Methodological Blind Spots in Machine Learning Fairness:

Lessons from the Philosophy of Science and Computer Science

Samuel Deng

Department of Philosophy, Columbia University

Advisor: Achille Varzi

UNI: sd3013

Phone: (626) 363 – 5118

April 1st, 2019

Acknowledgments

The completion of this thesis would not have been possible without the knowledge and skills, both technical and non-technical, that I have gained from the Department of Computer Science and Department of Philosophy at Columbia University. I am indebted to Professor Nakul Verma for introducing me to the exciting field of machine learning through his excellently taught class and our early discussions on machine learning fairness. I am also indebted to my friends at Columbia University and my family in California, who have invariably supported me in times of doubt and hardship through this year and the past three years of my undergraduate career.

Lastly, I am, of course, incredibly grateful for the supervision and wisdom of Professor Achille Varzi, whose guidance and instruction this year and the past three years have instilled in me a love of philosophy. I have learned, most of all, that philosophy is invaluable in keeping an open and curious mind. I will never forget one particular nugget of wisdom from his final lecture in *Metaphysics* (poorly paraphrased) – we should keep our realm of possibilities far wider than we could imagine. To use his ending quote from that last lecture: “There are more things in heaven and earth, Horatio, / Than are dreamt of in your philosophy.”

1. Introduction

In our current age of technological progress, artificial intelligence (AI) and machine learning (ML) technologies have become increasingly integrated into decisions previously made by humans. Though the original purpose of integrating ML technology into high stakes decision-making was to rid certain important decisions such as hiring or judicial procedures from human error and biases, it was soon clear that ML oftentimes proliferated and *increased* bias in decision-making (Hardt and Barocas). This spurred a surge in the machine learning research community to explore "fair machine learning," the paradigm of creating ML systems that detect and mitigate the hidden biases in ML-reliant decision-making. My aim in this paper is to explore the current state of "fair machine learning" from a philosophical point of view in order to tease out certain methodological "blind spots" that may benefit from philosophical investigation.

First, I must provide some standard definitions of the technical terms at hand. Though AI is an extremely broad and rapidly developing field in computer science, we borrow a simple definition from Russel and Norvig's classic computer science textbook on the subject: "[Artificial intelligence] is concerned with intelligent behavior in artifacts" (2). We might even narrow this definition to say that artificial intelligence is the capability of an artifact to seemingly perform cognitive tasks normally associated with human intelligence, such as visual perception or problem solving. With this definition, we can characterize the many subfields of AI with their human analogues: for instance, computer vision allows machines to perceive the visual world, or natural language processing allows machines to understand human language. However, I focus on one particular subfield of AI that drives most of its applications and research today: machine learning.

Machine learning (ML) is the specific subfield of AI focused on using datasets from the real world to allow machines to generalize, recognize patterns, and make inferences *without* specific and explicit instructions. That is, however, not to say that machine learning systems are autonomous or “conscious” – machine learning systems are mathematical models from real world data that, if made correctly, can generalize to new, unseen data using the patterns from seen data. For instance, suppose all of your interactions on an online shopping website are recorded as data – these include what you purchase, your viewed items, and the items you view but do not purchase. Using all this data, the website might implement an ML algorithm that, given all your previous clicks and views, shows you what product you *might* be interested in. This differs from the traditional notion of an algorithm in computer science (essentially, a set of human-made rules given to a computer to accomplish a specific goal) because the “rules” are not explicitly programmed from the start; instead, the “rules” are learned from the data. Instead of explicit instructions such as, “if user X clicked ‘Gardening Supplies’ more than 5 times, suggest ‘Garden Hose,’” machine learning algorithms “learn” these unique rules for each dataset. Despite this distinction, the term “algorithmic decision-making” will, in this paper, refer to decisions made by ML algorithms, as those will drive our main concerns for fairness.

So, what is the “fair machine learning” problem? In the contemporary boom of ML research and applications, algorithmic decision-making was once thought of as a solution to replace humans in different scenarios prone to human error and bias. News feeds, social networks, and online shopping sites, to name a few, populate before us with tailor-made recommendations that grow more and more uncanny. Hiring managers and job recruiters have begun delegating work to ML algorithms in efforts to reduce human error and increase efficiency (Whittaker et. al. 38). Even decisions once made by human judges are moving into the hands of algorithms, on the grounds that human judges have their own racial or social biases that may conflict with a just decision (Angwin et. al.). However, the adverse effects of algorithmic decision-making also go unnoticed. In each case, algorithmic decision-making proved to be flawed: the very recommendations designed to enrich social network participant experience end up biasing against women (Martineau 1), Amazon's AI hiring algorithm was found to be sexist (Whittaker et. al. 38), and tools in courts proved to be more random and discriminatory towards African Americans than human judges (Angwin et. al). These cases result from a variety of reasons, but the main reason these biases exist is because ML systems are made by *humans,* after all. Human bias exists in ML design, data, and the entire pipeline and, because of this, ML systems generalize to fit those biases. Machine bias stems from *human* bias. In a sense, ML holds a mirror up to humanity's own decision-making proclivities.

