factors in race produces a φ^{EOp} that comparatively invests even more in black students, leading to larger variance in budget allocations[87].

This takeaway shows that equality of opportunity can recommend justified differential treatment (so even when relaxing the Rawlsian condition by holding individuals responsible for their efforts, we still get a principled argument in support of equity); it can even be used to identify and characterize intersectional concerns. Indeed, it provides an argument against rhetoric stipulating that equal opportunity policies ought to be "color-blind", e.g. in admissions; those "color-blind" policies can be suboptimal if the goal is achieving equitable outcomes with respect to race, since to tackle racism requires recognizing race. Roemer's conclusions were published more than 20 years before the Supreme Court's decision to strike down affirmative action programs, partly in the name of "color-blindness"[97].

We can view Roemer's contributions as a natural extension of the line of inquiry begun by Rawls and Dworkin. He factors in individual effort as a core part of his model, while accounting for the fact that effort itself might intrinsically differ among different types of individuals. Later in this chapter, we will see that the work of Caterina Calsamiglia extends this issue by considering how individual decisions regarding expending effort depend on access to resources[17]. Hence we have seen a gradual evolution in thinking about fairness in both principle and practice over a timespan of centuries, progressing from origins as old as statistics itself to the modern day where there is a continuing effort to quantitatively describe concepts of equality and equity. With this established, it is about time to look at algorithmic fairness proper.

1.2 Into the World of Algorithms

The next transition in our story of fairness is the jump from studying economic mechanisms to algorithmic processes. We can consider running an algorithm as a further step down from the abstract perspectives above, similar to implementing an economic policy: both systematically assign individuals to outcomes. However, whereas economists might have more control over policy objectives and the tools they use, algorithmic fairness often works in a more constrained environment, considering a particular existing task for which an algorithm has been created. In some cases the degrees of freedom are even more limited: we may only be auditing a black-box algorithm without a way to inspect its implementation. At the same time, the specificity of the setting makes it easier to perform rigorous analyses that are closely tied to real-world case studies. Arguably, this is by design: an undeniable motivation for studying algorithms is their meteoric rise of use in consequential decision-making circumstances where the models have already been created[3,12,71]. The definitions we review thus are mainly concerned about how individuals are treated on the basis of the outcomes they have been given; we audit for fairness given the algorithm's results.

There are two overarching approaches to algorithmic fairness today: *individual fairness* and *group fairness* (sometimes called *statistical fairness*). At a high level, these approaches respectively interpret fairness as *treating similar individuals similarly*, and *ensuring groups of individuals receive similar aggregate outcomes*. To clarify these concepts we introduce some notation. We use P and E to denote

probability and expectation.[vi] Let V be the "world" of all possible individuals V as presented to an algorithm. In situations where we are evaluating an algorithm that has acted on a concrete set of individuals, we denote that set as X, a finite subset of V. To be precise, each element $x \in V$ is *not* a human but rather a *representation* of the human through a fixed-length vector of *features* available to the algorithm, $x = (x^1, \ldots, x^k)$. Each feature x^θ denotes an attribute θ of the individual whose value might be a scalar or one of a finite number of descriptive quantities, e.g. income and race respectively.[vii] The important takeaway is that when we talk about individuals in this thesis, we are actually *always implicitly talking about representations of those individuals!*

An algorithm implements a task of assigning individuals a particular outcome, or result, from a *result space* R. For instance, $R = \{0, 1\}$ could reflect a binary decision like "admit to college" or "don't admit"; $R = [0, 1]$ could reflect a probability such as "likelihood of defaulting on a loan", and in classification, R might be a probability distribution over a set of classes A, i.e. $R = \Delta A$, where ΔS is the set of probability distributions over a finite set S.[viii] The algorithm is then a deterministic map $f : V \to R$. X may be partitioned into several disjoint subsets X_i, which we call groups. If there are g groups, we can write $X = \bigcup_{i=1}^{g} X_i$. Importantly, our conception of groups does not allow for overlap. These groups are

[vi] We assume background familiarity with probability, which are discussed in almost every introductory statistics text, e.g. Wasserman[103].

[vii] That is, the way we think about non-quantitative features that do not have natural representations as numbers, such as race and gender, is agnostic to how those features might actually be encoded in bits to the algorithm (e.g. one-hot encoding). In other words, we do not care about details of data representation, e.g. we treat the feature of race as abstractly taking values in {`white`, `black`, `asian`, …}.

[viii] We give a more formal definition in Part 2.

often identified by having a value of a feature in common (e.g. race or economic stratum). This is analogous to Roemer's model of "types" of individuals.[ix]

Individual Fairness. The study of individual notions of fairness was jump-started by Dwork et al.'s seminal 2011 work "Fairness Through Awareness", often identified as the origin of modern algorithmic fairness discussions, which we call the "awareness" paper[28,108]. To make the concept precise we need a way to quantify what it means for individuals to be similar. Dwork et al. obtain this by assuming there is some task-specific function $d : V \times V \to \mathbb{R}$ that can be used to quantify distances between individuals; specifically, d should be a metric, so to guarantee "nice" properties about distances between individuals. Similarly, they consider the presence of a metric between elements in R, which they narrow to be the space of probability distributions over classes A in a classification setting: $D : \Delta A \times \Delta A \to \mathbb{R}$.[x] The fairness constraint is a bound on variation in distances:

$$\forall x_1, x_2 \in V,\ D(f(x_1), f(x_2)) \leq d(x_1, x_2)^{28}.$$

Related works following this framework consider alternate constraints on dis-

[ix] A definition of groups in X in terms of features can naturally extend to a more general partition of the overall space V. For instance, if groups are determined by race, we get a corresponding a partition of V based on the value of the race feature. Indeed, the awareness paper mentions this more general case in which a group is defined as a probability distribution over V.[28] We will be less general and focus on cases where X is finite and groups form a partition of X, instead of thinking in terms of an infinite space of individuals V and groups as distributions over V. However, we mention the definition of V to distinguish between a general infinite set of "all possible people" versus "the people under consideration by this algorithm".

[x] Metric spaces will be formally introduced in Chapter 2. The "awareness" can be interpreted via the idea that the metric is "aware" of how individuals differ. As a further remark, setting $A = \{0, 1\}$ gives an isomorphism between ΔA and $[0, 1]$ as sets of outcomes, by associating $r \in [0, 1]$ with the Bernoulli distribution with probability r in ΔA.

tances. For instance, (ϵ, δ)-*individual fairness* corresponds to the stipulation

$$\forall x_1, x_2 \in V,\ d(x_1, x_2) < \epsilon \implies D(f(x_1), f(x_2)) < \delta,$$

while other constraints bound the "additive distortion" between distances:

$$\forall x_1, x_2 \in V,\ |d(x_1, x_2) - D(f(x_1), f(x_2))| < \rho$$ [108,37].

Although this definition is ambitious, its core weakness is that it may not be obvious, or even tractable, how to define d given a task, especially if the given data about individuals is flawed in the first place. This may make these concepts implausible to work with. While some research has been done in estimating d from data and oracles such as a "human fairness arbiter", there has been comparatively little discussion about implementing individual fairness in practice [52,75,111,57].

Another view on individual fairness is inspired by legal notions of "protected" attributes of individuals, analogous to rhetoric about "color-blind admissions" and Dworkin's separation of relevant and irrelevant circumstances. Under this view, some features in the representation vector x might be deemed sensitive or protected (we use the terms interchangeably), such as race or gender, and so should not be explicitly used as inputs into decision-making processes [110].[xi] Pragmatically this translates to training an algorithm on only unprotected features by removing protected features from X. The question of what features are protected are usually assumed to be decided by practitioners, but could also be answered empirically, e.g. through surveying people [42]. Presumably as a contrastive reference, this

[xi] This corresponds with the intuition that an applicant's race or gender should be *irrelevant* and not affect a decision-making process given that the process ought to treat individuals equally. This is the broad idea behind color-blind admissions, which thus makes it ideologically opposed to affirmative action programs [62].

condition is sometimes called *fairness through unawareness*. One issue identified with fairness through unawareness is that even if a sensitive feature is not explicitly used as an input to an algorithm, there might be *redundant encodings*: there may be ways of determining discriminatory information from other features[71].[xii]

Kusner et al. strengthen this notion by adopting the contributions of casual inference, particularly the work of Judea Pearl, and the use of counterfactual models in law and social science, leading to what they call *counterfactual fairness*. At a high level, their setup considers algorithms to be representable as causal models, directed acyclic graphs of relationships between different variables, where a variable is a feature or a function of other features and variables. An algorithm is counterfactually fair if it returns the same output for an individual regardless of whatever values their protected features are set to in the causal model[83,64]. Refinements of this approach that simplify computations have been proposed, but overall this approach presents a modeling challenge at the outset where the question of how to choose a "correct" causal model relating features does not always have a clear answer[20]. Modifications to the metric-theoretic individual fairness constraint taking inspiration from this perspective have been proposed. For instance, letting V be written as $U \times W$ (and an individual $x = (u, w)$) and an

[xii] This is also noted in "Fairness Through Awareness", where the strategy is named "fairness through blindness": for instance, when considering user data on social media, "there is a very real possibility that membership in a given demographic group is embedded holographically in the history. Simply deleting, say, the Facebook 'sex' and 'Interested in men/women' bits almost surely does not hide homosexuality"[28,55]. In fact, Kusner et al. devise scenarios where causal models can be structured such that ignoring sensitive attributes actually perpetuates unfairness; the parallel here is comparable to Roemer's reasoning that achieving equality of opportunity actually requires an understanding, not ignorance, of individual types. Since effort for him might be dependent on type, to wholly ignore type would preclude the possibility of a faithful understanding of an individual's circumstances[64,88].

algorithm as a map $f : U \times W \to R$, where U and W represent unprotected and protected features respectively, *uniform individual fairness* definitions proposed by Xu and Strohmer bound the distortion between individuals accounting for arbitrary values of W: $\forall x_1, x_2 \in V$, $\sup_{z_1,z_2} D(f(u_1, w_1), f(u_2, w_2)) < d(x_1, x_2)$, or $d(x_1, x_2) < \epsilon \implies \sup_{z_1,z_2} D(f(u_1, w_1), f(u_2, w_2)) < \delta$.

Group Fairness. Group notions of fairness, inspecting how algorithms treat the groups X_i in aggregate, originate from an even more pragmatic viewpoint. As context, inspirations for this perspective include statistical tests for bias in models which have been used in policy and political science. For example, the concept of *disparate impact* is used in legal discussions, denoting a policy practice that exhibits a disproportionately adverse effect on protected groups (X_i that are defined by sharing a value of a protected feature). One mathematical formalization of this, the "80% rule" used by the United States Equal Employment Opportunity Commission, sets $R = \{0, 1\}$, interpreting $f(x) = 1$ as individual x being "selected" by the task. Given two groups $X = X_1 \cup X_2$, f violates this rule if

$$\frac{P(f(x) = 1 | x \in X_1)}{P(f(x) = 1 | x \in X_2)} \leq 0.8;$$

put into words, if the overall probability that f returns 1 differs significantly between the two groups[34].[xiii] Here, an aggregate statistic is used to quantify the idea that similar groups should be treated similarly.

[xiii] A related legal notion is disparate *treatment*, where individuals might explicitly be discriminated against based on protected attributes. Effectively, disparate impact can be viewed as an unintentional version of disparate treatment[110]. A definition of algorithmic fairness serving as a rigid formalization of disparate treatment is, however, harder to conceptualize. We might make a comparison to "fairness through unawareness", but as we have already seen, determining whether or not a protected attribute is used is not as straightforward as one might think, and there are also situations in which accessing protected attributes *helps* to accomplish fairness.

Most group fairness definitions essentially differ in the particular statistics they apply to measure and compare groups. Mehrabi et al. provide a summary of other commonly-used metrics, some of which we list here[71]. *Statistical parity*, also called demographic parity, considers the distribution of outcomes outputted by the algorithm on groups, and limits the divergence between distributions for different groups as measured by metrics between distributions. Thus members of groups are equally likely to land in a set outcomes, and observing an outcome indicates nothing about group membership[28]. In the case $R = \{0, 1\}$, it simplifies to equal proportions of selection, matching the 80% rule: $\forall i,\ P(f(x) = 1 | x \in X_i) = P(f(x) = 1)$. This makes the criteria one of the more well-known interpretations of fairness given its close link to the definition of disparate impact[8]. Corbett-Davies et al. provide a more granular version of this named *conditional statistical parity* which conditions on unprotected features: using the notation $x = (u, w)$ as above, the constraint is $\forall i,\ P(f(x) = 1 \,|\, u, x \in X_i) = P(f(x) = 1 \,|\, u)$[25].

Some definitions of group fairness apply to the setting where the task for f is to predict a quality or outcome of the individual that exists but cannot be measured; in these cases, we let $f(x)$ be the prediction for the actual value $y(x)$, respectively abbreviated as $\hat{y}$ and y whenever clear. Specifically, let $y, \hat{y} \in \{0, 1\}$. A true positive (TP) occurs when $\hat{y} = y = 1$, and a true negative (TN) occurs when both are 0. A false negative (FN) occurs when $y = 1$ but $\hat{y} = 0$, and a false positive (FP) occurs when $y = 0$ but $\hat{y} = 1$. All individuals fall into one of these four scenarios, and in terms of numbers $TP + FN = |\{x \in X : y(x) = 1\}|$ and $TN + FP = |\{x \in X : y(x) = 0\}|$ (knowing TP is sufficient to calculate

FN and vice versa, and similarly for TN and FP).[xiv] A natural interpretation of group fairness is then equalizing the prevalence rates of these four metrics, or functions of these metrics, among groups. Hardt et al. formalize definitions of *equal odds* and *equal opportunity* in a setting with $R = \{0, 1\}$: f achieves equal odds if $\forall r \in \{0, 1\}, i$, $P(\hat{y} = 1 | x \in X_i, y = r) = P(\hat{y} = 1 | y = r)$; in other words, the TP and FP rates should be equal across groups.[xv] Equal opportunity is a weaker condition that only requires TP (equivalently, FN) rates to be equal across groups; correspondingly, equalizing TN (equivalently, FP) rates is called *predictive equality*[44,100].[xvi] *Predictive parity* seeks to equalize the *positive predictive value*, or precision, of each group, defined as $P(y = 1 | \hat{y} = 1) = \frac{TP}{TP+FP}$, and *treatment equality* seeks to equalize the ratio of errors $\frac{FN}{FP}$[71].

A further special case in the setting of predicting outcomes occurs when

[xiv] The setting is equivalent to binary classification with $y = 0$ and $y = 1$ as the classes. This setup is often used in "screening problems" that select individuals from a group of candidates. In sentencing $\hat{y}$ might be a categorization of an individual as "high-risk" or "low-risk", or in diagnoses $\hat{y}$ might estimate whether or not a patient has a disease[59]. These quantities reflect some ground truth quality that is true or false but unknown to practitioners. Sometimes FNs are more consequential than FPs (e.g. when the FN is a false negative diagnosis) or vice versa (e.g. when the FP means an innocent defendant is punished). So working with definitions is not enough; thought should be given to the significance and stakes of the task. These concepts overall are sometimes referred to as the *confusion matrix* in the statistics literature[100].

[xv] The TP *rate* can be written as $P(\hat{y} = 1 | y = 1) = \frac{|\{x : x \in X, f(x) = y(x) = 1\}|}{|X|}$, conditioning on $x \in X_i$ when defined within group X_i; analogous expressions hold for the remaining rates. Thus when there are two groups $X = X_1 \cup X_2$, this can be written as $P(\hat{y} = 1 | x \in X_1, y = r) = P(\hat{y} = 1 \,|\, x \in X_1, y = r)$. The authors note that their definition extends for any general R, but the binary classification setting where $R = \{0, 1\}$ is most prevalent in practice and thus the focus of their analysis. Note the symmetry with predictive parity (defined later in the paragraph); this analyzes algorithmic labels $\hat{y}$ conditional on the actual truth y, while predictive parity does the opposite. Choosing the suitable direction of conditioning is a decision for practitioners to make.

[xvi] We can interpret equal opportunity as ensuring that true positives in groups are correctly "chosen" with equal probability. The authors provide examples where weakening the fairness constraint from equal odds to equal opportunity can lead to increased utility, thus presenting a situation where we may trade off the strength of fairness guarantees for better overall outcomes.

$y \in \{0, 1\}$ but $R = [0, 1]$, such that the output of the algorithm reflects a probability estimate, sometimes called a score, for the event $\{y = 1\}$ (e.g. the probability a defendant will re-offend within n years). *Calibration*, or test fairness, is an adaptation of predictive parity to this scenario, considering the fraction of correct positive predictions for each possible predicted score $r \in R$. In the case of two groups $X = X_1 \cup X_2$ this is the statement $\forall r \in [0, 1]$, $P(y(x) = 1 \,|\, f(x) = r, x \in X_1) = P(y(x) = 1 \,|\, f(x) = r, x \in X_2)$. *Well-calibration* is a stronger condition that hold that these probabilities should in fact be r, i.e. the algorithm's estimated probabilities are accurate. Since r could take on uncountably many values in principle, this is often dealt with in practice by applying this condition over intervals of r in a kind of discretization[112]. *Balance for the positive class* requires that individuals in the positive class in reality, where $y = 1$, should receive the same scores on average, for two groups: $E(f(x)|y(x) = 1, x \in X_1) = E(f(x)|y(x) = 1, x \in X_2)$, and *balance for the negative class* is the same but applied to the negative class, i.e. individuals with $y = 0$. These definitions also apply to arbitrary numbers of groups. Given the wealth of statistical concepts to examine, there are several more group fairness definitions which consider various other statistical criteria, a detailed review of which is given by Verma and Rubin[100].

More recent work has questioned what constitutes a group and how to select groups in the first place. For instance, simply using "race" as a criterion to create groups ignores issues of intersectionality and other possible salient ways of dividing individuals into groups. At the same time, it is sometimes unclear what level

of granularity defining a group should be (e.g. when grouping people by income or age, how large the brackets should be). Kearns et al. introduced the idea of "subgroup fairness", proposing techniques that can audit for fairness criteria over arbitrarily-many "subgroups" in an efficient manner[57]. Independently, Hébert-Johnson et al. proposed *multiaccuracy* and *multicalibration*, fairness conditions which consider some arbitrary collection of subsets $\mathcal{C} \in 2^X$, where 2^X denotes the set of subsets of X. Roughly, f is multicalibrated if it is calibrated for all of the subsets. In their strict definition, $\mathcal{C}$ could be any collection of subsets, making multicalibration unwieldly to work with in principle. To resolve this, the authors consider situations where $\mathcal{C}$ consists of only subsets that are computationally efficiently identifiable. They provide techniques to make a predictor multicalibrated with a runtime complexity bounded by the theoretical maximum size of a circuit testing for set membership in a certain element of $\mathcal{C}$. The authors also work with a more general prediction setting where individuals have true "latent" probabilities to be predicted ($y \in [0, 1]$); details of their exact definition of calibration are found in their paper[112]. With perhaps the most broad analysis out of the papers above, Dwork et al. showed that multicalibration can be seen as a particular instance of an *outcome indistinguishability* condition for predictors: general criteria where an algorithm's predicted probabilities cannot be separated from "true" probabilities by a class of distinguishing strategies of varying computational complexities[29].

Much of the literature on algorithmic fairness does not grapple with existing definitions or propose new ones, instead focusing on auditing or correcting real-life algorithms with respect to existing fairness criteria and devising new strategies to

train new algorithms obeying those criteria in practice. Logically, in a field concerned with real-life outcomes, work ought to extend beyond theory and consider reality. Indeed, the last section in our review includes work motivated by a controversial and publicized case study. The work reveals a troubling situation when evaluating how fairness concepts relate to each other in principle: that conflicts between definitions arise, and in fact are sometimes inevitable.

1.3 The Impossibility of Fairness

Given the wealth of fairness definitions above, we might hope that there exists some sort of unifying relationship between them. Unfortunately, this is not the case: parts of the literature critique and find fundamental limitations in the concepts we have just covered. Thematically, we identify three main categories of issues with algorithmic fairness: the inadequacy of concepts to ultimately prevent undesirable outcomes; contradictions and incompatibilities between different fairness criteria; and ontological critiques of the well-formedness of fairness criteria.

Inadequate fairness constraints. It is sensible to question whether or not, for instance, using simple statistical criteria upon groups sufficiently captures an intuitive sense of "fairness". In fact, some work goes a further step back and analyzes the limitations of working in an aforementioned local setting.

Caterina Calsamiglia, in her doctoral dissertation, continues in the tradition of Young and Roemer in analyzing the tension between local and global distributive justice. She interprets local equality of opportunity problems, i.e. the problem of making welfare based only on characteristics deemed relevant and independent of

irrelevant ones, as a decentralization of the global problem. She notes that policymakers are often limited in the scope of information they can access and have no good way to gauge effort.[xvii] Moreover, individuals often face trade-offs between investing effort in various parts of life (e.g. studying vs. part-time work), so effort levels cannot be viewed as independent in different areas; there is an issue of "dispersion of information and interrelation of local environments"[17]. Consequently, local problems typically do *not* coordinate to solve the global problem.

Calsamiglia proposes an example of two individuals P, R who study and play basketball, identical in "innate ability" or potential but with R having better access to educational resources. Both split effort between studying (e_S) and basketball (e_B) with respective resource levels r_S, r_B. Calsamiglia models an individual's measured percentile of ability a_S, a_B as a function of resources and time spent:[xviii] $a_S \propto r_S\sqrt{e_S}$, $a_B \propto r_B\sqrt{e_B}$, and the cost of studying and practicing as increasing with effort $c(e_S, e_B) \propto (e_S + e_B)^2$. Consider a college admissions officer and NBA recruiter adopting a Roemerian point of view to evaluate P and R based on their measured ability in studying and sports, ignoring the "irrelevant characteristics" of local resources (since those are beyond P and R's control), admitting an individual at the ath percentile of ability with probability a. P and R optimize for their welfare, $a_S + a_B - c(e_S, e_B)$. While the innate talent of P and R are equal, R has a higher marginal productivity in school (with respect to increasing odds of admission) because of better resources, so their optimal strategy is to spend

[xvii] This argument is much like that brought forth by the criticism of Matt Cavanagh, who documents empirical examples where applying global or local equality of opportunity leads to different policy recommendations[18].

[xviii] Since percentile values lie in $[0, 100]$, this function is implicitly clipped to lie in that range.

more time studying and less time on basketball relative to P. Consequently R is more likely to be admitted to school, and the individuals are predisposed to different outcomes despite having the same potential. Differences in irrelevant characteristics (resources) threaten equality of opportunity[17].

Suppose the admissions officer tries to achieve equal opportunity for P and R being admitted. If they do this by artificially *adding* to P's value of a_S to match R's in an affirmative action sense, then a discrepancy will arise in basketball. Individuals' efforts remain unchanged, but R is now disadvantaged, having equal odds of college admission but a lower probability of being recruited since P allocated a higher e_B. Alternatively, forcing P to spend additional effort until their a_S is the same as R's leads to their spending less time on basketball and lowers their welfare compared to the baseline scenario without intervention, making them worse off. So solving only the local problem (admissions) is counterproductive to other parts of the global problem (basketball). Calsamiglia shows that a way to guarantee concordance between the local and global setting is by making local mechanisms provide equality of rewards to effort, i.e. requiring two individuals with the same relevant characteristics and effort to be treated identically, and that under some technical conditions, decentralizing global equality of opportunity and equalizing rewards to effort are equivalent. In the example, one way to do this is scaling a_S by a multiplicative, not additive, factor to account for differences in resources[17].

Some critiques explicitly demonstrate how undesirable outcomes fail to be captured by certain fairness definitions. In the awareness paper Dwork et al. specifically target the insufficiency of statistical parity, for ensuring similar aggre-

gate distributions of outcomes over groups still leaves room for bias. For instance, an algorithm might choose individuals from a group for the sole purpose of claiming statistical parity without any intention of being treated fairly later on, or to "'justify' future discrimination... building a case that there is no point in 'wasting' resources" on the group, as has been seen through interviewing "quotas" in hiring[28]. Statistical parity also leaves open the possibility of discriminatory action within subgroups of each group. Implicit here is a critique about the conceptual limitations of group fairness and the undesirable "coarseness" of only considering how a group is treated overall, the motivation behind developments of subgroup fairness and multicalibration[57,112]. Concerningly, Liu et al. adopt a broader lens and examine how fairness criteria impact algorithms and the individuals they affect over time, showing that with repeated application of algorithms obeying fairness criteria it is possible for the welfare of groups to actually decrease over time. This critique targets a different shared nature of fairness constraints, in that they do not consider the long run: they apply to algorithms taken as implementing an isolated instance of a task, and do not take into account how they fare with repeated application, being greedy in some sense[67].

A common theme here is the idea of *locality* impeding global fairness by only seeing a limited picture. Calsamiglia shows that under an assumption that individuals implicitly distribute effort over multiple settings, being narrow-minded by ensuring a notion of fairness in one situation disregards other situations and can lead to worse overall utility. Temporally "local" applications of fairness do not necessarily lead to optimal results in the long term. Critiques of group fairness stem

from the fact that they lack the sufficient granularity to ensure non-discrimination, and defining a group may be unclear. Similarly, critiques of fairness through unawareness note that what a protected feature means is not clear. Further, even if fairness through unawareness were obeyed and no group or sensitive information could be observed, observe that many of the fairness concepts we have discussed, such as Roemer's evaluation of effort as relative to an individual's type and Calsamiglia's requiring an accounting for discrepancies in resources explicitly require us to consider individual attributes that might be barred from consideration. This is the classic equality versus equity debate: in situations where existing inequalities exist between groups, being blind to protected features could lead to the perpetuation of those inequalities and prevent equitable outcomes. In all these cases, locally good intentions fail to capture the global truth.

Incompatibilities between different fairness criteria. One of the best-known case studies in algorithmic fairness, and an example responsible for kick-starting much discussion in the field, ironically provides a well-known example of the *limits* of fairness definitions. It is found in the area of criminal sentencing.

In 2016, several years before the advent of large language models and discussions about artificial intelligence had entered mainstream discourse, the investigative journalism organization ProPublica published an article examining a sentencing tool developed by a company, Northpointe. This software, called COMPAS, was intended to aid with risk assessment efforts for criminals by computing scores predicting the likelihood of recidivism. Such risk assessment tools, including COMPAS, ProPublica claimed, despite being adopted at a growing rate

through the United States, displayed systematic racial bias, what they called "machine bias", in their decisions. Their conclusions came from analyzing data from arrests made in Florida, seeing how the algorithm would classify them as likely or unlikely to re-offend based on if the predicted score passed a certain threshold.[xix] Controlling for other factors, black and white defendants were classified with the same overall error rate, but non-recidivist black defendants were erroneously classified as high risk at almost twice the rate. In other words, the FP rate for the group of black individuals was disproportionately high and the FN rate for the group of white individuals was disproportionately low[5]. Similar reporting had been done by the Associated Press and The Marshall Project the year prior[9,96].

Northpointe published a lengthy objection and defense, characterizing the investigation as using inappropriate statistics. Calculating FPs and FNs, they argued, was nonsensical as such an action could only be done retrospectively (by definition, it requires future knowledge on whether or not an individual actually did re-offend, evidently unavailable at the time of sentencing). Instead, what mattered, they claimed, and what the COMPAS algorithm was trained to satisfy, was predictive parity. As we have seen, this meant that among black and white individuals, the positive predictive value was roughly the same, i.e. the probability that a defendant being classified as high-risk actually re-offending was roughly the same regardless of race[26]. A fundamental incongruity was clear: the

[xix] Using our notation, the algorithm itself produced values in the set $R = \{1, 2, \ldots, 10\}$, with "low risk" denoting scores up to 4. ProPublica's analysis deemed an individual as high risk, i.e. likely to re-offend, if their score exceeded 4. So they really considered a modified version of COMPAS with the outcome set $R' = \{0, 1\}$ by adding the post-processing step $r \mapsto I_{\{r>4\}}$, where I_A denotes the 0/1 indicator variable for event A[4].

same algorithm, applied on the same data but evaluated with different fairness criteria resulted in drastically different judgments. What was fair as judged by Northpointe turned out to display racial bias measured another way. ProPublica later refuted Northpointe's report, but noted that it was a known statistical issue that systems that satisfy equal accuracy among subgroups could still be unfair[4,66]. The debate has carried on: In 2020 Rudin, Wang, and Coker tried to reconstruct COMPAS (the full model is proprietary) and claimed that while their reconstruction violated Northpointe's claim that risk scores depended linearly on age[90], while another defenses of COMPAS came from Jackson and Mendoza[53].

Might it have been possible to, with access to the model behind COMPAS, make modifications to rectify ProPublica's concerning discoveries with Northpointe's fairness objective? Theoretical work published in response to controversy showed that in fact, the answer was *no*: these contradictions were inevitable. Soon after ProPublica's investigation was presented, Kleinberg et al. published an article providing a mathematical view into the issue at hand. They note that COMPAS also satisfied the fairness criterion of calibration (with respect to the groups of white and black defendants) at the cost of violating balance in the positive and negative classes, and that this violation must hold: unless groups have equal base rates, no algorithm exists satisfying calibration and balance simultaneously. They analyze calibration in a setting where the outcomes (risk scores) are predicted probabilities $R = [0, 1]$ as follows: by definition, for each group X_i where outcomes are collected into "bins" for each $r \in R$, the fraction of individuals from X_i who are in the positive class (re-offended in reality) should be r of the size of

the bin in expectation. Then the sum of scores given to X_i equals the number of individuals in X_i in the positive class, call it μ_i, and the following equation holds:

$$(|X_i| - \mu_i)r_{i,-} + \mu_i r_{i,+} = \mu_i$$

where $r_{i,-}$ and $r_{i,+}$ denote the average score given to individuals of the negative and positive class in X_i respectively. When comparing groups, their μ_i are equal if and only if their base rates are equal; otherwise, the above equation for each group yields a linear system in the variables $r_{i,-}, r_{i,+}$. Unless $r_{i,-} = 0$ and $r_{i,+} = 1$ always, which is attainable only if the positive and negative classes are trivially known and perfectly predictable (which would defeat the entire premise of needing the algorithm), then a solution to the system must have them set to varying values among groups, which by definition violates balance. This insight contextualized ProPublica's analysis: in COMPAS's case, inbalance for positive and negative classes manifested as disproportional FP and FN rates among groups[60].

Chouldechova concurrently formulated a similar impossibility result, focusing on a part of Northpointe's response observing that the base rates for black and white defendants were different, where the base rate (or prevalence) for group X_i is defined as the proportion of individuals in a group X_i with a "true" positive outcome, i.e. $P(y(x) = 1 \mid x \in X_i) = \frac{|\{x: x \in X_i, y(x)=1\}|}{|X_i|}$. Then a simple equation can be derived which relates the false positive and negative rates (FPR and FNR), PPV, and base rate (p) within any given group:

$$FPR = \frac{1}{1-p} \cdot \frac{1 - PPV}{PPV} \cdot (1 - FNR).$$

This immediately begets a similar conclusion: if base rates between groups differ,

then it is impossible to maintain the same PPVs, false positive rates, and false negative rates between groups[21]. Since base rates were indeed different in COMPAS's case, Northpointe, by choosing to equalize PPVs, implicitly ensured that false positive and negative rates would differ between races. Again, COMPAS made a tradeoff, choosing one version of fairness at the expense of another.

Notably, the quantities referenced in Chouldechova's analysis are not novel to algorithmic fairness, all coming from the statistics literature. Indeed, the problem she describes is general and has been encountered in other settings. Essentially, if some ground truth attribute intrinsically differs between groups, this imperils the mutual coherence of different versions of fairness. In the above setting, this attribute turned out to be differing base rates, explicitly used by Chouldechova and equivalent to the quantities μ_i used by Kleinberg et al. Berk et al. further delve into the connection between the above two results, concluding that "except for highly stylized examples," "the goal of complete race or gender neutrality is unachievable"[12]. This conundrum is not limited to criminal justice: it is just one way of mathematically substantiating the postulation in the Introduction that if bias is "baked in" and inherent in the data to begin with, then it is impossible to harness that data in a completely fair way. Thus, the foundational issue is working with data in reality that contains bias (Here, we would need these ground truth attributes to be equal across groups.)

As an example of this line of reasoning in a related field, Neil and Winship examine the problem of auditing for racial discrimination in policing through an example of sample data on police stops that differ for black and white people.

They identify three broad issues. First, in the "denominator problem", tests for fairness using ratios as statistical criteria may use unsuitable values as the denominators in these ratios (in this case, a disconnect between what an auditor thinks the police are doing versus what they are actually doing, e.g. in which situations they conduct stops), which can lead to mismeasurements of bias. Second, improper levels of data aggregation and stratification can mask or reverse trends in data, e.g. if different races are found in different proportions geographically and police conduct stops more frequently in places with more black individuals, even if police are unbiased when conducting stops conditional on place the overall pattern is biased.[xx] Third, they describe *infra-marginality*, whereupon different distributions of behavior between black and white people leads police to conduct stops differently depending on race. Discrimination in the form of this differential treatment may lead to a higher or lower overall TP rate (hit rate of police stops), so that statistical criterion reveals nothing about discrimination[76]. This last effect has been documented through examination of historical data[94].

The lack of harmony between fairness notions stresses a need for discretion when it comes to choosing *which* notions to satisfy, since there is no "all-encompassing" sense of fairness. However, some newer work has examined cases in which it is possible to achieve a compromise between different fairness criteria and overall utility by slightly relaxing conditions (e.g. in (ϵ, δ) individual fairness, increasing the pa-

[xx]This is the same idea captured by Simpson's Paradox. In fact, this phenomenon was investigated in an early investigation of bias in admissions, dating from the 1970s. Studying data from Berkeley graduate admissions revealed that female applicants were overall admitted at a lower rate than male applicants, showing bias. However, conditional on department women were admitted at a similar, if not higher, rate than men; the confounding aggregation factor was that women tended to apply to more competitive programs (for all applicants) than men[13].

rameters ϵ, δ), exploring a Pareto frontier between competing notions and utility in settings like learning specific classifiers or modifying outcomes post-hoc[108,68].

Ontological limitations of fairness. The last category of challenge to algorithmic fairness is an existential one which questions the validity of our conception of algorithmic fairness, particularly coming from the perspective of law.

The legal scholar and sociologist Issa Kohler-Hausmann has shared several influential critiques of our current understanding of fairness in practice. Her thrust of argumentation concerns the way that academia conceives of and treats sensitive attributes such as sex and race. For instance, in response to the Supreme Court's decision striking down affirmative action programs, she shared extensive critiques regarding the meaning of race-neutral admissions in itself. Among her numerous arguments, she highlights the fact that race often indirectly and unconsciously factors into large parts of the application process when admissions officers evaluate an applicant in light of their experiences and circumstances. It is intractable to simply "excise [race] and leave social and cognitive antimatter in its place"; achieving race neutrality cannot simply occur by pretending race does not exist, but requires an acknowledgement that the current state of the world is manifestly not equal on the basis of race, fundamentally affecting the application process and colleges' goals surrounding diversity[62,63].

These conceptual difficulties in thinking about race as a simple feature of individuals is the broad idea unifying her criticisms discussed here. At the core is the argument that attributes such as race and gender cannot only be thought of as features of an individual, but rather *constitutive* phenomena that are responsi-

ble for wholly determining one's lived experience, development, and identity (c.f. the impact of centuries of discrimination and ongoing prejudice when it comes to attributes like race, sex, gender, and socioeconomic status; being born as a black individual predisposes one to systematically different conditions, and behavior). As such, an individual's race is not independent, nor even correlated or confounded with, other features: it is in part *constituted* by all other features, and likewise *contributes* to those other features' meanings. To validly interpret the feature of "wears dresses" requires "situated cultural knowledge" of gender (and perhaps sexuality, ethnicity, etc.) and their associated social norms interacting in a "complex interrelationship"[61,76].

Consequently, notions of fairness treat the "thick ethical concept" of discrimination reductively. Treating race, for instance, as an arbitrary biological feature ignores the web of social constructs that make race the consequential feature it is today. Kohler-Hausmann specifically takes aim at counterfactual methods, such as those examined by Pearl and Kusner et al.. For instance, a popular technique in social science of detecting discrimination in a scenario is by manipulating a person's race *ceteris paribus* and considering if they would be treated differently. The issue here, Kohler-Hausmann claims, is not just that such a manipulation cannot be practically implemented,[xxi] but also that it would wholly transform the scenario such that the *ceteris paribus* assumption is a massive oversimplification; to really understand discrimination requires normative evaluations surrounding

[xxi] An example of discussion regarding the well-foundedness of using attributes like race and sex as Glymour and Glymour, who discuss how reasonable "hypothetical intervention[s]" ought to be viewed in the context of performing statistical inference[40].

relevant social meanings and practices[61]. To truly change the value of an individual's feature likely entails changing other features in a comprehensive way; in our language, if we had $x \in X_1$, simply making the change $x \in X_2$ without modifying the values of x itself might make it "out of distribution" and dissimilar to other individuals in X_1, because to properly imagine the process of "moving" x to X_1 requires an all-encompassing analysis of how the whole vector x would change[49]. In other words, context is deep and rich, making it intractable to manipulate features in isolation.

Finally, a question we can pose is how faithfully these definitions match the spirit of the tradition led by Rawls. Reuben Binns investigates this topic, warning against "the blunt application of fairness measures". He argues that while in principle both individual and group fairness share the same desire to fulfill egalitarian outcomes and treat individuals or groups consistently, they reflect different worldviews: that disparities arise from personal choices, or unjust structures, respectively. Furthermore, there is another kind of fairness in the form of *individual justice* that existing measures do not capture: the idea that a person should be assessed individually rather than on the basis of generalizations derived from others similar to that person. Such an idea can be empirically found "in calls for public administrators to exercise discretion rather than routine application of rules". Group fairness fails to capture this, as it follows the error of only testing for fairness in a generalized way (on the basis of group membership), while individual fairness makes an algorithm's outcome for an individual close to that of similar individuals (which admits a group generalization if we consider the group formed

by that individual with others who are "close" by the metric)[14].

1.4 Is Fairness Futile?

To sum our discussion up, we have presented an overarching narrative of an evolution of sorts: that of how academia thinks about fairness from principled origins to exact mathematical environments. It might also be apt to describe this narrative as one of adaptation, where the ideas at hand reduce their breadth in favor of depth. In a purely philosophical setting, there was free rein to define fairness on the broadest grounds; moving into the world of economics and law demanded the consideration of fairness as it applied within more particular situations; and faced with the algorithm, we arrive at the numerous precise mathematical and statistical definitions we have reviewed.

What to say of the bleak picture painted by the impossibility results of Kleinberg et al. and Chouldechova, or the conceptual critiques of Kohler-Hausmann and Binns? Foremost, the presence of impossibilities does not rend the project of algorithmic fairness vain. Rather, the fact that there are conflicting concepts can serve a reminder that thinking about justice in the real world, compared to on paper, can be a far more challenging affair, and should serve as a warning to practitioners to be thoughtful in what the goals of their algorithm are. As Binns writes, "philosophers, lawyers, and other humanities scholars deal with many different and conflicting notions of fairness that have been the subject of intractable debate for millenia"[14]. We might view these impossibility results as analogous to Gödel's impossibility theorem in pure mathematics, or uncomputability theorems

in the theory of computation; while express theoretical foundational constraints with a field, they do not mean that the field cannot be productive. Fundamental limitations can be worked with as one might with constraints. Indeed, Kleinberg et al. similarly point out that algorithms provide a new way to "let us precisely quantify tradeoffs among society's different goals"[59]. The action of defining what is fair has always been implicitly performed by legislators, admissions officers, economic advisers, and judges alike; when we design an algorithm this action is simply made explicit and gives the opportunity to think about tradeoffs between different kinds of fairness.

Similarly, the fact that the definitions we work with are flawed does not mean they are worthless, and does not preclude their ability to do good. Having a toolkit to inspect models and the types of decisions we are automating today is a marked improvement from an alternative where we have no way to perform critical analysis. Given that the field is especially nascent, with "Fairness Through Awareness" being barely over a decade old (at the time of writing), it is not unreasonable to expect the continual proposal of new definitions and results in the years ahead, some of which might resolve the essential tensions we have covered. The fact that there is much that is currently flawed and far more to be done should serve as a motivator for progress, rather than being a message of futility.

Of course, progress need not be made only through the introduction of new characterizations and measures, or the development of new techniques. Complementary to fairness, some parts of the literature engage in an epistemological project of defining exact kinds of bias and how they arise. The second part of this

thesis is comprised of our contributions on exactly this front.

2

Differential Expressiveness and Data Bias

In Part 1, we referenced situations in which existing data is biased inevitably led to the violation of some fairness criteria. For instance, in context of the COMPAS case study, Chouldechova's impossibility result took the form of an equation relating FPR, PPV, and FNR for a group, dependent on the base rate of the group. If base rates were approximately equal across groups, the discrepancies

between these three quantities could be kept small. Similarly, the idea of "fairness through unawareness" principally bars the possibility of differential treatment of groups (if the protected features correspond with obvious group classifications like race or sex) on the basis that group membership, if determined by protected features, should not matter in making a decision. This is *a priori* a reasonable naive assumption; if individuals all truly lived equally, no one racial group should re-offend at a higher frequency. However, this is conspicuously not the case. The data showed that black defendants re-offended at a higher rate, and it is not hard to conceive of several reasons why: the world we are in still contains injustice, e.g. systemically disparate economic situations that might drive the impoverished to re-offend, inequalities in access to safety nets, or discriminatory practices and racism in police actions. The source of the issue is the world we live in and its portrayal in data. Consequently, an algorithm necessarily faces a choice between how to be fair (and unfair) per Chouldechova, and fairness through unawareness ends up having an adverse impact by disallowing the design of algorithms that better combat existing inequalities through providing equity to specific groups (hence constituting differentiable treatment). In other words, data is the root of evil by running contrary to what we suppose is a just state of the world.

Indeed, the place occupied of data bias in discussions of algorithmic bias at large can be viewed as a preliminary and omnipresent concern. Simply looking at the design of an algorithm in itself does not provide a holistic picture if the behavior of the algorithm ultimately depends on qualities intrinsic to the data used to train it. Since data forms the cornerstone of the entire process of creating mod-

els (including, but not limited to, those using machine learning techniques) and trends in data propagate down to outcomes, any attempts to address algorithmic bias should not overlook opportunities, if applicable, to remediate problematic patterns or aspects of the data itself. While addressing data bias is not a panacea for eliminating algorithmic bias overall, tackling bias is certainly incomplete if data remains unscrutinized and therefore potentially problematic.

The contributions of this part are twofold. The first section introduces a new notion of data bias which we call differential expressiveness (DE), of which a primer was given in the Introduction. Our treatment will be mostly empirical; DE provides a new lens with which we can group together different concrete scenarios where bias in data has been documented or could arise. The second section considers how DE fits into existing frameworks of machine bias, integrating a discussion on related work. Indeed, while the previous part of this thesis focused on describing the evolution of fairness notions, there has also been work in algorithmic fairness seeking to describe and categorize the origin and treatment of bias throughout the machine learning lifecycle. This literature already proposes several different definitions of data bias, for which we will compare and contrast our formulation of DE. Furthermore, we will extend a framework presented by Friedler et al. for thinking about bias and fairness in their paper "On the (im)possibility of fairness" (which we refer to as the "(im)possibility paper") and discuss how DE fits within this framework[37]. The conclusion of this part will briefly address possibilities to audit for DE and review general approaches in the literature for inspecting and reasoning critically about data.

2.1 A Mathematical Language for Fairness

Before introducing DE, we define some mathematical vocabulary that will be used throughout the chapter, particularly surrounding how individuals and groups are represented in data. As we will see, algorithmic fairness inherits definitions given in real analysis and probability. We contribute a unified mathematical presentation of key concepts in the awareness and (im)possibility papers from first principles that adds clarity and rigor. We build upon the notation in Part 1. Recall that the overall space of (representations of) humans is V, from which we consider a finite set X corresponding to (representations of) individuals that are actually evaluated by an algorithm. X is split up into disjoint groups $X = \bigcup_{i=1}^{g} X_i$. In the case that we only care about one group X^* we can compare it to the rest of the population $X = X^* \cup (X \setminus X^*)$. Each element of X is a vector of abstract features x that form a representation for a human individual. An algorithm implements a task that maps (representations of) individuals to outcomes, $f : V \to R$.

The definition of individual fairness in the awareness paper and related individual fairness measures rely on the key idea of a metric space, consisting of a set of elements and a function called a metric to express distances between elements. This enforces more structure on V and consequently X.

Definition 1. Given a set M, a *metric* on M is a function $d : M \times M \to \mathbb{R}$ that obeys the following: for all $\forall a, b, c \in M$, (i) $(a, b) \geq 0$ with $d(a, b) = 0$ iff $a = b$; (ii) $d(a, b) = d(b, a)$; (iii) $d(a, c) \leq d(a, b) + d(b, c)$. If d is a metric for M, the pair (M, d) is called a *metric space*.

The constraints of the metric make it mathematically "nice" and roughly matching our intuition for what distance means in a set. Property (ii) is called the symmetric property (distance should not depend on order of measurement), and property (iii) is called the triangle inequality (the distance between two points should not be able to be shortened by going through a third point). A set M can have multiple metrics defined on it. In individual fairness definitions, we set $M = V$ and require d to be some task-specific metric. Any subset of a metric space equipped with the same metric is also a metric space, since conditions (i)-(iii) continue to hold for any subset of points; because $X \subset V$, we may also view the actual set of individuals under consideration (X, d) as a metric space. The elements of M can be more complicated objects, such as probability distributions—as it happens, the fairness through awareness condition relies on the existence of metrics in the space of probability distributions, since the setup there takes outcomes to be distributions $R = \Delta A$. Below are two examples.