One particular case that was particularly well-noted in the machine learning community and broke into general news was the COMPAS tool (Angwin et. al.). In an article that galvanized much of the current research in fair machine learning, ProPublica released reports that COMPAS, a software used to predict a risk score for recidivism (likelihood of committing a future crime), falsely labelled African American defendants as future criminals at twice the rate of white defendants. It simultaneously mislabeled white defendants as low risk more often than African American defendants. ProPublica found that the predictions were, indeed, false positives and false negatives from examining who *actually* recidivated two years after COMPAS gave these scores. These risk scores are meant to be unbiased, mathematically valid predictors of later behavior, and, thus, are used as brute fact by judges in many states in the US. Instead, COMPAS seemingly encoded an already prevalent racial bias into its decision making.

Because ML algorithms are increasingly implemented to deal with consequential decisions such as hiring, legal judgment, and policing, a nascent subfield in the technical machine learning community has developed, researching what it means for a machine learning algorithm to be fair. In recent computer science literature, varying definitions of fairness in mathematical terms have arisen as constraints on these algorithms. Reuben Binns argues in "Fairness in Machine Learning: Lessons from Political Philosophy" that, despite being internally and logically consistent through the math, each mathematical definition assumes a different *philosophical* view on what "fairness" means (1). For instance, should algorithms give every possible group the same probability for certain outputs? Should algorithms strive to reduce and minimize the negative impacts on more disadvantaged groups? Or maybe algorithms should make use of some ideal state where no discrimination or unfair disadvantages exist? From these varying definitions, we are reminded of much older thought in ethics and political philosophy on what exactly the terms fairness, discrimination, or justice mean in the first place (Binns 2). So, it seems that, in order to create fair ML algorithms, it may be helpful to study these definitions through the lens of philosophy. However, though Binns' characterizations of these various fairness criteria (more in Section 2) from the perspective of moral and political philosophy are certainly promising and eye-opening, I take inspiration to use philosophy in a different scope – examining the *methodological* assumptions for fairness in ML in the first place.

Therefore, in this paper, I will explore three main *methodological* assumptions in ML worth examining if we aim to achieve fairness. First, I provide a brief survey of the very basics of ML and the fairness definitions in the technical literature, explaining technical terms in the current state of ML fairness alongside brief non-technical analogues (Section 2). We will learn two important lessons from this: "the impossibility theorem" that states that it is mathematically impossible to implement more than one fairness definition at once, and the fact that attempting to solve the fair-ML problem through choosing a fairness definition necessarily forces us to commit to a strong normative view on fairness. This teases out a need for a shift in thinking for fair-ML – focus on the *methodology* instead of the existing solutions. I then argue that, instead of looking at the existing solutions and fairness criteria, it may be fruitful to turn our eyes to the *methodology* of the ML paradigm, design process, and implementation through the lens of philosophy (Section 3). In order to embark on this philosophical investigation, I make an argument that fairness is a *context-sensitive value,* an intuitive assumption that underlies the paper (Section 4). From this assumption, I propose three main methodological components of ML that, through philosophical scrutiny via the philosophy of science and computer science, conflict with *context-sensitive* fairness: abstraction, induction, and measurement (Sections 5, 6, 7). I argue that these are often overlooked and taken-for-granted "blind spots" for ML practitioners and the current field of fair-ML. Finally, I propose some very nascent suggestions for the field of fair-ML to address these latent blind spots (Section 8). The purpose of this paper is thus to use philosophy, an often-overlooked perspective in technical fields, to better understand and tease out latent issues in ML that might only see the light of day through philosophical investigation. Nonetheless, these latent issues have a serious impact on philosophers and non-philosophers, ML practitioners and non-practitioners, alike as algorithmic decision-making and ML systems grow more ubiquitous in the Digital Age.