Definition 2. Let M be a finite set and $\mu_1, \mu_2 \in \Delta M$ be two probability distributions defined on M. The *total variation norm*, sometimes called statistical distance, between μ_1 and μ_2 is defined as $D_{tv}(\mu_1, \mu_2) = \frac{1}{2} \sum_{a \in M} |\mu_1(a) - \mu_2(a)|$.[i] The *relative* ℓ_∞ *metric* between μ_1 and μ_2 is defined as $D_\infty(\mu_1, \mu_2) = \max_{a \in M} \log \left(\max \left\{ \frac{\mu_1(a)}{\mu_2(a)}, \frac{\mu_2(a)}{\mu_1(a)} \right\} \right)$.[ii]

To be precise, μ_1 and μ_2 should be thought of as probability measures on M,

[i] This definition has an analogue in measure spaces, generalizations of probability spaces defined below. Given two measures μ_1, μ_2 on M with event set Σ, $D_{tv}(\mu_1, \mu_2) = \sup_{E \in \Sigma} |\mu_1(E) - \mu_2(E)|$ [89].

[ii] For this to be well-defined we must assume μ_1, μ_2 give positive probability to all $a \in M$, or we ignore all a such that $\mu_1(a) = 0$ or $\mu_2(a) = 0$. We implicitly make this assumption.

which we fully define below. It is straightforward to show that D_{tv} is a metric on ΔA through applying definitions. The definition of D_∞ is one given by the awareness paper and it is also a metric on ΔA. Many other notions of metrics and "statistical distances" exist, which is briefly discussed in the Appendix. As a practical matter, however, it is worth noting that the ranges of these metrics can vary; the range of D_{tv} stays in $[0, 1]$, whereas D_∞ is unbounded. Dwork et al., who focus on these two metrics because they lend themselves to efficient computation, note that fairness through awareness requires that the hypothesized task-specific metric scale similarly between similar and different individuals to the choice of distribution metric[28]. In general, metric spaces let us exactly state what it means for distances to be distorted:

Definition 3. Let (M_1, d_1) and (M_2, d_2) be two metric spaces and $f : M_1 \to M_2$ be a map between them. For $K > 0$, f is *K-Lipschitz continuous* if it satisfies the *K-Lipschitz property*: $\forall a, b \in M_1$, $d_2(f(a), f(b)) \leq K d_1(a, b)$.

Definition 4. Let $f : M_1 \to M_2$ be a map as above. The *additive distortion* of f, ρ_f, is the least upper bound on the differences between distances induced by the map: $\rho_f = \inf_{a,b \in M_1} |d_1(a, b) - d_2(f(a), f(b))|$.

Definition 5. Let $f : M_1 \to M_2$ be a map as above. We say f is *(ϵ, δ)-continuous* if $\forall a, b \in M_1$, $d_1(a, b) < \epsilon \implies d_2(f(a), f(b)) < \delta$.

From Part 1, we can now identify fairness through awareness as enforcing Lipschitz continuity as applied to $M_1 = V$ and $M_2 = \Delta A$ with some choice of

metrics and $K = 1$,[iii] while alternate formulations of individual fairness apply constraints on distortion and what we have called (ϵ, δ)-continuity[28,37,108]. Some of these conditions are stronger than others, discussed in the Appendix.

It is all well and good to associate similarity between individuals with metrics. We may desire, however, to accommodate group fairness notions by thinking about formulations of "distances" between groups. For instance, these distances could quantify how dissimilar the aggregate outcomes assigned to different groups by an algorithm are, or how the data representations of groups differ. The following definitions introduce tools to make this idea exact by considering metric spaces additionally endowed with measures, borrowing from probability theory[37,89].

Definition 6. Given a set M and a set of its subsets Σ, a *probability measure* is a function $\mu : \Sigma \to [0, 1]$ satisfying $\mu(\emptyset) = 0$, $\mu(M) = 1$, and for any collection of disjoint sets $\{E_i\}$ in Σ, $\mu\left(\bigcup_i E_i\right) = \sum_i \mu(E_i)$. Note that μ is simply a formal representation of a probability distribution over M which maps subsets of M (events) to their probabilities. A *probability space* is a set M equipped with a probability measure μ, written (M, Σ, μ). If there is also a metric d on M, then M is a *metric probability space*, writable as (M, d, Σ, μ).[iv]

Definition 7. Given probability spaces (M_1, Σ_1, μ_1) and (M_2, Σ_2, μ_2), a *coupling measure* over the Cartesian product $M_1 \times M_2$ (with respect to μ_1 and μ_2) is a

[iii] The condition in the awareness paper explicitly notates the property is with respect to the metrics d_1, d_2 by writing it as the "(d_2, d_1)-Lipschitz property".

[iv] In order for a metric probability space to be well-defined, there are some technical conditions which we disregard, such as the fact that to make probabilities well-defined Σ should be a *σ-algebra* and M must be *compact*[1]. The foundations of probability carry over from measure theory; indeed, a probability space can just be viewed as a *measure space* with the restriction that the measure of the entire space is 1[103].

probability measure $\gamma : \Sigma_1 \times \Sigma_2 \to [0,1]$ such that for all $E \in \Sigma_1$ and $E' \in \Sigma_2$, $\gamma(E \times M_2) = \mu_1(E)$ and $\gamma(M_1 \times E') = \mu_2(E')$.

When we write "measure" in this thesis, we always implicitly mean "probability measure", and throughout we will interpret a probability measure μ as also denoting the distribution it represents. We will also keep Σ unwritten, as it describes a technical condition and does not contribute much to our idea. Usually there is a natural choice of Σ, such as when M is finite, in which case we choose $\Sigma = 2^M$.[v] The definition of a measure is also simplified when M is finite: a measure μ can be defined based on the probability "mass" it gives each element, provided that $\sum_{a \in M} \mu(a) = 1$[1].[vi] The probability measure of any set of elements is the sum of the masses of the elements. The constraint for a coupling measure γ on $(M_1, \mu_1) \times (M_2, \mu_2)$ then becomes $\sum_{m' \in M_2} \gamma(a, m') = \mu_1(a)$, and $\sum_{m \in M_1} \gamma(m, a) = \mu_2(a)$. This simpler definition is enough for us, as we disregard the theoretical infinite "universe" of individuals V and take M to be the finite set of individuals considered by an algorithm. Then we can think about distances between subsets in the context of those subsets representing groups of individuals.

Definition 8. Let (M, d) be a finite metric space. For two probability measures defined on M, $\mu_1, \mu_2 \in \Delta M$, the *Wasserstein distance* or *earthmover distance* between μ_1 and μ_2 is defined by $\mathcal{W}_d(\mu_1, \mu_2) = \min_{\gamma \in \Gamma} \sum_{a,b \in M} d(a,b)\gamma(a,b)$, where Γ is the set of all coupling measures over $M \times M$ with respect to μ_1 and μ_2.[vii]

[v] This is also the Σ implicit when talking about the set of probability distributions over a finite set A, ΔA, to make the distributions well-defined.

[vi] We are abusing notation slightly by writing for singleton sets $\mu(a) = \mu(\{a\})$.

[vii] In full generality, this Wasserstein metric is a discrete case of the p-Wasserstein metric

$\mathcal{W}_d$ is a metric on ΔM, so it is sometimes called the *Wasserstein metric*; a proof that it is a metric is given by Philippe and Wolfgang[23]. The Wasserstein distance extends naturally to describe distances between different metric spaces:

Definition 9. Let (M_1, d_1, μ_1) and (M_2, d_2, μ_2) be two finite metric probability spaces. The *Gromov-Wasserstein distance* between the spaces is defined by $\mathcal{GW}_d(M_1, M_2) = \min_{\gamma\in\Gamma} \sum_{(a,b)} \sum_{(a',b')} |d_1(a,b) - d_2(a',b')|\gamma(a,a')\gamma(b,b')$, where (a,b) and (a',b') are elements of $M_1 \times M_2$ and Γ is the set of all coupling measures over $M_1 \times M_2$ with respect to μ_1 and μ_2.

As it turns out, Mémoli shows that $\mathcal{GW}_d$ is a metric on the space of metric spaces, setting aside some technical details (e.g. up to isomorphism)[72].

Now we change our notation to specifically address our setup with a finite set of individuals X. Consider partitions of X into g groups, $X = \bigcup_{i=1}^{g} X_i$. ($X$ replaces M above, and the X_i are disjoint subsets.) Given a measure on X, the Wasserstein distance extends naturally to describe distances between groups.

Definition 10. Let X be a finite set with groups $\{X_i\}$. If X has a probability measure μ_X, then we define the *induced probability measure* for X_i, which is also a probability measure on X, as follows: for $x \in X$, $\mu_i(x) = \frac{\mu_X(x)}{\mu_X(X_i)}$ if $x \in X_i$ and 0 otherwise. If X does not have a measure, the induced probability measure is naturally defined by setting μ_X to be the uniform distribution that assigns weight $\frac{1}{|X|}$ to each element.

with $p = 1$: $\mathcal{W}_d(\mu_1, \mu_2) = \inf_{\gamma\in\Gamma} \left(\int_{a,b\in M} d(a,b)^p \gamma(a,b) \right)^{1/p}$. A similar generalization exists for Gromov-Wasserstein distances defined below that relies on the additional definition of metric couplings, analogous to (measure) couplings as we have defined them; see Mémoli[72].

Remark 1. Let X_i, X_j be groups in X. As a slight abuse of notation, we define $\mathcal{W}_d(X_i, X_j)$ to mean $\mathcal{W}_d(\mu_i, \mu_j)$, where μ_i and μ_j are the induced probability measures on X_i and X_j.

The takeaway is that we can define Wasserstein distances between groups using a measure on X. In the case that no measure is specified, the μ_i defined above represent uniform distributions supported on each X_i, like indicators for group membership; this is implicitly what we mean when we describe Wasserstein distances between subsets when no measure is specified.[viii]

There is an important distinction between how Friedler et al. and Dwork et al. use $\mathcal{W}_d$. For the latter, since they consider algorithms with $R = \Delta A$, there is a way to define a probability measure for X_i over A instead of X, based on the average distribution assigned to individuals in X_i by an algorithm.

Definition 11. Let X be a finite set with groups $\{X_i\}$ and let f be a map from X to ΔA. The *outcome probability measure* for X_i, which is a probability measure over A, is the average of outcomes over X_i, $\mu_i^{out} = \frac{1}{|X_i|} \cdot \sum_{x \in X_i} f(x)$.

We verify that outcome probability measures are probability measures in the Appendix. Outcome probability measures can be found directly from f; X need not have a metric or probability measure. Also, outcome probability measures lend themselves exactly to the fairness definition of statistical parity, by definition;

[viii] While our presentation only considers finite sets, these definitions also extend to a general case where X is continuous but measurable. For instance, we could conceive of X as a compact subset of $\mathbb{R}^n$, in which X_i are further subsets of X; then for an arbitrary subset $S \subseteq X$, $\mu_i(S) = \frac{\mu_X(S \cap X_i)}{\mu_X(X_i)}$. However, in continuous settings it is harder to think about individuals, since single points in continuous sets have measure zero. We stick with the more tangible case of having X finite for simplicity.

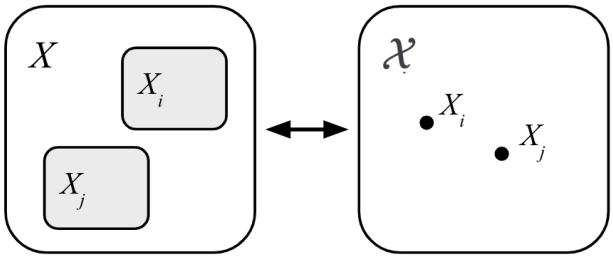

Figure 2.1: An illustration of the group space $\mathcal{X}$ that arises from the space of representations of individuals X with a partition into groups. Groups in X are elements in $\mathcal{X}$.

Dwork et al. define (ϵ-)statistical parity between two groups X_1 and X_2 as the condition $D_{tv}(\mu_1^{out}, \mu_2^{out}) < \epsilon$. By contrast, since induced probability measures are over X they have no direct link to statistical parity.

Finally, Friedler et al. consider, given $X = \bigcup_{i=1}^{g} X_i$, the set of groups, $\mathcal{X} = \{X_1, \ldots, X_g\}$. Each element in $\mathcal{X}$ constitutes a group in X, so we identify X_i as a subset of X as equivalent with X_i as a point in $\mathcal{X}$.

Proposition 1. *Suppose X has metric d_X and probability measure μ_X. Then the group space $\mathcal{X}$ as defined above is a metric probability space, with respect to the metric $d_{\mathcal{X}}(X_i, X_j) = \mathcal{W}_d(X_i, X_j)$ and the naturally-induced probability measure $\mu_{\mathcal{X}}$ defined by $\mu_{\mathcal{X}}(X_i) = \mu_X(X_i)$.*[ix]

Proof. See Appendix. □

We call $\mathcal{X}$ the *group space*, as compared to the individual space X; see Figure 2.1. Given a metric, probability measure, and group partition on X, $\mathcal{X}$ is induced (can be defined). Finally, consider two metric probability spaces (X, d_X, μ_X) and

[ix] While Friedler et al. do not explicitly define μ_X explicitly, this accords with the "natural way" to define a measure on $\mathcal{X}$.[37]

(Y, d_Y, μ_Y) with a map $m : X \to Y$, which we can interpret as a mapping from one representation of individuals to another. Assume that each individual in X can be identified with a counterpart in Y, that is, there is a bijection $X \to Y$. Then groups can also be consistently identified in both X and Y (X_i is mapped to Y_i), and we can compare the group spaces $\mathcal{X}$ and $\mathcal{Y}$.

Definition 12. Given two metric probability spaces (X, d_X, μ_X), (Y, d_Y, μ_Y) with g groups (i.e. subsets) and group spaces $\mathcal{X}$ and $\mathcal{Y}$, the *between-groups distance* is defined as $\rho_b(\mathcal{X}, \mathcal{Y}) = \frac{\mathcal{GW}_d(\mathcal{X},\mathcal{Y})}{\binom{k}{2}}$. The *within-group distance* is defined as $\rho_w(\mathcal{X}, \mathcal{Y}) = \frac{1}{g}\sum_{i=1}^{g} \mathcal{GW}_i(X_i, Y_i)$, viewing X_i as a metric probability space with metric d_X and the induced probability measure μ_i from μ_X, and similarly for Y_i. The *group skew* between $\mathcal{X}$ and $\mathcal{Y}$ is $\sigma(\mathcal{X}, \mathcal{Y}) = \frac{\rho_b(\mathcal{X},\mathcal{Y})}{\rho_w(\mathcal{X},\mathcal{Y})}$. In the case that $\rho_w(\mathcal{X}, \mathcal{Y}) = 0$ (when $\mathcal{X}$ and $\mathcal{Y}$ have identical structure), $\sigma(\mathcal{X}, \mathcal{Y})$ can be computed by adding $O(\delta)$ random noise to both numerator and denominator.

The intuition here is that $\mathcal{W}_d$ and $\mathcal{GW}_d$ quantify how "different" groups are in the context of metric probability spaces. The Wasserstein distance is called the earthmover metric because we can roughly visualize a probability distribution defined by a probability measure as a pile of earth; the metric then quantifies the minimum amount of "work" needed to complete the task of moving earth from one distribution to another, where the work factors in the amount of earth (expressed by the probability masses of the measure to be moved) and the distance moved (expressed by the metric).[x] While more abstract, the Gromov-Wasserstein

[x] In fact, the Wasserstein metric comes up in an applied mathematics problem expressing exactly this idea, the optimal transport problem, which can be solved via linear programming techniques[73].

distance carries the same idea. If two groups are "close" in a space, then the "work" done to move between the distributions is smaller as conveyed by $\mathcal{W}_d$, and similarly so when considering how "close" the representations of a group are in different spaces as conveyed by $\mathcal{GW}_d$.

Dwork et al. and Friedler et al. use the Wasserstein distance in categorically different situations. The former apply it in determining the degree to which an algorithm has "separated" groups in terms of the probability distributions the algorithm outputs, while the latter use it when comparing distances between groups across different spaces of representations. Particularly, they consider a bijective observation process, $m : X \to Y$, which modifies representations of individuals. Their definitions of between-groups and within-groups distances represent how the process globally distorts groups with respect to each other and the individual distortion of each group when going from X to Y. Friedler et al. take high group skew to be a sign that f treats groups more differentially since the overall distortion between groups is disproportionately high, thus possibly indicating bias in the observation process[37].

Before applying these concepts, we shift our attention to an empirical discussion of differential expressiveness.

2.2 Defining Differential Expressiveness

In this section we give a formal characterization of what we mean by differential expressiveness (DE). To our knowledge, while the concept behind DE is intuitive and widespread, no work in the literature thus far specifically describes it. One

overarching comment to make about DE is that it is a "local" concept, in that it only examines single features in isolation. Sources of bias can and do manifest in multiple features, and simply considering the values within one feature reveal only a fraction of the bigger picture. (In part, this is why more sophisticated mathematical tools like the ones defined above are introduced, in attempts to describe a more encompassing worldview.) Furthermore, we make no claims that our treatment of the idea of DE is comprehensive; it is certainly feasible to think there are more rigorous mathematical characterizations waiting to be written on this topic. Nonetheless, we hope our discussion lends itself to a more nuanced approach to how we think about the features we must choose to constitute human representations whenever we use an algorithm.

Consider the input data for an algorithm f to be the finite set of individuals $x \in X$, where each individual is represented by k abstract features in the data $x = (x^1, \ldots, x^k)$. The focus of DE is characterizing data bias, i.e. problems inherent in the k features that portray each individual, independent of what f may be. DE specifically focuses on individual features within the feature vector that serves as a representation of each individual x. As another minor abuse of notation, we associate a feature with its index within feature vectors. That is, if we use θ to denote a feature (e.g. race, income, etc.) we also treat θ as an index for the feature it represents, θ means the θ-th feature, x^θ, for all x. Let Θ be the set of values θ can take, e.g. a set of races $\{\texttt{white}, \texttt{black}, \ldots\}$ if θ is race, or $\mathbb{R}_{\geq 0}$ if θ is income. Recall our rough definition of DE from the introduction:

Definition 13. (DE, informal definition). A feature θ of a dataset of individuals

X is *differentially expressive* if its values cannot be consistently interpreted among different individuals in X.

There are two main senses in which we can characterize θ not being consistently interpretable. First, θ might take on systemically different values depending on the individual when we have no reason to expect those systemic differences to exist. Second, the same value of θ might mean categorically different things depending on the individual, and thus judging two individuals based on a value of θ in the same way would be unfaithful. While these characterizations might conceivably apply on the level of individuals,[xi] the fact that we want to discuss systemic differences manifestly lends itself to talk about DE in the context of groups (where the way we think about θ is different in each group) by considering the distribution of values of θ in that group. As shorthand, we refer to the distribution of values of a feature θ as just "the distribution of θ". Now we may more precisely say of these two characterizations that the first describes a situation where the distribution of θ within one group systemically differs from the rest of the population, and the second describes a situation where the members of a group interpret θ in a different way from others. There is a kind of symmetry between these two characterizations. In the first, we imagine that the values of θ should be distributed approximately equally between groups, but in reality they are not. In the second, we imagine that the meaning of θ is categorically different between groups, though values of θ could be distributed approximately equally between groups. In essence, there is

[xi] For instance, we could think of a scenario in the second characterization in which θ means something different to *everyone* in X, e.g. if some adversary has tainted the data collection process such that different data is presented the same way to the data curator. However, this example is a generic one reflecting issues with methodology.

a disconnect between a value we see of θ and the meaning of what we believe θ ought to really indicate.

Now we can be more precise in defining these two characterizations. We name them, respectively, *distributional DE* and *semantic DE*. The first relies on us having some way to quantify the extent to which two distributions are different. There are several different ways to do this, but for our definition we are agnostic to the particular choice. Additionally, in real-life situations we often encounter datasets with missing data, in which case we need some way to represent a value for θ corresponding to missing data. If Θ is the set of values that θ canonically takes, as shorthand we let $\tilde{\Theta} = \theta \cup \{\emptyset\}$, where $\emptyset$ represents the special case that the value of the feature is missing. A model might specify that feature values are Θ, whereas $\tilde{\Theta}$ is what we might see in practice with real data. Finally, in our definitions we only consider the case of two groups $X = X_1 \cup X_2$ for simplicity, but our definitions of DE can also apply to comparisons between two groups X_i and X_j, or between a group and the rest of the population X_i (in which case we can just set $X_2 = X \setminus X_i$).

Definition 14. Let θ be a feature in our set of representations of individuals X taking on possible values Θ in principle. Let $\tilde{\Theta}$ be the set of possible values of θ in practice as defined above. Let X_1, X_2 be two disjoint groups such that $X = X_1 \cup X_2$. Then for $i \in \{1, 2\}$, we define $\mu_i^\theta \in \Delta\tilde{\Theta}$ to be the (measure for the) *empirical distribution of* θ within X_i.

In the definition above we refer to the standard empirical distribution from probability. Suppose first that Θ is finite, which captures a wide variety of features

in practice. Some examples are exam scores, binary indicators, and discretizations of continuous measurements, e.g. where instead of raw income we work with income brackets or "bins". In this case, the empirical distribution is defined as, for $t \in \tilde{\Theta}$ as $\mu_i^\theta(t) = \frac{1}{|\tilde{\Theta}|} \sum_{x \in X_i} I_{\{x^\theta = t\}}$. In the case that Θ is infinite (e.g. $\Theta = \mathbb{Z}$ or contains real numbers), we can think of μ_i^θ as a mixed distribution where $\theta = \emptyset$ with probability $\frac{1}{|\tilde{\Theta}|} \sum_{x \in X_i} I_{\{x^\theta = \emptyset\}}$ and otherwise the distribution of θ is governed by its empirical cumulative distribution function over Θ.[103]

Distributional DE. Our first characterization of DE is essentially distributions of θ being different when they ought not to be. If this is the case, then something is wrong with our assumption, and that there may be a disconnect between our conception of the feature and how it actually manifests in groups. We can now describe this in mathematical language.

Definition 15. (DE, characterization 1.) Given a function D that measures differences between distributions $D : \Delta\tilde{\Theta} \times \Delta\tilde{\Theta} \to \mathbb{R}$, we say θ is (D, ϵ)-*distributionally differentially expressive* (distributionally DE) if $D(\mu_1^\theta, \mu_2^\theta) > \epsilon$.