2. Overview of ML and Fairness Definitions

In this section, I will provide a brief introduction to ML and survey the current definitions of fairness in the ML community to contextualize the rest of this paper, using some basic mathematics and probability. I will classify the existing *group fairness* definitions (where most of the research has been so far) into three main general classes: Independence, Separation, and Sufficiency. Though this is by no means exhaustive, many fairness definitions in the literature fall into these three main categories. I attempt to briefly survey two more main definitions after this: counterfactual fairness and individual fairness. I will conclude the section by explaining the "impossibility theorem," a mathematical constraint that states, roughly, that no two definitions can simultaneously hold for any nontrivial ML problem. From this survey, we gleam our first basic philosophical observation: each fairness definition poses a strong normative view on what fairness means in the philosophical sense, and, due to their incompatibility from the "impossibility theorem," any machine learning algorithm interested in fairness must necessarily include a normative decision on the nature of fairness.

Before surveying the fairness definitions, it is important to understand the main goal of ML and how a basic classification problem works. At the heart of machine learning is the task of using observed random variables X to predict the value of an unknown random variable Y. For instance, if we are given a loan applicant's credit history, credit score, and income (X variables), the algorithm gives some score or likelihood $\hat{Y}$ on how likely that applicant will pay off their loan. If we make a decision based on a certain threshold ("give loan if $\hat{Y}$ > 75% and reject otherwise"), then we have the basic *binary classification problem* – for every X, we output a result from the discrete set {-1, 1} (reject or accept). The goal of machine learning is to create a function $\hat{Y} = f(X)$ (a *classifier*) where $\hat{Y}$ predicts Y (the true *target variable*) accurately. A "good" $\hat{Y}$ is precisely the $\hat{Y}$ that maximizes:

$$\Pr\,[Y = \hat{Y}]$$

the probability of the true target variable Y matching the predicted value $\hat{Y}$**.** In our loan example, our machine learning algorithm thus gives us some function $\hat{Y} = f(X)$ learned from the data X about the applicant's financial history, and the closer $\hat{Y}$ (our prediction on whether the applicant pays off their loan) is to Y (whether the applicant *actually* pays off their loan or not), the better. In the binary case, if $\hat{Y} = Y$ then our ML algorithm succeeds for that particular applicant. We define $\Pr\,[Y = \hat{Y}]$ to be the *predictive accuracy* of our classifier. Thus, the goal for any ML algorithm or ML practitioner is to maximize predictive accuracy on the given data *X*. We will return to this notion later, but it is important to note here that ML practitioners focus on the standard of *accuracy* which, in itself, assumes that the desired outcome in all cases is ensuring that our predicted results match up with the patterns seen in already observed data.

Our formal fairness problem arises when algorithms use certain observed traits (let us denote them A) to maximize predictive accuracy of a classifier while human-made decisions that take these traits into account are seen as discriminatory or unjust (Barocas et. al.). In our loan example, we can take A to be *race of the applicant* and our classifier might be able to maximize $\Pr\,[Y = \hat{Y}]$ by strongly considering A in its decision-making process. This could result in rejecting a large number of a certain race of loan applicants while accepting a large number from another race. Why does this happen? An ML classifier can only learn from *training data* (the set of past labelled data), and if the training data contain a disproportionate rate of rejection for a certain race of loan applicants, the classifier will pick up on that pattern and generalize to classify future examples similarly. This disproportion in the training data comes from *human* biases – in our hypothetical example, loan applicants may have historically been rejected due to race, leading to the disproportion in the training data, leading to the classifier exhibiting the human bias. Thus, at the heart of the fairness problem lies this bottom line: a machine learning algorithm is *necessarily* only as good as the data you use to train it. The common misconception of ML and AI as objective is untrue – it is *objective* only in the sense of learning what the human behind it teaches (Zhong).

So, knowing this framework, we can arrive at a very naïve fairness definition: remove or ignore the attribute A in the dataset, and our ML classifier will be fair. This corresponds to our intuition that "justice is blind." However, this can result in no improvement to our machine learning task at best, and, at worst, can worsen the results (Hardt and Barocas). Because the datasets we work with in ML are necessarily huge (the larger the dataset, the more generalization a classifier can achieve), *other* features that correlate with attribute A encode the same human bias just as well as A may have. In our example, we may naively believe that we have arrived at a fair algorithm through simply discarding the *race* attribute A, but, in reality, attributes such as *geographic location* or *income* allow ML classification to achieve the same results. The failure of naively removing the attribute A in order to create a fair machine learning algorithm motivates the following fairness definitions.