Another suitable name might be "statistical DE". Choices for D include distance functions between distributions—e.g. D_{tv} or D_∞ as given in the awareness paper[28]. Picking an appropriate ϵ requires knowledge of D, since the ranges of these metrics are different. There are several other functions which can be used to quantify statistical distance, or metrics between distributions; see the Appendix.[xii]

Other choices for D could be adopted from the statistical literature. As an

[xii] For instance, while KL divergence is not a metric because it is asymmetric, there might be situations in which we want to exploit this property, by measuring the divergence of some μ_i^θ *relative to* the rest of the population.

example, suppose we make the modelling assumption that each μ_i^θ should follow some parameterized distribution $\mathcal{D}(\phi)$ (i.e. θ in each group follows a distribution family $\mathcal{D}$ over Θ), such as a Normal distribution over the real numbers or an interval thereof, or a Poisson distribution over the positive integers. Then when analyzing the distribution of θ between two groups, we could narrow our focus to consider only the x where $x^\theta = \emptyset$; fit relevant parameters to the empirical distributions of θ over the two groups to get estimates of the parameters ϕ_1, ϕ_2 of those distributions; and evaluate some notion of the distance between these two sets of parameters, such as via the wealth of statistical tests using p-values that exist for testing equality of distributions. The modelling assumption that each μ_i^θ follows a distribution family can be a strong one, however, and this method ignores the empty values $\emptyset$, which may be an important consideration. To be more general, other well-known statistical tests for equality of distribution that are agnostic to a particular choice of distribution, such as the Kolmogorov-Smirnov test, which can also be extended to support mixed distributions such as $\tilde{\Theta}$.[103] Since our discussion here remains in the abstract, we do not explore the mathematical details and results that follow from particular choices of D, but such an inquiry may prove to be a promising line of future theoretical work.

In some cases, we might be concerned with the applied problem for which the availability of a feature is unequal for different groups, i.e. the prevalence of the empty value $\emptyset$ is disproportionately high in one group. This might be indicative of pervasive bias in the data collection process whereupon one group might be underrepresented in existing circumstances from which measurements

of the feature are made; thus, the fact that *existence* of data for θ itself is not guaranteed to be consistent between groups is problematic. This is even more fundamental than issues pertaining to the distribution of θ, since it tells us that the process of even getting values for θ is flawed. However, we remark below that we can view this as a kind of distributional DE.

Definition 16. Let PE_1^θ and PE_2^θ denote the proportion of empty values of θ in X_1 and X_2, respectively: $PE_i^\theta = \frac{1}{|X_i|}\sum_{x\in X_i} I_{\{x^\theta=\emptyset\}}$. Then we say θ is ϵ-*existentially differentially expressive* (existentially DE) if $|PE_1^\theta - PE_2^\theta| > \epsilon$.

This is a simple definition that directly compares the proportions of empty values of θ. A weakness of this definition is that when the sizes of X_i are small, our values of PE_i^θ might be noisy due to small sample variability. In the Appendix we give a definition adapted from statistics that uses the two-proportion test to account for this. *Nota bene* that we can frame this "data existentiality" definition in terms of the distributional DE definition by specifically defining D as $D(\mu_1^\theta, \mu_2^\theta) = |\mu_1^\theta(\emptyset) - \mu_2^\theta(\emptyset)|$. So we view existential DE as a special case of distributional DE, rather than a wholly separate characterization.

Semantic DE. Now consider our second characterization of DE. To express the discrepancy between the values of θ that we observe in X versus the actual meaning of the values to individuals, we can conceive of θ as being a proxy measurement for some underlying feature, θ^*, that exists but is hidden. Let θ^* take values in Θ^*, which we assume does not contain empty values (θ^* will always be well-defined for an individual). For instance, θ^* could be "academic talent" and θ could be "GPA". We really mean to measure the feature θ^* but observe θ. We can

imagine some observation process that translates and crystallizes values of θ^* into observed values of θ as collected in X. This process might have statistical noise such that the same value of θ^* could materialize as multiple values of θ. More importantly, θ might be "coarse" and destroy information about θ^*, and individuals might view θ differently than they view θ^*, causing a semantic discrepancy. To be clear with notation, θ^* and θ are features; $t^* \in \Theta^*$ and $t \in \tilde{\Theta}$ are values the features can take.

We assume that all members of a particular group possess the same interpretative relationship between values of θ^* and θ. Specifically, let R_i^θ be the binary relation between Θ^* and $\tilde{\Theta}$ that associates, for members of X_i, values of θ^* and θ that are regarded as equivalent for the group. In other words, if $(t^*, t) \in R_i^\theta$, then for members of X_i, $\theta^* = t^*$ means the same thing as $\theta = t$.[xiii] We call R_i^θ the *semantic relation* of X_i. Now define for $t^* \in \Theta^*$ the *(semantic) image* of t^* as $R_i^\theta(t^*) = \{t \in \tilde{\Theta} : (t^*, t) \in R_i^\theta\}$. Similarly, define the *(semantic) preimage* of $t \in \tilde{\Theta}$ as $(R_i^\theta)^{-1}(t) = \{t^* \in \Theta^* : (t^*, t) \in R_i^\theta\}$. Here, $R_i^\theta(t^*)$ is the set of observed values of θ manifested by the unseen meaning $\theta^* = t^*$, while $(R_i^\theta)^{-1}(t)$ is the set of unseen meanings of θ^* corresponding with the observation $\theta = t$, according to members of X_i.[xiv] See Figure 2.2.

[xiii] We specify R_i^θ as a binary relation to capture a general sense of mappings between values. It is reasonable to think that due to noise or randomness in the process that transforms the feature θ^* into the feature θ, one value of θ^* corresponds to multiple values of θ—and vice versa: one value of θ could possibly mean many different things in truth (i.e. in terms of θ^*) to X_i and these several meanings have been conglomerated in our proxy, thus destroying information about the original expressiveness and variation of θ^*. This also accommodates the restricted setting where each value t of θ corresponds to one value of θ^*, such that there is a single "true" meaning of t for members of X_i, i.e. when R_i^θ is a function; and vice versa, when R_i^θ is injective.

[xiv] Our definitions have been in terms of $\tilde{\Theta}$ rather than Θ to also accommodate missing values. If missing data is a comprehensive issue that affects all individuals, it might be the case that

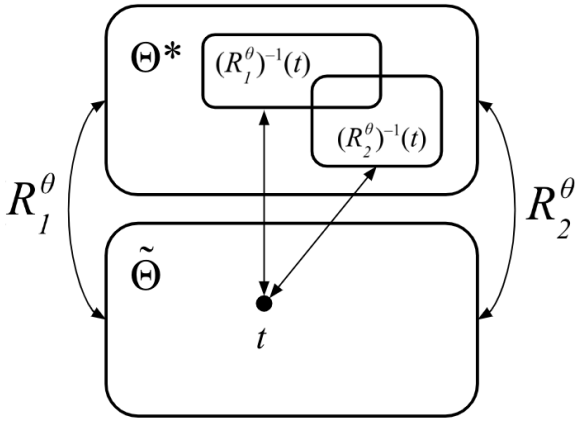

Figure 2.2: An example of the semantic relations R_1^θ and R_2^θ, which associate semantically equivalent values in Θ^* and $\tilde{\Theta}$ for each group. The preimages of one value t of θ could be multiple values in Θ^*, and could differ between groups (the root cause of semantic difference).

Our intuitive definition of "values of θ meaning different things to different groups" could take the form of the following condition: *there exists some* $t \in \tilde{\Theta}$ *such that* $(R_1^\theta)^{-1}(t) \neq (R_2^\theta)^{-1}(t)$. However, this is an incredibly weak condition, since it only requires that there is at least one value where the preimages are not identical. Especially if Θ^* and $\tilde{\Theta}$ are not finite sets, e.g. if they are intervals of real numbers, this also does not capture the idea that the extent to which preimages differ can vary. Certainly, having $(R_1^\theta)^{-1}(t)$ and $(R_2^\theta)^{-1}(t)$ differ by one element for one particular t is less concerning, as far as semantic difference is concerned, than having $(R_1^\theta)^{-1}(t)$ and $(R_2^\theta)^{-1}(t)$ be disjoint for all values of t. Hence, like our definition of distributional DE above, we suppose the presence of

$(R_i^\theta)^{-1}(\emptyset) = \Theta^*$ for all i. However, there could be scenarios where the extent to which we have missing data depends on the group, e.g. if less data is available for a disadvantaged group to evaluate because they have been methodologically overlooked, or if a more advantaged group has the power to purposefully redact data. Differential circumstances of data provenance and control could make the observation of a missing value, $\theta = \emptyset$, also mean different things based on the group!

some generic function $\mathfrak{D}$ that quantifies differences between subsets of Θ^*. Now we can precisely state our second characterization:

Definition 17. (DE, characterization 2.) Given a function $\mathfrak{D}$ that quantifies differences between sets of possible values of θ^*, $\mathfrak{D} : 2^{\Theta^*} \times 2^{\Theta^*} \to \mathbb{R}$, we say θ is $(\mathfrak{D}, \epsilon)$-*semantically differentially expressive* (semantically DE) if

$$\sum_{t \in \tilde{\Theta}} \mathfrak{D}\left((R_1^\theta)^{-1})(t), (R_2^\theta)^{-1}(t)\right) > \epsilon$$

if $\tilde{\Theta}$ is discrete, and

$$\int_{t \in \tilde{\Theta}} \mathfrak{D}\left((R_1^\theta)^{-1})(t), (R_2^\theta)^{-1}(t)\right)\, dt > \epsilon$$

otherwise.

Intuitively, we consider, for all possible values t of the proxy feature θ, the difference in the meaning of $\theta = t$ between the two subsets. If the aggregate "difference in meaning" over all t is large enough, we say θ is semantically DE.

Remark 2. We could generalize the definition of the semantic relation by saying there is a probabilistic process generating values of θ from values of θ^*, such that $(R_i^\theta)^{-1}$ is a *distribution* over Θ^*, rather than a set (under that perspective, our current definition associates a value of θ, $t \in \tilde{\Theta}$, with a uniform distribution over the set $(R_i^\theta)^{-1}(t)$).

Later in this thesis, we will also consider the fact that relationships between different θ^* and θ could be many-to-many, and our definition of R_i^θ is unable to capture this notion. An intriguing line of research might be to extend the concept of the semantic relation to capture these generalizations.

As with D in the definition of distributional DE, the choice of $\mathfrak{D}$ is purposefully left broad. A simple definition could be, given two subsets S_1, S_2 of Θ^*, $\mathfrak{D}(S_1, S_2) = m(S_1 \Delta S_2)$, where Δ here means the symmetric difference and m measures the size of the set (e.g. cardinality if the sets are finite). However, this ignores how far apart the elements exclusive to S_1 are from the elements exclusive to S_2.[xv] If Θ^* is endowed with a metric, an alternate choice for $\mathfrak{D}$ could be the Hausdorff distance, commonly used as a measure of how different two sets are[50]. Again, we will not delve into results from choosing particular $\mathfrak{D}$, but such an exploration would be a natural extension of this idea.

We can get a more concrete definition of semantic DE if we add a restriction that R_i^θ should be injective, so that every value $t \in \tilde{\Theta}$ is associated with one value $t^* \in \Theta^*$;[xvi] then for all t, t^* is the unique "true meaning" of $\theta = t$ for individuals in X_i. In this case, we can write a simpler definition of semantic DE that does not have to consider differences of sets of values in Θ^*:

Definition 18. Suppose that R_i^θ is injective, such that $(R_i^\theta)^{-1}(t)$ is a singleton set $\{t^*\}$ (or empty).[xvii] Identify $(R_i^\theta)^{-1}(t)$ with this value t^* if it exists. Given a function $\mathfrak{d}$ that quantifies differences between possible values of θ^*, $\mathfrak{d} : \Theta^* \times \Theta^* \to$

[xv] For instance, suppose $\Theta^* = \mathbb{N}$ represents "level of talent", $S_1 = \{1, 2, 3, 4\}$, $S_2 = \{1, 2, 3, 5\}$, and $S_3 = \{1, 2, 3, 10\}$. It is natural to think that S_3 and S_1 are "further apart" than S_1 and S_2, but the definition of $\mathfrak{D}$ does not capture this.

[xvi] E.g. by assuming we are in the special case where there are no missing values, so that $(R_i^\theta)^{-1}(\emptyset) = \{\}$.

[xvii] The preimage of t could be empty since we do not stipulate that R_i^θ be surjective. In that case, we can just disregard that value of t.

$\mathbb{R}$, we say θ is $(\mathfrak{d}, \epsilon)$-semantically DE if

$$\sum_{t \in \tilde{\Theta} \,:\, \text{both } (R_i^\theta)^{-1}(t) \text{ exist}} \mathfrak{d}\left((R_1^\theta)^{-1}(t), (R_2^\theta)^{-1}(t)\right) > \epsilon$$

if $\tilde{\Theta}$ is discrete, and

$$\int_{t \in \tilde{\Theta} \,:\, \text{both } (R_i^\theta)^{-1}(t) \text{ exist}} \mathfrak{d}\left((R_1^\theta)^{-1}(t), (R_2^\theta)^{-1}(t)\right)\, dt > \epsilon$$

otherwise.

As injective relations are relations, this is a special case of Definition 17.[xviii]

These characterizations of DE are not mutually exclusive. A feature θ could be distributionally DE as measured between groups, while also being an unfaithful proxy for some underlying feature. As long as we think of θ as a proxy, we can evaluate θ using D based on the data we see while also conceiving of some $\mathfrak{D}$ (or $\mathfrak{d}$) to hypothetically compare and contrast θ against an underlying feature. One way to give a unified relationship between these forms of DE is as follows: whereas distributional DE describes what is problematic about the features we observe at face value, semantic DE concerns what might be problematic about an unseen process that produces these features. In particular, distributional DE might be a consequence of semantic DE, if the reason distributions of a feature differ between groups is that the groups interpret the feature differently. We can formalize this idea by saying that though θ^* might have the same distribution in all groups, the R_i^θ distort those distributions to cause distributional DE in θ.

[xviii] Notice that all functions are relations: $g : \Theta^* \to \Theta$ can be written as the relation $\{(t^*, t) : g(t^*) = t\}$. A further restriction we could make is specifying that R_i^θ be an injective function, which is easier to conceptualize. However, simply requiring that R_i^θ be a function from Θ^* to $\tilde{\Theta}$ is not sufficient, as preimages might not be singleton sets. E.g. if $\Theta^* = \Theta = \mathbb{R}$ and R_i^θ is the squaring function, we have $(R_i^\theta)^{-1}(1) = \{-1, 1\}$.

Proposition 2. *Suppose X_1 and X_2 share the same distribution for values of the underlying feature θ^*, $\mu^* \in \Delta\Theta^*$.*[xix] *Then the empirical distributions of θ for each group are determined by their semantic relations. Specifically, μ_i^θ is defined for all $t \in \tilde{\Theta}$ as $\mu_i^\theta(t) \propto \mu^*\left((R_i^\theta)^{-1}(t)\right)$. In the opposite direction, the following relationship holds: for all i and $t^* \in \Theta^*$, $\mu^*(t^*) \propto \mu_i^\theta\left(R_i^\theta(t^*)\right)$ (assuming $\tilde{\Theta}$ is discrete; otherwise, the above holds with the sum replaced by an integral).*

At this point we have introduced several measures on several different spaces; a breakdown is given in the Appendix. We can equivalently write $\mu_i^\theta(t) \propto \sum_{t^* \in \Theta^*} \mu^*(t^*) I_{\left\{t^* \in (R_i^\theta)^{-1}(t)\right\}}$ and $\mu^*(t^*) \propto \sum_{t \in \tilde{\Theta}} \mu_i^\theta(t) I_{\left\{t^* \in (R_i^\theta)^{-1}(t)\right\}}$, observing that the events $\{t \in R_i^\theta(t^*)\}$ and $\{t^* \in (R_i^\theta)^{-1}(t)\}$ are equal. The statements above use proportionality constraints rather than equalities, because we need to normalize to get a measure. In particular, we can write

$$\mu_i^\theta(t) = \frac{\mu^*\left((R_i^\theta)^{-1}(t)\right)}{\sum_{t' \in \tilde{\Theta}} \mu^*\left((R_i^\theta)^{-1}(t')\right)};\ \mu^*(t^*) = \frac{\mu_i^\theta\left(R_i^\theta(t^*)\right)}{\sum_{t^{*\prime} \in \Theta^*} \mu_i^\theta\left(R_i^\theta(t^{*\prime})\right)}$$

swapping the sum with an integral if needed. Proofs follow from definitions: $\theta = t$ iff $\theta^* \in (R_i^\theta)^{-1}(t)$, so the overall probability mass put by μ_i^θ on t should be proportional to the sum of the probability masses put by μ^* over the set $(R_i^\theta)^{-1}(t)$. The opposite direction proceeds analogously.

There is a duality in how semantic relations are used when thinking about the use of proxy features. In the "forward direction", they give one explanation for how distributional DE arises as described above. In the "backward direction", we get constraints on how R_i^θ and μ_i^θ are related.

[xix] As a technical note, if $|X_1| \neq |X_2|$ then it may not be possible for the groups have identical empirical distributions. We can think of μ^* as being a common data distribution from which "crystallized" values of θ^* would be drawn across the two groups, if we could observe θ^* directly.

Theorem 1. *(Distributional DE arising from semantic DE) Suppose θ is a proxy feature for the underlying feature θ^*, and the underlying distribution $\mu^* \in \Delta\Theta^*$ is the same across groups. Let $D : \Delta\tilde{\Theta} \times \Delta\tilde{\Theta} \to \mathbb{R}$ be a function that measures differences of distributions "atomically", i.e. given distributions $\mu_1^\theta, \mu_2^\theta \in \Delta\tilde{\Theta}$, we can write $D(\mu_1^\theta, \mu_2^\theta) = \sum_{t\in\tilde{\Theta}} d(\mu_1^\theta(t), \mu_2^\theta(t))$ for some function $d : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, replacing the sum with an integral if appropriate (e.g. if $\tilde{\Theta}$ is not discrete).*

Let $c_i = \sum_{t\in\tilde{\Theta}} \mu^*\left((R_i^\theta)^{-1}(t)\right)$. Then θ is (D, ϵ)-distributionally DE if and only if θ is $(\mathfrak{D}, \epsilon)$-semantically DE, where $\mathfrak{D} : 2^{\Theta^*} \times 2^{\Theta^*} \to \mathbb{R}$ is defined as $\mathfrak{D}(S_1, S_2) = d\left(\frac{1}{c_1}\mu^*(S_1), \frac{1}{c_2}\mu^*(S_2)\right)$ given arbitrary subsets $S_1, S_2 \subseteq \Theta^*$.

Proof. To first gain some intuition, note that we have chosen D and $\mathfrak{D}$ to accord with the definitions of distributional and semantic DE above. When we say D measures differences of distributions atomically, we mean that the way it computes differences between two probability measures representing distributions is by summing over a function d applied to the values of those measures when evaluated on single elements in $\tilde{\Theta}$. For instance, if $\tilde{\Theta}$ is finite, then the total variation norm D_{tv} is atomic with $d(a, b) = \frac{1}{2}|a - b|$, since to compute D_{tv} is just to sum the absolute differences of probability masses. However, the relative ℓ_∞ metric is not atomic, since it takes a maximum over events. If $\tilde{\Theta}$ is continuous, then D could be computed by integrating some function of probability densities over all possible values in $\tilde{\Theta}$.

Suppose θ is (D, ϵ)-distributionally DE. Observe that

$$\begin{aligned} D(\mu_1^\theta, \mu_2^\theta) &= \sum_{t \in \tilde{\Theta}} d\left(\mu_1^\theta(t), \mu_2^\theta(t)\right) \\ &= \sum_{t \in \tilde{\Theta}} d\left(\frac{1}{c_1}\mu^*\left((R_1^\theta)^{-1}(t)\right), \frac{1}{c_2}\mu^*\left((R_2^\theta)^{-1}(t)\right)\right) \\ &= \sum_{t \in \tilde{\Theta}} \mathfrak{D}\left((R_1^\theta)^{-1}(t), (R_2^\theta)^{-1}(t)\right). \end{aligned}$$

where we applied Proposition 2 in the second line and our definition of $\mathfrak{D}$ in the third line. Thus the condition $D(\mu_1^\theta, \mu_2^\theta) > \epsilon$ is equivalent to the condition $\sum_{t \in \tilde{\Theta}} \mathfrak{D}\left((R_1^\theta)^{-1}(t), (R_2^\theta)^{-1}(t)\right) > \epsilon$. These are, respectively, the definitions of (D, ϵ)-distributional DE and $(\mathfrak{D}, \epsilon)$-semantic DE. If $\tilde{\Theta}$ is not discrete we can just replace the sum with an integral. □

This theorem puts into mathematical language what we just described: when θ is a proxy for θ^*, we can link our two DE definitions by regarding differences in the feature distributions μ_i^θ as a consequence of semantic differences as conveyed through the R_i^θ. Figure 2.3 gives an illustration of this idea.

To summarize, we have just presented a mathematical treatment of DE. Instead of the concepts in the previous section, our formulation examine distributions of individual features. Distributional DE goes along with a description of the world we see, whereby a feature means the same thing among all groups but the data we have for the feature is unexpectedly distributed differently, or does not exist, for different groups. Semantic DE is a more normative definition, where, regardless of the distribution of a feature, it cannot be interpreted consistently because how it is interpreted by different groups is different—we need context (in

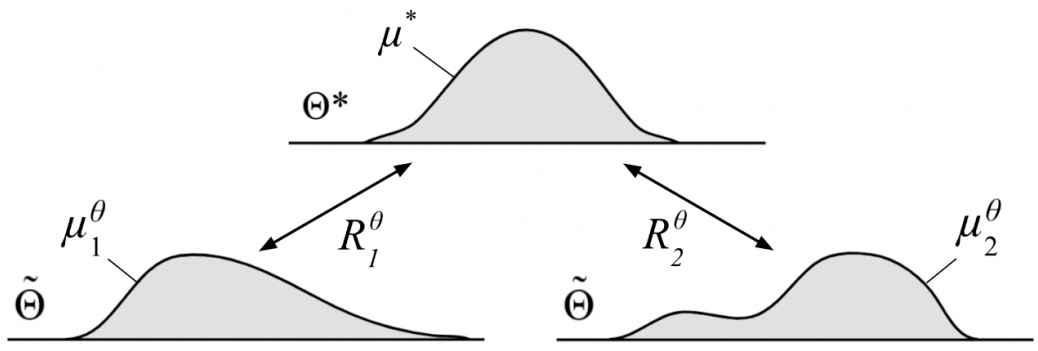

Figure 2.3: A illustration casting distributional DE as a consequence of semantic DE. Assume here for sake of illustration that Θ^* and $\tilde{\Theta}$ are one-dimensional and continuous. The diagrams represent histograms of values. We take the underlying feature in Θ^* to be distributed equally, following μ^*, among two groups X_1 and X_2. However, the ways in which these groups interpret the proxy in Θ differ, which leads to different empirical distributions μ_1^θ and μ_2^θ.

the form of R_i^θ) to understand the feature thoroughly. We also showed that in some cases we can frame distributional DE as a consequence of semantic DE.

It appears that semantic DE is a more versatile and general conception to work with. However, there are two snags. First, a feature θ might not be a proxy for anything—it could just represent the base truth, such as "height" or "weight", making the semantic setup unnecessary. Second, while we can diagnose distributional DE from X in practice—indeed, we just need to look at the empirical distributions of θ across groups—it is harder to empirically reckon with semantic DE by definition, because the reason we use proxy features is likely because the underlying feature θ^* is impossible to measure itself. While we can talk about R_i^θ in theory, finding it in practice might be intractable because of the broadness of the constraints linking μ^*, μ_i^θ, and R_i^θ above; namely, R_i^θ is underspecified. Because of this epistemic limitation, it may be more helpful to think of semantic DE as a more

theoretical tool, while thinking of distributional DE as a diagnostic measurable tool to assess our expectations that a feature ought to be distributed similarly across groups. As we have said in the Introduction, since these conditions are properties of data itself, they can be considered before factoring in the algorithm. If data is differentially expressive, than the algorithm will likely inherit the bias and misunderstanding that DE entails.

2.3 Differential Expressiveness in the Wild

DE is far from just a formal proposition; it provides us with a new perspective for thinking about several instances of data bias in the real world. Here we consider case studies and hypothetical examples of bias in which our notion of DE serves as an insightful lens. We group our discussions into broad thematic categories to illustrate the idea that DE as a phenomenon of data bias is quite widespread, underscoring a need to think critically about the origin and significance of the data we gather for any algorithm.

Hiring. We start out by building on an example in the awareness paper and framing it as an instance of semantic DE. Consider a world with two groups X_1 and X_2 of students belonging to different cultures. Suppose X_1 is a minority group, and that in their culture the most talented students are steered toward STEM subjects at a higher rate than the overall population in X_1. Conversely, suppose that in X_2 the most talented students are steered by cultural norms toward finance and economic at a higher rate than the overall population X_2. The overall level of engagement with STEM courses among both groups is equal. Consider the task

of a consulting firm wishing to hire prospective graduates. Wishing to attract the top talent, they notice that among the overall population X, the students showing the most promise study finance and economics since X_2 is the majority group. Thus, they choose the feature θ to be "level of engagement with finance and economics coursework" as a proxy for talent, used in screening candidates.

Evidently this is suboptimal; the firm should really be selecting for STEM majors among group X_1. (If culture is, however, a protected feature, we could imagine that, adopting a true fairness through unawareness approach, the firm has no way of differentiating between groups, locking in this suboptimality.) This scenario provides a clear illustration of θ being semantically DE, because the meaning of "person x studies economics" very clearly differs depending on the group x is in. Moreover, in this scenario, semantic DE does not cause distributional DE, as we specified above that the distribution of this feature in each group is equal.

We may describe this situation using the mathematical framework above. Let θ^* express talent, and suppose $\Theta^* = \Theta = \{0, 1, \ldots, 10\}$. Suppose we do not have problems with empty values, so we only need to think about Θ and not $\tilde{\Theta}$. Let a higher value of θ^* indicate a higher level of talent. Suppose the students are at a strange institution where they must take exactly 10 courses in total, with the only options being STEM and finance/economics courses, and let θ be the number of finance and economics courses taken. Let μ^* be the uniform distribution over Θ^*, and define the semantic relations as follows: $R_1^\theta = \{(t^*, t) : 0 \leq t^* \leq 10, t = 10 - t^*\}$ and $R_2^\theta = \{(t^*, t^*) : 0 \leq t^* \leq 10\}$. The semantic relations here represent simple bijections. Whereas in X_2 θ is identical to θ^* and therefore a perfect

proxy (there is no difference in meaning), in X_1 meaning is inverted: higher talent corresponds to *lower* values of θ. It follows, however, that $\mu_1^\theta = \mu_2^\theta$, both being the uniform distribution over values of Θ. A test for semantic DE might notice that preimages of θ under R_1^θ and R_2^θ differ significantly. As a sample calculation, since R_i^θ are bijections here we can use the simpler definition of semantic DE using $\mathfrak{d}$. Set $\mathfrak{d}(t_1^*, t_2^*) = \frac{1}{|\Theta^*|}|t_1^* - t_2^*|$. Then, as a sample calculation,

$$\begin{aligned}\sum_{t \in \tilde{\Theta} \,:\, \text{both } (R_i^\theta)^{-1}(t) \text{ exist}} \mathfrak{d}\left((R_1^\theta)^{-1}(t), (R_2^\theta)^{-1}(t)\right) &= \sum_{t^*=0}^{10} \frac{1}{11}|(10 - t^*) - t^*| \\ &= \frac{1}{11}\sum_{t^*=0}^{10} |10 - 2t^*|\end{aligned}$$

which evaluates to $\frac{60}{11}$. We may say that relative to our choice of $\mathfrak{d}$, we benchmark this as a strong signal of semantic DE given the construction of our example (choosing ϵ, of course, depends on $\mathfrak{d}$).

What can the firm do in this case? Perhaps they can learn, through interviews with students or campus research, about the presence of these two groups, and reject the use of θ as a proxy given their goal. All proxies are flawed, but some may be more useful than others. Alternatively, if the firm gained access to data on group membership, they might be able to tweak their algorithm to interpret θ conditional on group. This is an example of where differential treatment of groups leads to better outcomes for both parties; although our conception of differential treatment might be tied to harmful discriminatory practices, it can also be used for good—it is what equity entails.

Admissions. One of our running examples throughout this thesis has been

thinking about fairness and bias in the context of education—specifically from the perspective of admissions. Consider the task of admitting students to a program, college, etc., based on features we can gather about them. Here, we take X to be a representation of students used to train an admissions algorithm, and $R = \{0, 1\}$ to signify a "admit"/"don't admit" recommendation as the outcome. Let X comprise disjoint groups X_1 and X_2, where X_1 is a group of historically and presently disadvantaged students and X_2 is the remainder of students. There are several ways we could think of X_1 in real life. For instance, X_1 could signify (i) students in racial or ethnic minorities (a racial analysis); (ii) students from lower socioeconomic strata (a class-based analysis); or (iii) students in public rather than private schools, or in school districts receiving less public funding rather than more (an economic resource-based analysis). Say the college's admissions officers wish to measure and train their algorithm upon is "effort", or "current academic grit". However, since there is no way to directly measure this, they rely on the proxy features of a student's GPA and set of AP exam scores in X.

One key way of applying DE is that it describes the effect of structural inequalities and injustices. We propose two interpretations linking inequality and data. The first interpretation states that the impact of inequality is an all-encompassing, systemic perturbation that modifies the innate qualities and dispositions—the underlying features—of an individual. The second interpretation states even if individuals maintained the same innate qualities over time, the fact that inequalities put individuals in lived experiences under structurally different circumstances manifests in differences in the distributions of features we can observe.

To better articulate these two views, consider the task setup above. The first view might tell us that we are doomed from the start. Group X_1, as a consequence of structural inequality, will innately contain differences of distribution of effort. One feasible explanation for how this might arise is through differences in group aptitudes. For instance, while students from lower economic classes might be born with the same innate potential for academic grit, a cumulative lack of access to rigorous education over their schooling can lead to the current distribution of grit being lower for these students, which consequently decreases average GPA/AP test-taking in the group. Thus, even if we could measure this underlying feature perfectly and this feature has a consistent meaning across groups, it would still exhibit distributional DE as a result of the world in which these students have been raised and educated.[xx]

We might make a similar argument for effort: that structural disadvantages have reduced the intrinsic willingness of individuals in X_1 to pursue opportunities and work hard, from disillusionment or group norms that prioritize sustenance over ambition. Bringing in Part 1, this aligns with the Roemerian argument that individuals of different "types" have different distributions of effort, and so we must compare an individual relative to their type. However, this way of thinking should be treated very carefully, since it is reminiscent of, and may perpetuate, toxic rhetoric that certain groups are *innately* more "lazy" and less hardworking.[xxi]

[xx] Of course, it is invalid to, for example, set X_1 to be individuals with lower grit and raise the alarm of distributional DE. In that situation, we cannot make the assumption that the distribution of grit is equal across groups; in fact, we specifically chose our groups to defy that assumption.

[xxi] This is exactly the message of some pervasive racial stereotypes. Christine Reyna provides a discussion of these attitudes in education[86].

To avoid this hazardous thinking, we should stress that the reason differences of effort arise is (per Roemer) because of soul-crushing circumstances and barriers in the world today, not because individuals are born inherently predisposed to expend less effort than others—to establish distributional DE in effort (differential effort) as a product of an individual's environment rather than their character. For consider one of the underlying features, "current academic grit". While this is not a proxy for anything, we could still conceive of its relationship with the feature of "academic potential at birth". Across groups, we expect potential at birth to be equally distributed, but the effect of being raised in differential circumstances would create systemic differences. In other words, the transformation from academic potential at birth to current academic grit is different among groups, and in the same way semantic differences cause distributional DE, compounded differences in average life experiences cause distributional DE. So we could in fact still adopt a similar theoretical way of reasoning as we did above with R_i^θ, except now R_i^θ do not convey the effect of choosing a proxy but rather the withering impact of inequality over time.

Adopting the second view, assume *contra* Roemer that baseline effort and academic grit are still equally distributed across X_1 and X_2. Then the reason we might observe differences in GPA and AP exam scores is because their use as proxies distorts and misrepresents effort and grit, so that taking their values as reflections of effort and grit is semantically unfounded.

The discussion in the introduction of this thesis gives us one explanation why this might be the case. A wealthier and poorer student with equal academic

capacities might each enroll in the same set of AP classes and receive the same marks, but when it comes to taking the exam itself, which is what our feature measures, there may be a split in behavior where the poorer student does not enroll in tests. An obvious reason why is the financial barrier of tests, or the availability of tests themselves. Some rural or underfunded school districts may simply not offer AP tests altogether, leading to empty values for the feature of AP exam scores. This case illustrates existential DE.

This is, however, not the only reason (especially considering that some schools might offer financial subsidies). Outside of school, the wealthier student could have access to a swath of test prep materials and tutors helping to boost their confidence for test-taking. Moreover, their peers and community might view AP test-taking as the norm, so signing up for an exam is expected, not exceptional. On the other hand, if the poorer student lacks access to outside resources or knows less about test-taking details (e.g. signup deadlines), they would likely take less exams, even if they had equal opportunity to register. Finally, Calsamiglia's work provides one more way to look at this issue; while students have equal level of effort proper, students with less resources, optimizing for their own utility, will end up diverting less of that talent toward academic pursuits. Then the value of "5 AP exams taken", while exceptional for a poorer student, might be below average for a wealthier student, even if they have the same level of academic potential—a clear form of semantic DE.

There is evidence to back this reasoning up. It is a documented fact that disparities in levels of AP test-taking across the US exist across students of differ-

ent races, even when controlling for a school's level of access to AP programs[19]. Additionally, the availability of AP courses differs among urban and rural areas, and despite the existence of subsidy programs, the financial cost of registration can act as a deterrent to taking the exam for low-income students, even if they have completed the corresponding course[46,58].

The psychologist Richard Weissbourd has studied these issues with admissions work in education.[xxii] Referencing past collaboration with the Common App, he notes that some admissions officers already assess GPA and AP exam records differently based on a student's background, and make attempts to contextualize an individual's circumstances beyond the canonical set of application criteria. For instance, he proposes, instead of looking at an individual's community service record, to also take into account part-time work responsibilities. Voluntourism, for instance, is only a viable option for those who can afford travel, and it is arguably no better of a reflection of one's altruism as is caretaking for a family member. So it is an arbitrary decision to only consider the first category of activity in admissions and not the second; the feature of "number of volunteer hours" can be semantically DE based on socioeconomic class if it is used as a proxy for "community engagement". Overall levels of community engagement might be equal for all, but the proxy is distorted because the way in which community engagement might manifest for poorer students is in working jobs or supporting family rather than volunteering. Moreover, rural students often have a less diverse set of accessible extracurriculars to choose from, which may adversely affect the distinctiveness

[xxii] Remarks here are taken from an interview done with Dr. Weissbourd in January 2024.

and attractiveness of their applications, as an example of distributional DE for the feature "engagement in extracurriculars".[xxiii] Finally, he suggests that given that wealthier students have more access to resources to help their essay-writing process and a lower counselor to student ratio (making it easier to get nuanced recommendation letters)—an example of distributional DE for the feature "level of application assistance"—admissions officers may want to consider other forms of evaluation other than an essays and recommendations, e.g. through videos. There are also intersectional concerns, such as the fact that poorer students in rural areas face even more barriers than just rural students, that our model with two groups does not quite capture.

Education and Assessments. Consider a situation where school administrators are trying to select for students to place in an accelerated program in grade school. Instead of academic achievement, the quality of interest here might rather be academic potential. Suppose a school decides that quantitative skills form an important part of academic potential, and measure this in part by considering a student's self-reported interest in STEM subjects and their prior coursework therein. This feature of a student's interest might also express DE; as there may be students who possess sharp quantitative skills who have simply prioritized other areas of study. For this group of students, a low level of interest in STEM does not indicate a lack of quantitative strength, as the school might believe. Consequently, this feature faces an issue where the extent to which it is semantically correlated with quantitative skills differs between different kinds of students. In

[xxiii] This has been empirically studied in Ontario schools, as reported by People for Education[7].

the real world, some school districts indeed screen for "gifted child" programs in elementary school using a private suite of assessments[27].

Broadly, let X be students at a school and say our group of interest is the subset of neurodivergent individuals. These individuals might struggle with standard test environments (e.g. timed and written) causing artificially negative impacts in performance. In this case, we might assume our feature of "test performance" stands as a proxy for "academic talent". In reality, this feature can also be viewed as a proxy for "comfort with standard test-taking environments". This latter underlying feature exhibits distributional DE. The cause of this DE is not historical injustice, but bad educational policy, wherein administrators make a faulty assumption that neurodivergent individuals thrive just as well in standard test-taking scenarios as other students.

As Weissbourd notes, academic assessments, like most facets of the education system, are designed with the typical student in mind, often blindsiding students with disabilities or different neurological profiles and learning styles. International students often lack more than just language proficiency, but the cultural currency to properly engage with peers which can impact, for instance, their performance in group projects. Individuals have different ways of understanding language, such that among two kindergartners with an equal level of "English language mastery", one may do better on a phonics test and the other better on a literacy task. Past studies have also pointed out differential problems with the SAT, possibly the most important test for high schoolers, by observing that certain questions function differently for black and white students owing to hypothesized

cultural differences[36,91]. Essentially, Weissbourd argues, every time an assessment is given to students the administrator necessarily deems the assessment a trustworthy proxy for a background attribute of students that they are trying to measure. But any assessment will necessarily be "coarse" by nature of trying to boil an individual's skills and abilities down to a number, and will intrinsically elevate certain capacities over others. A chief lack of understanding of students' specific and complex strengths and weaknesses leads to schools artificially hampering students' performance. Ultimately, Weissbourd proposes, it is an understanding of individual potential, rather than standard measures of achievement, that good teachers should prioritize, and that means we ought to come up with new ways of assessment *away* from standardization—in our words, new, improved, and more transparent proxy features.[xxiv] Because individuals learn in a multitude of ways, bias (DE) is inevitable; the best we can do is to recognize and minimize it.

Risk Prediction. We consider three very different kinds of risk scenarios. In all three, however, we can think of an algorithm whose outputs are risk or propensity scores, where higher values mean higher risk.

Our first example considers COMPAS as mentioned in the Introduction and Part 1. We think of X as the training data for COMPAS, and the two groups here as black and white defendants. Both of the impossibility results from Kleinberg et al. and Chouldechova express trade-offs and competing relationships between different fairness conditions as arising from an issue in the data—specifically, the

[xxiv] For instance, he espouses the following criterion for predicting student success: "look at the students to whom others naturally go to for help". This is an excellent heuristic on paper, but it is hard to think about encoding this into a measurable observation is another problem entirely.

base recidivism rates are different between the groups. This generates a tradeoff between calibration and balance (Kleinberg et al.), and between FPR, PPV, and FNR. From our perspective of DE, these impossibility results both express the message that the consistency of fairness conditions is predicated on an absence of distributional DE; that is, the expectation that base rates of recidivism should be the same between the two groups. To make this more precise, imagine that there is a binary feature θ indicating whether or not the individual reoffended. We would assume that the distribution of these features (proportion of ones) among X_1 and X_2 should be the same, but it is not, since black defendants were on average more likely to reoffend[21,60]. There is no immediate obvious choice of a feature for which we can view the indicator of reoffending as a proxy. The issue is that because of a complex web of factors, conceivably including the lasting impact of discriminatory and biased societal institutions as well as the constitutive impact, per Kohler-Hausmann, of what it means to be a black offender vis-à-vis being a white offender (where black individuals, once released, have comparatively less robust safety nets, are more alienated, or disproportionately policed, etc.), the distribution of reoffending is systemically higher for black individuals[61]. Fixing this differentially expressive reality falls in line with the long and laborious task of achieving racial justice.

Could Northpointe do better? As Part 1 explained, θ is not actually part of the algorithm's input data X, since it is what the algorithm is actually trying to predict. Knowing θ would be akin to having an oracle that could see the future, trivializing the problem. Consequently, our treatment of DE in this example does

not have a direct link to the algorithm, as the differentially expressive feature is "outside" the data. We might want think of ways in which the algorithm could treat defendants as if they were in a better world where racial discrepancies did not exist. But this is not Northpointe's objective; their goal, as with most decision-making efforts, is to best predict what would happen in the real world, rather than an ideal world, to forecast what is actually likely to happen[26]. Thus, diagnosing DE is separate from making a normative evaluation for how undesirable DE is for the algorithm. In some cases, we just want to understand and follow the data, even if the data is biased.[xxv] If DE tells us that the state of the world itself is unjust, that gives us a vision to do better, but that does not necessarily morally oblige us to scrap, or redesign, the algorithm.

Our second example examines a troubling situation presented by the political scientist and writer Virginia Eubanks as told in her book *Automating Inequality*, where she profiles a risk assessment tool, the AFST, used in Allegheny County, Pennsylvania. This tool is used as a screening tool for maltreatment reports by outputting a risk level for child abuse. Suppose X_1 and X_2 represent lower-class and middle/upper-class households in the county respectively. Eubanks identifies a critical flaw in the training data: there is far more data available for X_1 than for X_2. Some of this is inevitable: some social programs from which data is collected are only available to lower-income households, so even if wealthier families wanted

[xxv] Consider also, in the case of college admissions, the goal of measuring the underlying attribute "aptitude at handling college-level examinations". A student's AP exam, while distributionally DE, would not be semantically DE with respect to this underlying feature—AP exams are designed to measure just that. So although inequalities in AP access still exist, this feature would be the best feature that an admissions officer wishing to measure a student's college preparedness could hope to get.

to use those resources, they would not be able to access them. On the other hand, middle- and upper-class families using private insurance and living in secluded suburbs (where abuse can be easily hidden) and give less data known by the government: "the professional middle class would not stand for such intrusive data gathering." Compounding this fact is the fact that black and biracial families were reported to hotlines at disproportionately high rates (likely due to discriminatory attitudes or less privacy in their communities). As Eubanks observes, "when automated decision-making tools are not built to explicitly dismantle structural inequities, their speed and scale intensify them"[32].

This is another case of distributional DE, where data provenance is skewed to "oversample the poor", and as a consequence allows the algorithm to treat lower-income families with more scrutiny. There is also semantic DE: if a household has been reported by two hotline calls, this is likely much more troubling for a wealthy, secluded family than for a poorer, less secluded one. Because more cases of child abuse go unreported in wealthier households, there is a positive feedback loop: the algorithm is "harsher" on lower-income families by conflating poorer households with households more prone to abuse, leading to more police investigations on those families, leading to even more data being gathered for the poor. At the same time, data pertaining to the interaction of a household with public programs becomes contorted with the family being low-income, for there is an existential DE issue: sufficient data simply doesn't exist for wealthier demographics, and the model does not know how to treat this lack of information. In this case study, confronting DE goes hand in hand with confronting a dilemma of data collection.

In addition to thinking of the fairness of this algorithm's behavior, we may think about whether or not it is principally fair or practically useful to collect this data from low-income households when the wealthy do not, and would not want to, offer comparable data.

For our last example, briefly consider the situation of a bank assessing the credit risk of an applicant. Like with COMPAS, there is a relationship between distributional DE and the problematic origins of data, if we consider historical discriminatory practices performed by human lenders, of which various reviews exist[51,69]. The challenge here is distinguishing legitimate individual circumstances from blatant discriminatory outcomes. It is sensible here to think that, unlike what historical data seems to indicate, we ought to consider applicants for loans, credit cards, and mortgages without negative or positive racial connotations, situating us into an "ideal" scenario. However, learning how to adjust training data to achieve this by correcting for discrimination is not a straightforward task, and bias might still unconsciously propagate through the algorithm, as some recent studies have discovered[10].

Medicine. The history of medical treatment is not exactly an equitable or equal one, either. However, it seems that we might have better grounds to hope for fairer and less problematic data in this setting, since we only evaluate individuals based on biological attributes, free from the destructive effects of discrimination and human bias, for the purpose of accurately diagnosing some health conditions.

The picture in reality is less rosy. Current medical practices, like almost all other fields, carry over a legacy of prejudice and discrimination. As one example,

existential DE is an issue where, much like we have discussed above, there is far less health data available for minorities, which impacts the accuracy of diagnoses. Studies, for example, have shown that a lack of clinical studies involving lack patients have led to systemically higher rates of error when testing for health conditions[82,106,70]. Another example is presented by Obermeyer et al., who find that a commercial algorithm used to assess health risk used a truly dubious proxy feature: levels of healthcare spending as a proxy for need. This interpretation lent itself to semantic DE, as it completely ignored the fact that healthcare spending depends not only on need but also ability to pay and practitioners' decisions on where to allocate funds. Spending a few thousand dollars on a procedure is probably reflective of a much higher level of need if the patient is low-income rather than high-income. Additionally, spending on black patients was historically low, such that the algorithm falsely concluded that black patients were healthier than equally sick white patients and reduced the number of black patients flagged to receive extra care by over half[82].

A more interesting situation to consider is one in which there is, to turn the tables, an incorrect assumption of the presence of semantic DE. What happens when practitioners think that the value of a feature should be interpreted differently based on group membership, e.g. a patient's race, when in reality the feature does not have differential meaning? Vyas et al. categorize instances of these problems in medicine, in the context of algorithms producing risk scores for certain medical conditions. Some of these scores are then applied to "thresholding" tests, in which individuals test positive if some feature of theirs falls above a calculated

threshold.[xxvi] In fact, these issues have manifested in medical settings, where such thresholding tests are applied to patients' medical data as risk screens in a number of different areas such as cardiology, organ transplantation, and obstetrics. In many of these tests, black patients' scores are given offsets before being compared to set thresholds, allegedly to account for racial and genetic differences—i.e. a hypothesized semantic difference between races of scores[101].

There has unsurprisingly been criticism on the validity of these "point correction" adjustments. A common justification for these offsets is some claim about biological processes differing between races, such as black patients having more muscle mass. While some since-discarded measures had no empirical justification for this belief and relied on invalid data and methodology (e.g. eugenics), Vyas et al. find that other justifications at the time were backed up by reasonable empirical logic. However, finding a correlation between race and clinical outcomes is not sufficient grounds to use race as a predictor (there may be a confounding variable), and later studies have found that there is more genetic variation within race than across race, decreasing the supposed utility of using race as a feature[101]. Additionally, the optics of artificially applying corrective adjustments is a dubious one, and is not unlike an affirmative action policy in that it might make patients feel treated in an overly reductive and crude way based on their race.