The first formal fairness definition is *Independence* (also known as *demographic parity*), which, mathematically speaking, requires the sensitive attribute A to be statistically independent of the predicted value $\hat{Y}$ (Barocas et. al. 43):

$$\Pr\left[\hat{Y} = \hat{y} \mid A = a\right] = \Pr\left[\hat{Y} = \hat{y} \mid A = b\right] \quad \forall \hat{y} \in \{-1, 1\}$$

$$\hat{Y} \perp A$$

The first equation above is the definition of Independence for a simple binary classification problem, and the second expression is an alternative way to phrase Independence: the predicted value $\hat{Y}$ is *independent* of A. Simply put, Independence requires that, given any two values of A, the probability of our classifier returning either -1 or 1 (acceptance or denial, yes or no) is equal. For our loan example, this means that our classifier is constrained to have the same probability of denying applicants for all individuals that are black as the probability of denying applicants for all individuals that are white (the same applies for accepting applicants). This fairness definition allows *group fairness* – for any two groups of the sensitive attribute A, we equalize the odds of acceptance. Already, we see a normative claim on fairness – fairness works on the level of entire group attributes. If the rate of acceptance for *all groups* of a certain attribute are equal, then we have fair decision-making. This echoes the *collectivist egalitarian* motive behind affirmative action and similar practices that equalize rates of acceptance for various groups of people (Gajane and Pechenizkiy 3). This leads to our main problem with Independence – by constraining the rates of entire *groups* to be equal, we forego an intuitive sense of individual fairness or meritocracy. For instance, suppose we have two groups *a* and *b*. Then, suppose a company carefully hires from group *a* at some rate *p* > 0, keeping in mind attributes that likely allow individuals from group *a* to perform well on the job. To achieve the constraint for Independence, the company can carelessly and "lazily" select applicants from group *b* without much thought at rate *p* as well (Barocas et. al. 44). Acceptance rates are completely identical, fulfilling Independence and group fairness, but group *b* likely will perform much worse. This *collectivist egalitarianism* goes against our intuitions of meritocracy and efficiency (Gajane and Pechenizkiy 3), should we buy into those normative stances.

The second formal fairness condition is *Separation* (also known as *equalized odds* or *equality of opportunity*), where the predicted value $\hat{Y}$ is independent of the sensitive attribute A, *given* the *actual* value Y (Barocas et. al. 45):

$$\Pr\left[\hat{Y} = \hat{y} \mid Y = y, A = a\right] = Pr\left[\hat{Y} = \hat{y} \middle| Y = y, A = b\right] \quad \forall \hat{y} \in \{-1, 1\}, \forall y \in \{-1, 1\}$$

$$\hat{Y} \perp A \mid Y$$

The first equation above is the definition of Separation for a simple binary classification problem, and the second expression is an alternative way to phrase Separation: the predicted value $\hat{Y}$ is *independent* of A *given* the true value Y. Because we add in the random variable Y for the *true value* now, we judge a subtly different scope of possibilities – where Independence had no regard for the true value, Separation assumes we have the true value Y. And, while Independence constrained only the predicted values $\hat{Y}$ to be equivalent across any two sensitive groups, sufficiency constrains the *false positive rates* and *false negative rates* to be equivalent across any two sensitive groups (Barocas et. al. 46). That is, sufficiency has the following two constraints:

$$\Pr\left[\hat{Y} = -1 \middle| Y = 1, A = a\right] = Pr\left[\hat{Y} = -1 \middle| Y = 1, A = b\right]$$

$$\Pr\left[\hat{Y} = 1 \mid Y = -1, A = a\right] = Pr\left[\hat{Y} = 1 \middle| Y = -1, A = b\right]$$

Separation thus focuses on the classifier's results insofar as we know the *actual* results of the candidates. For example, the Separation constraint on our loan problem would measure all the candidates that *actually* defaulted on their loans for both races (say, after observing their payments on a loan for a year or two) and then ensure that the classifier's rate of mistakes on those candidates (where the classifier *predicted* they would not default) is equal across the two groups. These are the false positive rates. The same applies for the candidates that *actually* did not default on their loans. Those are the false negative rates. Thus, this penalizes the "lazy" classification seen in Independence, as the constraints are not as easily achieved as in Independence. However, the added constraint goes *against* our possible intuitions of a "positive" justice where social gaps are closed over time (Zhong). To exemplify this, suppose again that we have two groups *a* and *b* and our task is to hire applicants for a job. Suppose that group *a* has 100 applicants with 58 qualified applicants. Group *b* has 100 applicants with only 2 qualified applicants. To satisfy the Separation constraint, the company can hire 30 applicants total, with 29 applicants from group *a* but only 1 applicant from group *b.* If we subscribe to a "positive" notion of justice where disadvantaged groups should gain advantages over time, Separation allows the opposite – instead, Separation pays close attention to the gaps between groups and seeks to reproduce the gap in its classification. Thus, in cases where there is already a correlation (perhaps from preexisting historical or social conditions) of the positive outcome with a certain group more than another, Separation acknowledges that correlation and continues to mirror it. This echoes the famous argument from Rawls for *fair equality of opportunity*, where positions are formally open and meritocratically allocated (Arneson).