[xxvi] Here is another methodological issue that thresholding tests can have through the lens of distributional DE: say the thresholding test is based on a feature for which the distributions of its values are different between different groups (e.g. there is group heteroskedasticity, or one group has their features distributed roughly uniformly while another has features distributed in a bell curve). Then the thresholding test might not work as expected for the groups; some groups may test positive at disproportionate and inappropriate rates. This would violate, for instance, the fairness definition of calibration. Because each group's distribution of the feature differs, it is inappropriate to apply one threshold across the board.

Not unlike our discussion of education, one path of improvement might just be to search for alternative features to use in place of race—better proxies for the biological "truth" of a patient. There is a remarkable example of this being done in practice, as reported in *JAMA*. A previous method to estimate the risk of kidney failure (being a set of estimating equations rather than an algorithm proper, but still corresponding with our framework) had incorporated race as a predictor. By removing race in favor of a new feature—the levels of cystatin C, a protein—the performance of the method remained roughly the same, and better for certain patients[15,43]. While just one example, this gives some proof that there are ways to address DE, or other bias problems, by performing a critical evaluation and modification of the features in X.

In the above, we have addressed different scenarios where our characterizations of DE apply. For a feature θ, we can get distributional DE without semantic DE if θ is not viewed as a proxy for anything but we still have unexpected differences in distribution across groups, such as with differential effort in education and admissions; we can think of situations where we have semantic DE but not distributional DE, such as in our example about two cultures treating the study of STEM versus finance and economics differently; and the feature could be affected by both, such as in the case of features used in AFST. Sometimes we may act as realists, optimizing for accuracy even if we acknowledge that our algorithms learn and perpetuate a biased world, and we simply accept DE. However, it usually ought to be productive to think more rigorously about the proxies we choose and how we can compare a real and "ideal" world. This is the focus of the next section.

2.4 An Extended Framework of Data Bias

In this section we review and build upon Friedler et al.'s (im)possibility paper. They provide a framework for thinking about the sources of bias by describing an observation process that, in terms of our vocabulary, produces proxy measurements from underlying measurements, and discuss how notions of individual and group fairness factor into this process. The purpose of this section is to present their contributions, making use of the mathematical setup that was given at the start of this part, while simultaneously extending their framework to accommodate our idea of differential expressiveness.

We have already talked about two main ideas in this part and how they interact with DE. First, we can conceive of a terrible transformation from an ideal world we hypothetically could, and should morally aim to, live in, without the effects of historical discrimination, economic inequality, prejudice, etc., to the flawed real world. Second, we often find ourselves providing to the algorithm proxy features for underlying features that are themselves not measurable. Our extended characterization of data bias gives us a structured way to categorize these ideas. We will present our extended framework directly, and note how it adapts from the (im)possibility paper[37].

Recall that we defined V as a universe of representations of individuals operated on by an algorithm, from which X is a finite set. To mathematically characterize X, we assumed it had a metric and probability measure. The key innovation here is to think of V and X as one of many *representation spaces* which

are related by transformations. For consistency with our previous development of differential expressiveness, we will treat all the spaces here as finite to preserve the setup. That is, the set of elements in each space is finite, each point corresponds to how an individual is represented *in the context of that space*, and groups are defined from a partition of the entire space into disjoint sets. However, Friedler et al. present their definitions more generally to cover the case where each space could be infinite, provided that the groups still form a partition of the space into disjoint sets. This is because the definitions given at the start of this part like probability measures, Wasserstein and Gromov-Wasserstein distances, and the group space have analogues in continuous settings. (In fact, these generalized definitions were mentioned in footnotes in the first section of this part.) Thus, while our presentation will only concern finite spaces, it should be understood that these concepts generalize to arbitrary metric spaces, should we wish to be more abstract.

A representation space, first and foremost, is just a generalization of how we viewed V and X, i.e. a metric probability space of representations of individuals.

Definition 19. A *representation space* is a finite metric probability space (S, d_S, μ_S), where S is a set of representations of individuals. Each element $s \in S$ that is a representation of an individual takes the form of a vector of features. If μ_S is not specified, we take it to represent the uniform distribution over elements of S.

We say "representation space" to emphasize the fact that each space uses features to *represent* individuals. For shorthand, we usually omit writing d_S and μ_S. Since our spaces are finite, we continue to implicitly define the set of events as 2^M (to formally define μ). Groups are partitions of the spaces. The act of defining

a representation state already marks a large conceptual leap from thinking of an individual as a fully-fledged human being to thinking of them as a vector.

We begin by formally stating the two main ideas above. The first idea is that, as we just stated, there is a comprehensive difference between an "ideal" and "real" world. If we compared a representation space for the "ideal" world with a representation space containing the exact same features for the "real" world, those spaces would most likely be different because the values of the features for each individual would differ.

Definition 20. Let the *ideal world* denote an imagined state of the world where any systemic disparity does not and has never existed. These disparities include discriminatory institutions, individual biases, and unjust differential access to resources and opportunities. Denote the universe of values for any possible feature of any individual in the ideal world as $\mathcal{D}^*$. Similarly, denote the universe of values for any possible feature of any individual in the real world as $\mathcal{D}$.

We write $\mathcal{D}^*$ as conveying the universe of data for individuals in an ideal world, as opposed to $\mathcal{D}$. This chasm between what an individual's feature values would be in an ideal world and what they are in reality is very loosely motivated by Dwork et al.'s idea of a "better world", where they analyze modifications to existing predictors trained on real-world data so that the modified predictors perform as if they were trained a transformation of the world that improves it (e.g. achieves balance for the positive class)[31]. While Dwork et al. constrain the form of the possible transformations under consideration to rigorously analyze their modifications, we adopt a more conjectural definition. The main question is

where we set a boundary between an unacceptable and acceptable disparity; what kinds of inequality do we still tolerate? For instance, we should do away with sexism in $\mathcal{D}^*$, but we should not deny our biology by desiring that all individuals have equal distributions of physical characteristics. For our attempt to draw the line, we propose that we should eliminate whatever a political philosopher would feasibly deem socially unjust.[xxvii] So in $\mathcal{D}^*$ segregation, redlining, and gerrymandering do not exist, but people could still be neurodivergent or disabled.

The second idea is that, as already introduced when defining semantic DE, there is a difference between an underlying feature and a proxy feature chosen to estimate it. Recall that we expressed the relationship between an underlying feature θ^* and a proxy θ in terms of the semantic relations for each group, R_i^θ. The overall mechanism obtaining values of θ from θ^* is then determined by the semantic relations (as given in Proposition 2), which we described as an "observation process that translates and crystallizes values of θ^*". Here, we will consider this process applied to an entire vector at once, giving a transformation from a representation space of underlying features to a representation space of proxy (observed) features.

We can now define our representation spaces, of which there are four in total, along with the transformations of metric probability spaces between them.

Definition 21. Fix a task with outcomes in the result space R for which we aim to design an algorithm f. The *underlying space*, U, is the representation space in the real world $\mathcal{D}$ where representation vectors contain underlying features, the

[xxvii] Since each philosopher has a different conception of what exactly is unjust, this definition is still a bit ambiguous. Given Part 1, we think it is most apt to go with a Dworkinian view: an ideal world is one where *there are no large-scale factors disadvantaging individuals which exist beyond their control and cannot be attributed to differences in innate endowments.*

"desired" qualities and attributes of the individual that the task truly wants to consider. The *proxy space*, X, is the representation space in the real world where representation vectors contain the corresponding proxy features, i.e. the features we actually observe. Let U^* and X^*, the *ideal underlying space* and *ideal proxy space*, denote the analogs of U and X under the ideal world, $\mathcal{D}^*$.

A *transformation* is a map between representation spaces taking an individual (and group)'s representations in one space to their representation in the other. Let the *proxy transformation* or *observation transformation* f_p be the map $U \to X$ representing the process of observing values of proxy features from the underlying features, and the *ideal proxy transformation* f_p^* be its analogue $U^* \to X^*$. Let the *reality transformation* f_t be the map $U^* \to U$ representing the process of going from the ideal to the real world among the underlying features.

We will also assume R has a metric and measure, such that it can be viewed as a representation space (where the representation is the algorithm's output). See Figure 2.4 for a visualization of these spaces. U^* and U have the same features whose values might differ, and similarly for X^* and X. In the case that a feature is not a proxy for anything, we can view it as a trivial proxy for itself and present in both U and X (and their ideal counterparts).

We have defined our notation to make this setup compatible with our earlier definition of an algorithm as a map $X \to R$, since X still represents the actual data given to the algorithm. However, we can now conceive as f as the last stage in a larger algorithmic pipeline given f_r and f_p.[xxviii] All the fairness definitions in

[xxviii] One aspect of the pipeline we have not considered here is data sampling, i.e. the process of

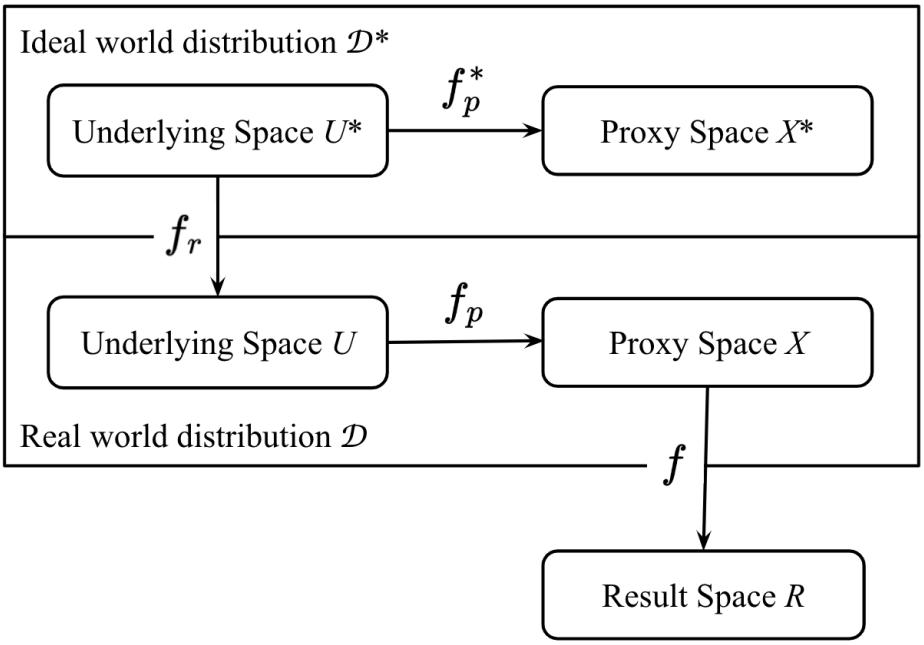

Figure 2.4: The different representation spaces we consider for a task. The pairs (U^*, U) and (X^*, X) share the same set of features for their representations of individuals. U^* and X^* have values of features that would exist in the ideal world, whereas U and X have values of features that exist in the real world. f denotes the algorithm; all other functions are transformations.

Part 1 are stipulations on f, since they do not consider the background processes responsible for generating data. There is now more to explore.

As a remark on notation differences, Friedler et al. defines U, X, and R as the "construct space", "observed space", and "decision space" respectively. However, they do not consider U^* or X^*. The aim of some tasks might be to estimate an underlying feature in U or U^* given X. To adopt an example we have already covered, in the context of admissions we can take U^* and U to represent academic

getting our finite set of individuals X from a universe of individuals V. Improper sampling could serve as another possible source of bias if some groups are misrepresented by their members in X. We could think of this as occurring between f_p and f, if U^* and U are interpreted as infinite universes of representations. There is a large literature in statistics and the social sciences tackling the issue of sampling bias, so it is not our focus[107].

grit in the ideal and real world, X to be GPA or AP exam scores, and R as the binary variable denoting an admission recommendation. We will now summarize the main takeaways of the (im)possibility paper in this setup.

Since Friedler et al. do not consider an ideal world, their chief concern is characterizing f_p and the way in which it can express bias. They immediately consider three ways in which f_p can be problematic: (i) it could add noise to features, (ii) it could express proxy features in terms of functions of multiple underlying features, and (iii) it may add meaningless proxy features that are independent from underlying features. We can compare this to how we interpreted the observation process in terms of semantic relations R_i^θ. Suppose θ is a proxy for θ^*. We may express (i) by saying that the semantic image for a value of θ^* is a set of multiple values corresponding to adding random noise that produces variation in observed values of θ. However, for (ii) and (iii), our semantic relations only consider isolated pairs of underlying and proxy features, and not possible many-to-many relationships, such as if θ were a function of several θ^*. So although our framework is enough to categorize semantic DE, it may fail to capture the full complexity of the overall transformation f_p.[xxix]

The main thrust of the (im)possibility paper is to describe core assumptions of fairness definitions in terms of relationships between U and X. Instead of considering an algorithm as just f acting on X, they advocate to consider it as acting on U by analyzing the map $f \circ f_p : U \to R$. This conveys how an algorithm

[xxix] We mentioned this limitation briefly when introducing semantic DE. If there are n^* underlying features and n proxy features, we might extend our analysis to consider all $n^* \cdot n$ semantic relations, one for each pair of underlying feature and proxy feature.

treats the "true" character of individuals. They also articulate two main axioms, for which we provide our own interpretations and observe new results.

Remark 3. U and X fulfill the "what you see is what you get" axiom (WYSIWYG) if the additive distortion of f_p is small (less than some ϵ). WYSIWYG is assumed to ensure that individual fairness constraints on f translate into comparable individual fairness constraints on the overall map $f \circ f_p$.

The WYSIWYG axiom naturally goes along with individual fairness discussions since the technical condition it relies on, additive distortion, is defined in terms of how f_p acts on individuals. The remark essentially explains that since individual fairness definitions bound the extent to which f distorts distances from X to R, assuming WYSIWYG also guarantees that the overall distortion from U to R as given by $f \circ f_p$ will be small. In particular, say that f_p satisfies (ϵ, δ)-individual fairness. Choose some $\epsilon' < \epsilon$. Then if we assume the WYSIWYG axiom with parameter ϵ', $f \circ f_p$ satisfies $(\epsilon - \epsilon', \delta)$-individual fairness as a map $U \to R$. To prove this, notice that since the algorithm bounds the distortion of f_p, any two individual representations in U which differ by at most $\epsilon - \epsilon'$ are mapped to points in X which differ by at most ϵ; the fact that the outcomes differ by at most δ then follows by the definition of individual fairness. Following the same strategy of tracking distortions of distances as we move from U to X to R (we omit full proofs), if we assume WYSIWYG with parameter ϵ' and constrain f to have distortion at most ϵ, then the overall map $f \circ f_p$ will have distortion at most $\epsilon + \epsilon'$. If we instead constrain f to be K-Lipschitz as defined in the Appendix, then $f \circ f_p$ will satisfy $(\epsilon, K(\epsilon + \epsilon'))$-individual fairness.

Remark 4. Consider a partition of individuals in our representation spaces into groups. Let $U_1, \ldots, U_k$ be the subsets of U corresponding to the groups. Then U fulfills the "we're all equal" axiom (WAE) if the Wasserstein distances on U between pairs of groups in U are small, i.e. for all U_i, U_j, $\mathcal{W}_d(U_i, U_j) < \epsilon$.

Additionally, consider the group spaces $\mathcal{U}$, $\mathcal{X}$, and $\mathcal{R}$ defined based on U, X, and R respectively. We say that f_p admits ϵ-*structural bias* if the group skew between the two spaces, $\sigma(\mathcal{U}, \mathcal{X})$, exceeds ϵ, and we similarly say that f admits ϵ-*direct discrimination* if $\sigma(\mathcal{X}, \mathcal{R}) > \epsilon$. Conversely, if $\rho(\mathcal{U}, \mathcal{R}) < \epsilon$, we say the overall map $f \circ f_p$ is ϵ-*nondiscriminatory*.

Recall from our mathematical setup that given a partition into groups, Wasserstein distances on a representation space, as well as the group space, can be defined using the space's measure μ. WAE essentially assumes that all groups are "close" to each other in U, having no innate differences in the values of their underlying features, and the following results give a mathematical treatment of the notion of groups being treated differentially by the observation process f_p (structural bias) and/or the algorithm that produces results f (direct discrimination) using group skew. This allows us to think about group discrimination in a modular way, and tells us that in order to guarantee overall nondiscrimination over groups (as measured by $\sigma(\mathcal{U}, \mathcal{R})$). Indeed, they present the following result: if we impose the constraint on f that the distances between groups in R must be small (a kind of group fairness constraint), i.e. $W_d(R_i, R_j) < \epsilon$ for all pairs of groups R_i, R_j in R, then assuming the WAE axiom with parameter ϵ', the overall map $f \circ f_p$ is $\frac{\max(\epsilon, \epsilon')}{\delta}$, where δ is the noise term added in the calculation of group skew. We

defer the proof to the (im)possibility paper where it can be found[37].

Now let us consider and compare our contributions to this framework. First, since our formulation of differential expressiveness is in terms of groups, it does not engage with the individual fairness results presented by Friedler et al.. However, the WAE axiom is strikingly similar to our discussion of semantic DE, in which we assumed that the distribution of an underlying feature, μ^* should be the same across groups (see Proposition 2). The difference is that the WAE axiom considers groups at large as subsets of U, whereas our formulation of μ^* is for an individual underlying feature. If we apply our assumption of equal μ^* across *all* underlying features in the representation space U, then with high likelihood the pairwise Wasserstein distances between groups in U should be small.[xxx]

The main new contribution of our framework comes with the inclusion of the ideal world $\mathcal{D}^*$, which allows us to think about an additional transformation f_r that, for us, happens even before f_p. f_r, for us, represents the process of moving from the ideal to the real—what we imagine as a regrettable and soul-crushing process by introducing inequality and injustice. This gives practitioners an additional option to consider when it comes to conceptualizing the overall mechanism that their algorithm implements. Instead of thinking about an algorithm as a map from data about proxy features to outcomes $X \to R$, the (im)possibility paper opens up the possibility to think about an algorithm as implementing a larger mechanism from underlying, "true" attributes to outcomes $f \circ f_p$, which opens up the possibility of considering how the process of observing coarse and

[xxx]To numerically quantify this, however, we need to consider factors like different sizes of the groups, which requires probabilistic analysis we do not perform here.

blunt proxy features instead of underlying ones might introduce bias before the algorithm even enters the picture. With our framework, the practitioners now have a third option: considering their algorithm as the last step in a mechanism $f \circ f_p \circ f_r$, if the starting point is to evaluate individuals based on the underlying features they might have in an ideal world as signified by U^*.

Just as Friedman et al. applied the WYSIWYG axiom to f_p, we can also consider applying the axiom to f_r. If we assume that WYSIWYG holds for f_r, then we can derive similar results to the above wherein making f obey individual fairness constraints begets individual fairness guarantees for the entire pipeline $f \circ f_p \circ f_r$. This would be excellent; this would mean that a mechanism is not only individually fair with respect to underlying features in the real world, but it would also be fair with respect to underlying features in the ideal world—in a way, the algorithm would treat individuals like their best selves.

However, assuming WYSIWYG holds for f_r is, as we have discussed earlier in this thesis, likely a mistaken assumption. Because f_r represents all the injustice and evils of the real world, it likely causes significant distortion of distances, making the guarantees we might have for $f \circ f_p$ fail to carry over to $f \circ f_p \circ f_r$. Furthermore, instead of assuming WAE holds in U, we assume it holds in U^*. It is sensible to think that the introduction of discrimination that comes with f_r will cause it to exhibit high group skew, causing WAE to no longer hold in U. This follows the argument we presented for distributional DE when analyzing examples: sometimes the underlying feature itself is unequally distributed, against our expectation, because of the legacy of disparity and how that might selectively

affect cultural norms and character dispositions among groups. In other words, f_r is the vector in which distributional DE gets introduced in our world. As a result, WAE might not hold in U, and we get neither individual nor group fairness guarantees for the overall mechanism $f \circ f_p \circ f_r$.

How might practitioners address this? One approach is to simply dismiss U^* as providing a theoretical vantage point, and interpreting f_r a way of expressing grievances for the state of this world but not an actual transformation to tackle. As we have already said, in some cases the algorithm is meant to best predict what happens in the real world rather than some imagined ideal which, although more morally just, would give a worse-performing algorithm right now. In other words, ignore f_r and work with the reduced framework $U \to X \to R$ with overall mechanism $f \circ f_p$, as in the (im)possibility paper. This is a legitimate strategy; after all, algorithms aim to meet a teleological goal. Nonetheless, this framing might be helpful in some circumstances (e.g. when we want to assess an individual's innate potential and behavior), whereupon we might take interest in the line of work that was introduced with Dwork et al.'s "better worlds" paper and will hopefully continue to develop.

There remains one unaddressed part of our extended framework: that of the map $f_p^* : U^* \to X^*$. As we defined it, this is analogous to the process of observing proxies even under an ideal world. We include this to highlight the fact that even under ideal circumstances, algorithm designers and data collection processes might still face the issue of having to pick proxies. Even if there is equality of educational opportunity, teachers will likely still need to hand out academic assessments, by

virtue of them being a far more easy way to measure a student's level of competence and academic ability than, say, individually working and investing time into every student.[xxxi] Similarly, we will still be unable to measure an abstract underlying feature such as "grit", relying on proxies like GPA and AP exam scores. Proxy features will still by nature remain coarse and rough approximations—and this means that some of the case studies of semantic DE that we have discussed above, such as neurodivergent individuals being disadvantaged in normal means of assessing students, and exceptional students of different cultures being more inclined to study in different fields, can still happen. While moving to $\mathcal{D}^*$ will solve systemic issues, it is *not* a panacea for DE entirely. The conclusion as represented by f_p^* is that to detect and mitigate differential expressiveness entirely is a deceivingly complex endeavor. In short, we associate distributional DE with f_r, indicative of the manifestation of systemic harms and unfairness in reality, and associate semantic DE with f_p and f_p^*, indicative of the methodological hazards that come with every choice of proxy, conscious or not, that we make.

2.5 Towards Remediating Differential Expressiveness

We finish our thesis by connecting DE to other characterizations of data bias that have been proposed in the literature. Our characterization is most similar to three kinds of harm that Suresh and Guttag describe in their framework: historical, measurement and aggregation bias. In historical bias, the world *as it was* or

[xxxi] Of course, teachers *ought* to act this way and we would expect that they do in an ideal world, but in situations where schools need some way of reporting a standardized and easily-understandable measure of ability across hundreds if not thousands of students, grades and marks may very well still be a necessity.

as it is reflects intrinsically harmful outcomes, which aligns with our conception of distributional bias (as conveyed through f_r). In measurement bias, features aimed at approximating unobserved constructs are measured differently among groups, causing differences in measurement methods and accuracies or plainly oversimplifying the construct. Semantic DE resembles this idea, whereupon even with representative and equal sampling strategies, features do not properly convey what they are meant to convey. Similarly, aggregation bias occurs when the same model is used on underlying groups which ought to be treated differently; the same variable means different things to different subgroups, making the overall model biased toward a majority group and/or suboptimal for all groups[98]. This accords nicely with our understanding of semantic DE, although they consider this bias as arising during the deployment, rather than training, of the model.

Ntoutsi et al. categorize different ways in which bias can be understood, mitigated, and accounted for. Our entire discussion of DE lives within one section of their presentation: understanding the "socio-technical" causes of bias, which as defined for them constitutes an investigation of the processes that generate data[79]. DE describes at its core problems surrounding the origin of data, whether those origins are historical or methodological. Finally, we mention that Mhasawade et al. conceive of a similar separation between the "world as it should and could be" and the "world as it is", calling the process that takes us from the former to the latter "retrospective injustice", or "societal bias"[82]. This is yet another way of presenting f_r from our framework, and demonstrates to us that the framing tool of thinking about how things *ought* to be as compared to how they are right

now has been considered before. The entire set of circumstances where DE arises is, in a sense, "pre-algorithm", and thus only constitutes a small fraction of all the different situations where algorithmic bias can arise. Understanding DE will not completely characterize data bias, but we nonetheless hope that it gives an insightful new perspective with which to view it.

To finish, we note that differential expressiveness is also fundamentally a very applied problem, in that it is all about information gathered into datasets. Hence, one practical way to be aware of bias in data is if datasets are rigorously analyzed as part of the curation process, and appropriate information and warnings are disseminated to those using data to train their models. There exists work, albeit limited, on this front. A 2018 paper by Gebru et al. has pointed out a lack of standardization in the process for documenting datasets, and proposes the idea of *datasheets for datasets*, standardized templates for communicating matters of data provenance, that is, providing documentation around how datasets were created and including relevant metadata and methodological details. The hope is to use these templates to increasing data transparency, stewardship, accountability, and reproducibility[38]. Holland et al. propose a very similar framework that suggests including information about dataset provenance, simple statistics, and pair plots among others in the form of "dataset nutrition labels". Happily, these forms of rigorous documentation have seen adoption by the machine learning community, and have accompanied some of the latest datasets, including new benchmark assessments for large language models[47,92]. Fabris et al. put these frameworks to use by generating documentation for over two hundred datasets intended to be

used in machine learning fairness research[33]. They particularly note that a few datasets are disproportionately used in the literature despite containing problematic attributes (e.g. noisy data, coding mistakes, etc.), including the very dataset used in ProPublica's analysis of COMPAS. Ultimately, perhaps one of the best ways to stay alert of and efficiently address differential expressiveness and data bias at large is to be more thorough and thoughtful in thinking about the data we take for granted before the training process even begins.

We have already included above some strategies to reckon with DE. Sometimes we will just choose to live with DE, so to build algorithms more suited to perform well in the real world. We may also strive to look for better proxies that supersede more dubious ones. We might think about ways to reconstruct a better world by modifying our existing algorithms, while being tactful not to accidentally implement questionable measures in the name of equity (as has happened in medicine). Our definition of distributional DE has also been presented using statistical language that straightforwardly lends itself to tests for differences of distributions. In fact, auditing for distributional DE is equivalent to the problem of testing for a difference between distributions, e.g. by using the Kolmogorov-Smirnov test. In the Appendix we also frame existential DE in terms of a statistical test. The only additional thing that is needed to qualify our applying a statistical test is a normative judgment that we expect the distributions of a feature not to differ across groups. Statistical inference, of course, is a whole other realm in itself worthy of further study, so our use of statistical testing is only a cursory suggestion. Beyond these suggestions, a lively sector of algorithmic fairness focuses on strategies for

pre-processing data, and several works have already shared their treatments of the issue of representing individuals to an algorithm[16,56,111]. This thesis, however, has concentrated on describing the issue of DE, rather than solving it and/or evaluating if these such existing methods do (or do not) counteract DE. It may well be the case, however, that some of these methods already contain the key to addressing differential expressiveness. Performing a review with current state-of-the-art methods for modifying data is just one of many other natural next steps to take given the material we have presented.

3

Conclusion

Prima facie, we might say the task of thinking about fairness is an intuitive one. In social settings it may often be natural to deem situations and actions fair or unfair based on our instinct, and perhaps equipped with a few distributive justice principles from philosophy class. One thematic objective of this thesis has been to show that in the context of complex real-world scenarios involving algorithms, this

intuition verily does *not* carry over. Part 1 of this thesis addressed fundamental fairness definitions themselves when giving a "story" of the evolution of the field. As we saw through impossibility results and critiques of the definitions, even defining fairness itself is a difficult task, and the field has not adopted one canonical definition of fairness to follow. Part 2 of this thesis investigated the area of data bias. In exploring the phenomenon of differential expressiveness, it also seems that the rudimentary goal of getting trustworthy and unbiased data in the first place is a demanding one, due to the challenges we encounter when considering the impact of historical and present disparity as well as the implicit coarseness that comes with picking proxy features to represent underlying attributes.

This being said, it might seem that the tone of this thesis is a rather pessimistic one casting efforts to achieve algorithmic fairness as hopelessly obstructed by competing definitions and flawed data. To be very clear, this is not the intention. The message here is not to cry wolf about how wicked the data around us is, but rather draw attention to the deceiving complexity of fairness and justice while suggesting new perspective to think about existing issues. We think DE is, at the minimum, provides an interesting mathematical spin on an existing problem, and, more optimistically, shines a new light on the difficulty of prescribing a *meaning* to a feature by connecting it with other problems in data bias. Distributional DE can describes an issue with injustices that exist in the world today, whereby attaining the same value of a feature requires disproportionate effort among different groups. On the other hand, semantic DE arises as a consequence of the feature itself signifying different things to individuals in different groups.

To end this thesis, we address a few categories of shortcomings and omissions. Firstly, while our broad description of an algorithm as implementing a task and acting on vectors of features representing individuals plausibly seems quite general, it is actually rather restrictive considering the vastness of the field of machine learning. Many application areas of machine learning models do not fit into this structure of producing outcomes for individuals, but may still be viable settings to discuss differential expressiveness, or extended characterizations thereof. For instance, algorithmic bias has been well-studied in the setting of natural language processing (e.g. word embedding models enshrining gender bias), especially given the rise of large language models. We could try translating DE to be defined in this situation as the phenomenon whereby the internal world model of a language model disproportionately associates certain groups X_i with categories of words or concepts C_i, when in reality we think that each C_i should equally represent all X_i, or equivalently, each X_i should have a representation in C_i that reflects an ideal world. For instance, a C_i could be the semantic category of a career, an element of the set $\{\texttt{medicine}, \texttt{politics}, \texttt{theoretical computer science}, \texttt{athlete}, \dots\}$, while each X_i correspond to demographic groups which may currently be underrepresented among certain C_i (so we want the distribution of concepts to follow an idealized distribution, not the real-life distribution), e.g. the partition of humans into men, women, and non-binary individuals. As another example, existential DE seems like a relevant concept when considering historical case studies of facial recognition models being far better at processing images with light skin tones versus dark skin tones, due to a lack of diverse samples during training (so there

is missing data for racial minorities in the training set).

While defining our setup, there are more limitations we have flagged. Simply taking the groups X_i to be partition of X might not properly express the impact of intersectionality whereby we want to evaluate fairness with respect to different group partitions simultaneously (e.g. race as well as socioeconomic class); this is a motivation for the definition of multicalibration we roughly described in Part 1. We could extend our presentation to cover the general case where X and the different representation spaces are not necessarily finite. In our interpretation of X, we could consider more general scenarios in which entities x do not represent humans but rather, e.g., companies, nation-states, political parties etc., so long as there is some meaningful notion of forming groups. This could also open up for our consideration a wider range of scenarios in which the use of algorithms has taken root, such as in biology and voting. The presentation of semantic relations, aside from being somewhat notationally cumbersome, faced two opportunities for generalization as we noted: changing semantic images and preimages to be probability distributions over feature values rather than just sets of values, and to also accommodate the fact that the relationship between underlying and proxy features could be many-to-many. Our work, like most of algorithmic fairness, also is quite local: we defined DE on the level of individual features, which might lead us to miss sight of the bigger and constitutive picture on the scale of entire feature vectors representing individuals.

As the impossibility results might suggest, algorithmic fairness is a field teeming with definitions and relationships between them. However, the link between

Parts 1 and 2 of this thesis is not especially strong aside from sharing core concepts; we did not really consider how differential expressiveness works with the different definitions of individual and group fairness in Part 1. (E.g. if feature θ is semantically DE, what does that tell us about predictive parity, balance, calibration, etc? Are there scenarios in which DE is not caught by fairness notions, or conversely results we can show where DE will be "caught" by a notion?) Similarly, there are likely more connections between the two characterizations we gave of DE, and of our characterization of DE and other frameworks of bias that are waiting to be discovered. Additionally, we have not focused on developing strategies to detect DE and fix it through data, aside from the brief references at the end of Part 2. Nor have we analyzed real datasets to examine more examples of how DE appears "in the wild". For the most part, our ideas currently remain just that—ideas and proposals, not yet fully-fledged and empirically tested theories proper. Understanding this, this thesis only presents a start.

While this thesis belongs in the field of algorithmic fairness, there was very little to talk about algorithms proper. As mentioned, we have not talked about broad swathes of the field, including the literature on how to achieve and guarantee fairness in practice. There are even more related concepts that we have not touched on. Aside from datasheets for datasets, other proposed ways of making it easier to catch and fix algorithmic bias think about how we might change the way humans interact with the data representing them and the algorithms that process that data, c.f. the nascent fields of machine learning interpretability, explainable AI, and human in the loop approaches within the study of human-computer

interaction. In the age of big data, other scholarship has studied the privacy implications with assembling datasets of unprecedented scale, raising ethical concerns about what kinds of underlying and proxy features we *ought* to use, c.f. the philosopher Helen Nissenbaum[77,30]. To think about algorithmic fairness is to enter a vast forum of discussion where striving to be utterly comprehensive is a futile and unnecessary attitude.

The heart of this work revolves around a few core themes. There is the long history of defining fairness as dealt with in political philosophy, law, economics, and computer science. There is the idea of representing individuals through data and features to an algorithm. There is the idea that every data curator and algorithm designer implicitly makes a choice of a proxy feature when collecting data, and that there is almost always some element of bias we should be aware of. There is the idea of comparing an ideal, hypothesized world with the one we live in. And there is the juxtaposition between equality and equality, similar to the tradeoffs between different conceptions of fairness. Differential expressiveness naturally leads us to adopt a more equitable view, insofar as the fact that meaning differs across groups implores us to be aware, rather than blind, to group membership (which may be sensitive) in hopes of producing better outcomes for everyone.

At the same time, we should not mistakenly lead us to think that data is the root of all evil, and ridding datasets of bias will be a panacea. As Guttag writes, the statement "data is biased" is not false, but "treats data as a static artifact divorced from the process that produced it... long and complex grounded by historical context and driven by human choices and norms". Data is a living thing,

and we have the power to change it inasmuch as it holds the power to affect us by way of the algorithms that are trained on it. We hope our discussions, conceptual and otherwise, mark progress towards a further rigorous way of thinking about data and fairness, and also help us reflect on the very real challenges that still exist in today's world for which biased data is just a symptom. Humans are far more than trends, biases, and numbers, and the challenge upon us is to ensure that the increasingly potent algorithms we design understand that too.

4

Appendix

4.1 Metrics on Probability Distributions

In the mathematical preliminaries of Part 2 we mentioned that D_∞, defined in "Fairness Through Awareness", was a metric on ΔA[28]. While this metric has been mentioned in the statistics literature it appears to be less discussed than D_{tv} and a proof that it is a metric was not included in the paper. So it may be a helpful

exercise to show that it is indeed a metric.

Proposition 3. *The relative ℓ_∞ metric D_∞, as defined for elements in ΔA where A is a finite set, is a metric.*

Proof. Consider any $P, Q, R \in \Delta A$. We want to show D_∞ satisfies the three criteria defining a metric. For criteria (i) note that for any event $a \in A$ $\max\left(\frac{P(a)}{Q(a)}, \frac{Q(a)}{P(a)}\right) \geq 1$, with equality iff $P(a) = Q(a)$. If $P \neq Q$, there must be some $a \in A$ on which they assign different probabilities, so $D_\infty(P, Q) > 0$. If $P = Q$, then $\forall a \in A$, $\max\left(\frac{P(a)}{Q(a)}, \frac{Q(a)}{P(a)}\right) = 1$, so $D_\infty(P, Q) = 0$. Criteria (ii) follows by definition of D_∞.

To show the triangle inequality (iii), let $a^* = \arg\max_{a \in A} \log\left(\max\left\{\frac{P(a)}{R(a)}, \frac{R(a)}{P(a)}\right\}\right)$, and WLOG suppose $\frac{P(a^*)}{R(a^*)} > \frac{R(a^*)}{P(a^*)}$. Then

$$\begin{aligned}
D(P, R) &= \log\left(\frac{P(a^*)}{R(a^*)}\right) \\
&= \log\left(\frac{P(a^*)}{Q(a^*)} \cdot \frac{Q(a^*)}{R(a^*)}\right) \\
&= \log\left(\frac{P(a^*)}{Q(a^*)}\right) + \log\left(\frac{Q(a^*)}{R(a^*)}\right) \\
&\leq \log\left(\max\left\{\frac{P(a^*)}{Q(a^*)}, \frac{Q(a^*)}{P(a^*)}\right\}\right) + \log\left(\max\left\{\frac{Q(a^*)}{R(a^*)}, \frac{R(a^*)}{Q(a^*)}\right\}\right) \\
&\leq \max_{a \in A} \log\left(\max\left\{\frac{P(a)}{Q(a)}, \frac{Q(a)}{P(a)}\right\}\right) + \max_{a \in A} \log\left(\max\left\{\frac{Q(a)}{R(a)}, \frac{R(a)}{Q(a)}\right\}\right) \\
&= D_\infty(P, Q) + D_\infty(Q, R).
\end{aligned}$$

as desired. □

Dwork et al. also mention a relationship between D_{tv} and D_∞, namely that $D_{tv}(P, Q) \leq 1 - \exp(-D_\infty(P, Q)) \leq D_\infty(P, Q)$. Proofs for this inequality are available online and for brevity will not be reproduced here[48].

These two metrics are just part of many metrics used to express distance between statistical distributions. A good overview of such distances in general is given by Gibbs and Su[39]. It is worth noting that some commonly-used ways of conveying "distance" between distributions are not, in fact, metrics, such as the KL divergence:

Definition 22. Let A be a finite set and $P, Q \in \Delta A$ be two probability distributions defined on A. The *Kullback-Leibler (KL) divergence* of P from Q is defined as $D_{KL}(P\|Q) = \sum_{a\in A} P(a) \log\left(\frac{P(a)}{Q(a)}\right)$.

KL divergence is not a metric. As a simple counterexample, let $A = \{a_1, a_2\}$, with the elements assigned probabilities of $(0.1, 0.9)$ and $(0.5, 0.5)$ by P and Q respectively. Then $D_{KL}(P\|Q) \approx 0.029$ but $D_{KL}(Q\|P) \approx 0.222$, violating symmetry. However, for example, the *Jensen-Shannon divergence*, which symmetrizes KL divergence, is a metric, defined between P and Q as $\frac{1}{2}(D_{KL}(P\|Q) + D_{KL}(Q\|P))$.

The takeaway of this to our work is that if practical computational considerations are not the focus, there is a wealth of metrics to draw from that formalize the notion of distances between distributions. An interesting line of work could examine the specific properties and benefits of using certain notions of distance over others, e.g. if they satisfy nice mathematical or computational properties. However, our analysis mainly considers the abstract, meaning we are overall not concerned about particular choices of metrics.

4.2 Definitions of Continuity

Consider a map between metric spaces. We can consider commonly-used definitions of what it means for f to be continuous:

Definition 23. Let (M_1, d_1) and (M_2, d_2) be two metric spaces and $f : M_1 \to M_2$ be a map between them. f is *continuous* if $\forall a \in M_1$ and $\forall \delta > 0$, $\exists \epsilon > 0$ such that for any $b \in M_1$, $d_1(a, b) < \epsilon \implies d_2(f(a), f(b)) < \delta$. f is *uniformly continuous* if $\forall \delta > 0$, $\exists \epsilon > 0$ such that $\forall a, b \in M_1$, $d_1(a, b) < \epsilon \implies d_2(f(a), f(b)) < \delta$.[1]

The idea is that when approaching a point $a \in M_1$, f should approach arbitrarily close to $f(a)$. Mathematical convention usually switches the places of ϵ and δ above, but we present this for consistency with the definition of (ϵ, δ) continuity (and so (ϵ, δ) individual fairness)[108]. Uniform continuity implies continuity as the differences in definitions are a change in quantifier order: if f is uniformly continuous, then for any chosen δ, the according ϵ satisfies continuity at any $a \in M_1$ (whereas continuity may require ϵ to be dependent on a).

Proposition 4. *If $f : M_1 \to M_2$ is K-Lipschitz continuous, then f is uniformly continuous (and hence continuous).*

Proof. Given arbitrary δ, set $\epsilon = \frac{\delta}{K}$. Then given any point $a \in A$, for any $b \in A$ satisfying $d_1(a, b) < \epsilon$ we have by Lipschitz continuity $d_2(f(a), f(b)) \leq K\epsilon = \delta$, satisfying uniform continuity by definition. □

What we have defined as (ϵ, δ)-continuity (with respect to d_1 and d_2) in Part 2 does not imply continuity or the Lipschitz condition. The idea is that if we

look at small neighborhoods of width ϵ in M_1, any two points mapped by f within this neighborhood should diverge by less than δ; however, those points do not necessarily get arbitrarily close. For instance, choose some $\epsilon, \delta^* > 0$, let $M_1 = M_2 = \mathbb{R}$ with the standard (absolute value) metric, and consider f defined as $f(a) = 0$ if $a \leq 0$ and $\frac{\delta^*}{2}$ otherwise. f satisfies (ϵ, δ^*)-continuity since no two values of f differ by more than $\frac{\delta^*}{2}$. However, f is not continuous: let $\delta = \frac{\delta^*}{3}$, then for any $\epsilon > 0$, if we let $a = 0$ and $b = \frac{\epsilon}{2}$ we have $|a - b| < \epsilon$, but $|f(a) - f(b)| = \frac{\delta^*}{2} > \delta$. It follows that f cannot be uniformly continuous or K-Lipschitz continuous. However, we do have a result in the opposite direction.

Proposition 5. *If $f : M_1 \to M_2$ is K-Lipschitz continuous, then for any $\epsilon > 0$, f is $(\epsilon, K\epsilon)$-continuous.*

The result immediately follows from the definition of the Lipschitz property. For if we are guaranteed that any $a, b \in M_1$ such that $d_1(a, b) < \epsilon$ satisfy $d_2(f(a), f(b)) < K\epsilon$, we have immediately met the condition of $(\epsilon, K\epsilon)$-continuity. In fact, uniform continuity also implies (ϵ, δ)-continuity, though the relationship is more general. For consider any $\delta > 0$. Uniform continuity will give us ϵ such that for any two points a, b, $d_1(a, b) < \epsilon \implies d_2(f(a), f(b)) < \delta$, i.e. exactly (ϵ, δ)-continuity. So we can write:

Proposition 6. *If $f : M_1 \to M_2$ is uniformly continuous, it is also (ϵ^*, δ^*)-continuous for all ϵ^*, δ^* in the set $\{(\epsilon, \delta) : \epsilon = UC(\delta)\}$, where we denote UC as the function that, given a value of δ, outputs a suitable corresponding value of ϵ from the definition of uniform continuity for f.*

Continuity does not, however, imply (ϵ, δ)-continuity. As a counterexample, again take $M_1 = M_2 = \mathbb{R}$ with the standard metric, and consider $f(a) = a^2$. f is continuous as for any $a \in A$ and $\delta > 0$, set $\epsilon = \min(1, |\frac{\delta}{2a+1}|)$; then any b within ϵ of a will satisfy $|f(a) - f(b)| = |a^2 - b^2| = |a + b| \cdot |a - b| < |2a + \delta| \cdot \frac{\delta}{|2a+1|} \leq |2a + 1| \cdot \frac{\delta}{|2a+1|} = \delta$. However, suppose by way of contradiction that f is (ϵ^*, δ^*)-continuous for some fixed ϵ^* and δ^*. Choose any $a > \frac{\delta^*}{\epsilon^*}$, and consider the two points $a, a + \frac{\epsilon^*}{2} \in M_1$. Clearly the distance between these two points is less than ϵ^*. However, we have that $|f(a) - f(a + \frac{\epsilon^*}{2})| = (a + \frac{\epsilon^*}{2})^2 - a^2 = (2a + \frac{\epsilon^*}{2}) \cdot \frac{\epsilon^*}{2} > 2a \cdot \frac{\epsilon^*}{2} = a\epsilon^*$. By the choice of a, this is greater than δ^*, and so it cannot be the case that f is (ϵ^*, δ^*)-continuous for any ϵ^* or δ^*.

K-Lipschitz continuity is essentially the strongest condition out of what we have discussed here, since it implies (uniform) continuity as above and also guarantees us a version of (ϵ, δ)-continuity. In this way it provides us some further intuitive evidence for why it might be natural to gravitate to a Lipschitz condition in the definition of fairness through awareness. One other thematic takeaway is that the innovation in conceiving of individual fairness is the application of well-understood mathematical structures to wholly new domains, allowing us to carry over established results into new paradigms.

4.3 An Alternate Definition of Existential DE

Recall the definition for existential DE given in Part 2: following the setup, θ is ϵ-*existentially DE* if $|PE_1^\theta - PE_2^\theta| > \epsilon$, where $PE_i^\theta = \frac{1}{|X_i|} \sum_{x \in X_i} I_{\{x^\theta = \emptyset\}}$ is the proportion of values of θ in X_i that are $\emptyset$. We mentioned that this definition

suffers from small sample variability, in which small group sizes might increase the variability of PE_i^θ and thus make the difference $|PE_1^\theta - PE_2^\theta|$ more likely to exceed ϵ from statistical noise.

To adjust for this, we can borrow from the literature on statistical inference. Suppose that we model, for an individual x in group X_i, the likelihood that $x^\theta = \emptyset$ as p_i, determined by the group: $x^\theta \sim \text{Bern}(p_i) \,|\, x \in X_i$. In other words, the proportion of empty values for θ within X_i is p_i. We expect that p_i should be equal across groups; if this is not the case, then there is some issue with getting values of θ across groups. Checking if $p_1 = p_2$ is a situation that the *two-proportion Z test* is designed for, which gives us an alternate definition.

Definition 24. Let PE_1^θ and PE_2^θ denote the proportion of empty values of θ in X_1 and X_2, respectively, and let $PE^\theta = \frac{1}{|X|}\sum_{x \in X} I_{\{x^\theta = \emptyset\}}$ denote the proportion of empty values of θ overall. Let $n_1 = |X_1|$ and $n_2 = |X_2|$. Define z as

$$z = \frac{PE_1^\theta - PE_2^\theta}{\sqrt{PE^\theta(1 - PE^\theta)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

Then we say θ is ϵ-*existentially statistically DE* if $P(|Z| > z) < \epsilon$, where $Z \sim \mathcal{N}(0, 1)$ is a standard Normal random variable.

Specifically, z as above is the two-tailed test statistic for testing the hypothesis $H_0 : PE_1^\theta = PE_2^\theta$. Assuming H_0, it can be shown that z follows the standard Normal distribution. Thus, extreme values of z, both positive and negative, are an indication that it is likely that PE_1^θ and PE_2^θ are different. Typically in statistical testing, ϵ as we have defined it here is referred to as a *p-value*, and $\epsilon = 0.05$ is the threshold for determining whether or not z is significant.[104] Also, since the

definition of existential statistical DE only requires us to know PE_1^θ and PE_2^θ, which can be obtained through the empirical distributions of θ within the groups (μ_i^θ), we can express the quantity $P(|Z| > z)$ as given above in terms of a function D of μ_1^θ and μ_2^θ. Consequently, just like existential DE, this is also a special case of distributional DE.

4.4 Miscellaneous Results

Here are some proofs that were omitted from the mathematical setup in Part 2.

Proposition. *Let X be a finite set with groups $\{X_i\}$ and let f be a map from X to ΔA. Then the outcome probability measure, defined for X_i as the empirical average of outcomes over X_i, $\mu_i^{out} = \frac{\sum_{x \in X_i} f(x)}{|X_i|}$, is a probability measure on A.*

Proof. μ_i^{out} assigns to $a \in A$ the mass $\frac{1}{|X_i|} \sum_{x \in X_i} (f(x))(a)$, interpreting $f(x)$ as a probability measure in ΔA. Thus it defines the probability "mass" of singleton elements in A. Then to prove $\mu_i^{out} \in A$, it suffices to show that the sum of these masses, i.e. the measure of A, equals 1, which also follows from $f(x)$ being a probability measure:

$$\begin{aligned}
\mu_i^{out}(A) &= \sum_{a \in A} \mu_i^{out}(a) \\
&= \sum_{a \in A} \left(\frac{1}{|X_i|} \sum_{x \in X_i} (f(x))(a) \right) \\
&= \frac{1}{|X_i|} \sum_{x \in X_i} \sum_{a \in A} (f(x))(a) \\
&= \frac{1}{|X_i|} \sum_{x \in X_i} 1 \\
&= 1
\end{aligned}$$

which is what we wanted. □

Observe that outcome probability measures are members of ΔA (in contrast with induced probability measures, which are over X). We also note that in the more general setting where groups are probability distributions over V (which we do not directly address), Dwork et al. write $\mu_i^{out} = E_{x \sim X_i} f(x)$, replacing an empirical average for an expectation[28].

Proposition. *Suppose X has metric d_X and probability measure μ_X. Then $\mathcal{X}$, the group space, is a metric probability space with respect to the metric $d_{\mathcal{X}}(X_i, X_j) = \mathcal{W}_d(X_i, X_j)$ and the naturally-induced probability measure $\mu_{\mathcal{X}}$ defined by $\mu_{\mathcal{X}}(X_i) = \mu_X(X_i) = \sum_{x \in X_i} \mu_X(x)$.*

Proof. We take $\mathcal{W}_d(X_i, X_j)$ to mean $\mathcal{W}_d(\mu_i, \mu_j)$, where μ_i is the induced probability measure for X_i. Observe that if we identify X_i with μ_i, we get an isomorphism of sets between $\mathcal{X}$ and the set of induced group probability measures $\{\mu_i\}_{i=1}^g$,

which is a subset of ΔX. The fact that $d_{\mathcal{X}}$ is a metric over $\mathcal{X}$ then follows from the above fact that $\mathcal{W}_d$ is a metric over ΔX and subsets of metric spaces are metric spaces.

We have also specified in the definition of $\mu_{\mathcal{X}}$ how it assigns probability "mass" to the singleton elements of $\mathcal{X}$ directly in terms of μ_X: $\mu_{\mathcal{X}}(X_i) = \mu_X(X_i) = \sum_{x \in X_i} \mu_X(x)$. Since the X_i form a partition of X, it is immediate that $\mu_{\mathcal{X}}(\mathcal{X}) = \sum_i \sum_{x \in X_i} \mu_X(x) = \sum_{x \in X} \mu_X(x) = 1$. In essence, $\mu_{\mathcal{X}}$ is a more quantized version of μ_X, and its definition is nothing more than a notational trick using the fact that we consider X_i both as subsets of X and as elements of $\mathcal{X}$. This makes $\mu_{\mathcal{X}}$ a probability measure on $\mathcal{X}$. □

Here we can also extend this concept to handle the more general case where X is continuous instead of finite and X_i are also continuous subsets by considering the general definition of the Wasserstein (or earthmover) distance, $\mathcal{W}_d(\mu_1, \mu_2) = \inf_{\nu \in U} \int_{a,b \in M} d(a,b)\nu(a,b)$. As μ_X is then a measure over a continuous set, the definition $\mu_{\mathcal{X}}(X_i) = \mu_X(X_i)$ remains unchanged.

4.5 Catalog of Probability Measures

Assume X is a finite set of representations of individuals, split into groups $\{X_i\}$. Let $\mathcal{X}$ denote the group space where each each element identifies a group X_i. θ denotes a feature within each representation taking values in Θ. A task is a map $f : X \to R$.

We have encountered various different (probability) measures in this thesis which we summarize here. Recall that as a probability distribution can be defined

by a probability measure, in this thesis we have treated a measure μ as synonymous with the distribution it represents. Also note that we take "distribution of θ" to mean "distribution of values of θ over $\tilde{\Theta}$".

- $\mu_X \in \Delta X$ is the probability measure with which X is endowed. If it is not defined *a priori*, then we take μ_X to be the measure for the uniform distribution, $\mu_X(x) = \frac{1}{|X|}$.
- Given μ_X, $\mu_i \in \Delta X$ is the *induced probability measure* for X_i, defined as $\mu_i(x) = \frac{\mu_X}{\mu_X(X_i)}$. It is used to define Wasserstein distances between groups.
- Assume $R = \Delta A$, where A is a finite set of classes. Given f, $\mu_i^{out} \in \Delta A$ is the *outcome probability measure* for X_i, defined as the empirical average of outcomes in the group, $\mu_i^{out}(a) = \frac{1}{|X_i|} \sum_{x \in X_i} (f(x))(a)$.
- Given μ_X, $\mu_{\mathcal{X}} \in \Delta \mathcal{X}$ is the probability measure on the group space $\mathcal{X}$, defined as $\mu_{\mathcal{X}}(X_i) = \mu_X(X_i)$ where X_i is treated on the left-hand side as an element of $\mathcal{X}$ and on the right-hand side as a group (subset of X).
- Given a feature $\theta \in \tilde{\Theta}$, $\mu_i^\theta \in \Delta\tilde{\Theta}$ is the probability measure of the empirical distribution (over values) of θ within X_i.
- Given an underlying feature $\theta^* \in \Theta^*$, $\mu^* \in \Delta\Theta^*$ is the probability measure for the underlying distribution, a distribution (over values) of θ^* which we assume is the same across groups. Then the empirical distribution of a proxy feature θ, μ_i^θ, can be determined given μ^* and the semantic relation R_i^θ for the group X_i.

References

[1] Abbott, S. (2015). *Understanding Analysis*. Springer.

[2] Allingham, M. (2024). Distributive justice. https://iep.utm.edu/distributive-justice/.

[3] Alves, G., Bernier, F., Couceiro, M., Makhlouf, K., Palamidessi, C., & Zhioua, S. (2023). Survey on fairness notions and related tensions. *EURO Journal on Decision Processes*, 11, 100033.

[4] Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016a). How we analyzed the COMPAS recidivism algorithm.

[5] Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016b). Machine bias. ProPublica.

[6] Arnold, C. (2022). How biased data and algorithms can harm health. https://magazine.jhsph.edu/2022/how-biased-data-and-algorithms-can-harm-health.

[7] Attygalle, K., Hodgson-Bautista, J., & Hopson, R. (2023). Inequities persist: Extracurriculars, clubs, activities, and fundraising in ontario's publicly funded schools.

[8] Barocas, S. & Selbst, A. (2016). Big data's disparate impact. *California Law Review*, 104.

[9] Barry-Jester, A. M., Casselman, B., & Goldstein, D. (2015). The new science of sentencing.

[10] Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2019). Consumer-lending discrimination in the fintech era.

[11] Bazarbash, M. (2019). *FinTech in Financial Inclusion: Machine Learning Applications in Assessing Credit Risk*. Technical report, International Monetary Fund.

[12] Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2017). Fairness in criminal justice risk assessments: The state of the art.

[13] Bickel, P. J., Hammel, E. A., & O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from berkeley. *Science*, 187(4175), 398–404.

[14] Binns, R. (2019). On the apparent conflict between individual and group fairness.

[15] Boulware, L. E., Mohottige, D., & Maciejewski, M. L. (2022). Race-Free Estimation of Kidney Function: Clearing the Path Toward Kidney Health Equity. *JAMA*, 327(23), 2289–2291.

[16] Calmon, F. P., Wei, D., Ramamurthy, K. N., & Varshney, K. R. (2017). Optimized data pre-processing for discrimination prevention.

[17] Calsamiglia, C. (2005). *Decentralizing Equality of Opportunity and Issues Concerning the Equality of Educational Opportunity*. PhD thesis, Yale University.

[18] Cavanagh, M. (2002). *Against Equality of Opportunity*. Clarendon Press.

[19] Chatterji, R., Campbell, N., & Quirk, A. (2021). *Closing Advanced Coursework Equity Gaps for All Students*. Technical report, Center for American Progress.

[20] Chiappa, S. & Gillam, T. P. S. (2018). Path-specific counterfactual fairness.

[21] Chouldechova, A. (2016). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments.

[22] Clayton, A. (2020). How eugenics shaped statistics. https://nautil.us/how-eugenics-shaped-statistics-238014/.

[23] Clement, P. & Desch, W. (2008). An elementary proof of the triangle inequality for the wasserstein metric. *Proceedings of The American Mathematical Society - PROC AMER MATH SOC*, 136, 333–340.

[24] College Board (2021). SAT program results capture impact of COVID on class of 2021. https://newsroom.collegeboard.org/sat-program-results-capture-impact-of-covid-on-class-of-2021.

[25] Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness.

[26] Dieterich, W., Mendoza, C., & Brennan, T. (2016). *COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity*. Technical report, Northpointe Inc. Research Department.

[27] Durham District School Board (2019). Gifted program. https://www.ddsb.ca/en/programs-and-learning/gifted-program.aspx.

[28] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2011). Fairness through awareness.

[29] Dwork, C., Kim, M. P., Reingold, O., Rothblum, G. N., & Yona, G. (2020). Outcome indistinguishability.

[30] Dwork, C. & Mulligan, D. K. (2013). It's not privacy, and it's not fair. *Stanford Law Review*.

[31] Dwork, C., Reingold, O., & Rothblum, G. N. (2023). From the real towards the ideal: risk prediction in a better world. In *4th Symposium on Foundations of Responsible Computing (FORC 2023)*: Schloss Dagstuhl-Leibniz-Zentrum für Informatik Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

[32] Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.

[33] Fabris, A., Messina, S., Silvello, G., & Susto, G. A. (2022). Algorithmic fairness datasets: the story so far. *Data Mining and Knowledge Discovery*, 36(6), 2074–2152.

[34] Feldman, M., Friedler, S., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact.

[35] FFEIC (2022). Number of mortgage applications in the united states in 2021, by race and outcome (in 1,000s). https://www-statista-com.ezp-prod1.hul.harvard.edu/statistics/1175267/mortgage-application-volume-by-race-and-outcome-usa.

[36] Freedle, R. (2008). Correcting the SAT's Ethnic and Social-Class Bias: A Method for Reestimating SAT Scores. *Harvard Educational Review*, 73(1), 1–43.

[37] Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2016). On the (im)possibility of fairness.

[38] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., au2, H. D. I., & Crawford, K. (2021). Datasheets for datasets.

[39] Gibbs, A. L. & Su, F. E. (2002). On choosing and bounding probability metrics. *International Statistical Review / Revue Internationale de Statistique*, 70(3), 419–435.

[40] Glymour, C. & Glymour, M. R. (2014). Commentary: Race and sex are causes. *Epidemiology*, 25(4).

[41] Gosepath, S. (2021). Equality. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2021 edition.

[42] Grgic-Hlaca, N., Zafar, M. B., Gummadi, K. P., & Weller, A. (2016). The case for process fairness in learning: Feature selection for fair decision making.

[43] Gutiérrez, O. M., Sang, Y., Grams, M. E., Ballew, S. H., Surapaneni, A., Matsushita, K., Go, A. S., Shlipak, M. G., Inker, L. A., Eneanya, N. D., Crews, D. C., Powe, N. R., Levey, A. S., Coresh, J., & Chronic Kidney Disease Prognosis Consortium (2022). Association of estimated GFR calculated using race-free equations with kidney failure and mortality by black vs non-black race. *JAMA*, 327(23), 2306–2316.

[44] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning.

[45] Heilweil, R. (2020). Why algorithms can be racist and sexist. https://www.vox.com/recode/2020/2/18/21121286/algorithms-bias-discrimination-facial-recognition-transparency.

[46] Hengtgen, K. & Morales, K. L. (2022). *The Leaky Pipeline of Advanced Placement Testing*. Technical report, Urban Institute.

[47] Holland, S., Hosny, A., Newman, S., Joseph, J., & Chmielinski, K. (2018). The dataset nutrition label: A framework to drive higher data quality standards.

[48] (https://mathoverflow.net/users/83321/magic a) (2018). Total variation and relative ℓ_∞ metric. MathOverflow. URL: https://mathoverflow.net/q/293998 (version: 2018-03-06).

[49] Hu, L. & Kohler-Hausmann, I. (2020). What's sex got to do with machine learning? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20: ACM.

[50] Huttenlocher, D. P., Klanderman, G. A., & Rucklidge, W. J. (1993). Comparing images using the hausdorff distance. *IEEE Transactions on pattern analysis and machine intelligence*, 15(9), 850–863.

[51] III, C. N. (2008). The shadow of credit: the historical origins of acial predatory lending and its impact upon african american wealth accumulation. *University of Pennsylvania Journal of Law and Social Change*.

[52] Ilvento, C. (2020). Metric learning for individual fairness.

[53] Jackson, E. & Mendoza, C. (2020). Setting the Record Straight: What the COMPAS Core Risk and Need Assessment Is and Is Not. *Harvard Data Science Review*, 2(1). https://hdsr.mitpress.mit.edu/pub/hzwo7ax4.

[54] Jagtiani, J. & Lemieux, C. (2019). The roles of alternative data and machine learning in fintech lending: Evidence from the lendingclub consumer platform. *Financial Management*, 48(4), 1009–1029.

[55] Jernigan, C. & Mistree, B. F. (2009). Gaydar: Facebook friendships expose sexual orientation. *First Monday*, 14(10).

[56] Kamiran, F. & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1), 1–33.

[57] Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness.

[58] Kim-Christian, P. & McDermott, L. (2022). *Disparities in Advanced Placement Course Enrollment and Test Taking: National and State-Level Perspectives*. Technical report, Urban Institute.

[59] Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2019). Discrimination in the Age of Algorithms. *Journal of Legal Analysis*, 10, 113–174.

[60] Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores.

[61] Kohler-Hausmann, I. (2019). Eddie Murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. *Northwestern University Law Review*, 113.

[62] Kohler-Hausmann, I. (2023a). What does 'race neutral' admissions mean? *Yale Law School, Public Law Research Paper*.

[63] Kohler-Hausmann, I. (2023b). What just got banned? Acting on the basis of race and treating people as equals. *Arizona Law Review*, 66.

[64] Kusner, M. J., Loftus, J. R., Russell, C., & Silva, R. (2018). Counterfactual fairness.

[65] Lamont, J. & Favor, C. (2017). Distributive Justice. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2017 edition.

[66] Larson, J. & Angwin, J. (2016). Technical response to northpointe.

[67] Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., & Hardt, M. (2018). Delayed impact of fair machine learning.

[68] Lohia, P. K., Ramamurthy, K. N., Bhide, M., Saha, D., Varshney, K. R., & Puri, R. (2018). Bias mitigation post-processing for individual and group fairness.

[69] Lynch, E. E., Malcoe, L. H., Laurent, S. E., Richardson, J., Mitchell, B. C., Helen, & Meier, C. (2021). The legacy of structural racism: Associations between historic redlining, current mortgage lending, and health. *SSIM Population Health*.

[70] Manrai, A. K., Funke, B. H., Rehm, H. L., Olesen, M. S., Maron, B. A., Szolovits, P., Margulies, D. M., Loscalzo, J., & Kohane, I. S. (2016). Genetic misdiagnoses and the potential for health disparities. *New England Journal of Medicine*, 375(7), 655–665.

[71] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022). A survey on bias and fairness in machine learning.

[72] Mémoli, F. (2011). Gromov–wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics*, 11(4), 417–487.

[73] Mémoli, F. (2014). The gromov–wasserstein distance: A brief overview. *Axioms*, 3(3), 335–341.

[74] Mhasawade, V., Zhao, Y., & Chunara, R. (2021). Machine learning and algorithmic fairness in public and population health. *Nature Machine Intelligence*, 3(8), 659–666.

[75] Mukherjee, D., Yurochkin, M., Banerjee, M., & Sun, Y. (2020). Two simple ways to learn individual fairness metrics from data.

[76] Neil, R. & Winship, C. (2019). Methodological challenges and opportunities in testing for racial discrimination in policing. *Annual Review of Criminology*, 2(Volume 2, 2019), 73–98.

[77] Nissenbaum, H. (2004). Privacy as contextual integrity. *Washington Law Review*, 79(1).

[78] Nozick, R. (1974). *Anarchy, State, and Utopia*. Basic Books.

[79] Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.-E., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., Kompatsiaris, I., Kinder-Kurlanda, K., Wagner, C., Karimi, F., Fernandez, M., Alani, H., Berendt, B., Kruegel, T., Heinze, C., Broelemann, K., Kasneci, G., Tiropanis, T., & Staab, S. (2020). Bias in data-driven ai systems – an introductory survey.

[80] Ochigame, R. (2020). The long history of algorithmic fairness. https://www.phenomenalworld.org/analysis/long-history-algorithmic-fairness/.

[81] Osoba, O. & IV, W. W. (2017). *An intelligence in our image: The risks of bias and errors in artificial intelligence*. Rand Corporation.

[82] Panch, T., Mattie, H., & Atun, R. (2019). Artificial intelligence and algorithmic bias: implications for health systems. *Journal of Global Health*.

[83] Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, 3(none), 96 – 146.

[84] Rawls, J. (1971). *A Theory of Justice*. Belknap Press.

[85] Rawls, J. (2001). *Justice As Fairness: A Restatement*. Harvard University Press.

[86] Reyna, C. (2000). Lazy, dumb, or industrious: When stereotypes convey attribution information in the classroom. *Educational Psychology Review*, 12(1), 85–110.

[87] Roemer, J. (1996). *Theories of Distributive Justice*. Elsevier.

[88] Roemer, J. & Trannoy, A. (2015). *Equality of Opportunity*, chapter 4. Elsevier.

[89] Roth, S. (2016). *Modern Discrete Probability: An Essential Toolkit*. Cambridge University Press.

[90] Rudin, C., Wang, C., & Coker, B. (2020). The Age of Secrecy and Unfairness in Recidivism Prediction. *Harvard Data Science Review*, 2(1). https://hdsr.mitpress.mit.edu/pub/7z10o269.

[91] Santelices, M. V. & Wilson, M. (2010). Unfair Treatment? The Case of Freedle, the SAT, and the Standardization Approach to Differential Item Functioning. *Harvard Educational Review*, 80(1), 106–134.

[92] Sawada, T., Paleka, D., Havrilla, A., Tadepalli, P., Vidas, P., Kranias, A., Nay, J. J., Gupta, K., & Komatsuzaki, A. (2023). Arb: Advanced reasoning benchmark for large language models.

[93] Shi, S., Tse, R., Luo, W., D'Addona, S., & Pau, G. (2022). Machine learning-driven credit risk: a systemic review. *Neural Computing and Applications*, 34(17), 14327–14339.

[94] Simoiu, C., Corbett-Davies, S., & Goel, S. (2017). The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics*, 11(3), 1193 – 1216.

[95] Stojanović, B., Božić, J., Hofer-Schmitz, K., Nahrgang, K., Weber, A., Badii, A., Sundaram, M., Jordan, E., & Runevic, J. (2021). Follow the trail: Machine learning for fraud detection in fintech applications. *Sensors*, 21(5).

[96] Sullivan, E. & Greene, R. (2015). States predict inmates' future crimes with secretive surveys.

[97] Supreme Court of the United States (2023). Students for fair admissions, inc. v. president and fellows of harvard college. https://www.supremecourt.gov/opinions/22pdf/20-1199_hgdj.pdf.

[98] Suresh, H. & Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21: ACM.

[99] United States Sentencing Commission (2023). 2022 federal sentencing statistics. https://www.ussc.gov/research/data-reports/geography/2022-federal-sentencing-statistics.

[100] Verma, S. & Rubin, J. (2018). Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, FairWare '18 (pp. 1–7). New York, NY, USA: Association for Computing Machinery.

[101] Vyas, D. A., Eisenstein, L. G., & Jones, D. S. (2020). Hidden in plain sight — reconsidering the use of race correction in clinical algorithms. *New England Journal of Medicine*, 383(9), 874–882. PMID: 32853499.

[102] Wang, X., Zhang, Y., & Zhu, R. (2022). A brief review on algorithmic fairness. *Management System Engineering*, 1(1), 7.

[103] Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference*. Springer.

[104] Webb, R. (2019). *Mostly Harmless Statistics*. Portland State University Library.

[105] Wenar, L. (2021). John Rawls. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2021 edition.

[106] Williams, D. R. & Rucker, T. D. (2000). Understanding and addressing racial disparities in health care. *Health Care Financing Review*.

[107] Winship, C. & Mare, R. D. (1992). Models for sample selection bias. *Annual Review of Sociology*, 18(Volume 18, 1992), 327–350.

[108] Xu, S. & Strohmer, T. (2024). On the (in)compatibility between group fairness and individual fairness.

[109] Young, H. P. (1995). *Equity: In Theory and Practice*. Princeton University Press.

[110] Zafar, M. B., Valera, I., Rodriguez, M. G., & Gummadi, K. P. (2017). Fairness constraints: Mechanisms for fair classification.

[111] Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. In S. Dasgupta & D. McAllester (Eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research* (pp. 325–333). Atlanta, Georgia, USA: PMLR.

[112] Úrsula Hébert-Johnson, Kim, M. P., Reingold, O., & Rothblum, G. N. (2018). Multicalibration: Calibration for the (computationally-identifiable) masses.

This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. This thesis was based on a PhD dissertation template that has been released under the permissive MIT (X11) license, and can be found online at github.com/suchow/Dissertate.