The third formal definition of fairness is *Sufficiency* (also known as *predictive rate parity*), where the *actual* value Y is independent of the sensitive attribute A, *given* the *predicted* value $\hat{Y}$ (Barocas et. al. 48):

$$\Pr[Y = y \mid \hat{Y} = \hat{y}, A = a] = Pr[Y = y | \hat{Y} = \hat{y}, A = b] \quad \forall \hat{y} \in \{-1, 1\}, \forall y \in \{-1, 1\}$$

$$Y \perp A \mid \hat{Y}$$

This relates to the idea of *calibration,* the notion that, between two groups, the rate of *actual* positive examples in the positively classified portion are equal (Barocas et. al. 48). In our example, if we take all the loan applicants of group *a* that our classifier predicted would not default and compare them to all the loan applicants in group *b* that our classifier predicted would not default, the *actual* percentage of applicants that did not default (say, after waiting a couple of years and checking) would be equal. Sufficiency tries to constrain the prediction from our classifier for a candidate to accurately reflect their true value. In a similar argument to Separation, sufficiency also goes against our intuitive notion of "positive" justice, possibly widening the gap between two groups a and b (Zhong).

Each of the above three definitions of fairness are definitions for "group fairness" (also called "observational fairness") in the machine learning literature (Zhong). These fall into the category of "group fairness" because the constraints imposed on the classification task work on the level of $A = a$ or $A = b$. Regardless of the individual applicants in each group, if the final rates are somehow balanced at the group level, we deem our classifier "fair." However, we saw in each example that, upon closer inspection of the individual applicants in each definition with some concrete numbers (the hiring illustrations), these classifiers could *still* make decisions that go against some notion of fairness or another. Recent literature has proposed a definition for *individual fairness* which attempts to address this (Dwork et. al.), on the basic principle that similar people should be treated similarly. This seems intuitive, but the major problem is that we must somehow *quantify* similarity in people; individual fairness requires us to somehow measure the "closeness" of two individuals in a mathematical space, which becomes an even hairier problem when we introduce the notion of the sensitive attribute A. Another alternative to the "group fairness" paradigm is *counterfactual fairness* which leverages the notion of a causal graph (Russel et. al.). Essentially, counterfactual fairness allows us to see the *effects* of eliminating sensitive attribute A and the counterfactual "world" it produces. However, counterfactual fairness is also far from being a panacea to the issue. It imposes the daunting task in practice of creating an accurate causal graph of the variables at play, which would involve the task of indicating hard-to-find and complex causal connections between human factors such as race, income, and gender, to say the least.

So, this discussion naturally leads to the question – why can we not simply try satisfying all three fairness definitions at once? If we can satisfy all three of the intuitive normative assumptions in the "group fairness" definitions by imposing each constraint, why do we not just solve our problem with imposing all three constraints? As it turns out, the literature proves an "impossibility theorem" – no two fairness criteria could hold at the once for a particular problem, unless under very trivial, unrealistic conditions (Kleinberg et. al.). It is because of this "impossibility theorem" that designing an ML algorithm ultimately ends with a normative choice. Indeed, in response to ProPublica's article on the COMPAS recidivism algorithm, those responsible for COMPAS claimed that COMPAS followed the fair measure of "predictive parity," which was in tension with ProPublica's conception of fairness (Courtland). By creating an ML algorithm, one assumes a definition of fairness (or, possibly, the lack thereof), which carries along its normative baggage. This brings the technical ML problem into the realm of philosophy – if the choice of fairness definition on a machine learning algorithm is a normative question, then how do notions such as justice, fairness, egalitarianism and discrimination in the philosophical realm address algorithmic decision-making (Binns 1)? And, more broadly, what other philosophical implications must be considered aside from this normative choice of definition?

3. Intermission: Turning Towards Methodology

In the previous section, we saw that fairness in machine learning is predominantly composed of "fairness definitions" that ultimately lead to a strong commitment to a normative choice on the definition of fairness. These choices each have their philosophical counterparts – for instance, the definition of *separation* entails *equality of opportunity* while *independence* proceeds from a commitment to *collectivist egalitarianism* (Gajane and Pechenizkiy). At this point of this paper, we reach a juncture in our philosophical evaluation of fairness in ML.

On one hand, we may go the route of grappling with political and moral philosophy, an approach excellently taken by Reuben Binns in "Fairness in Machine Learning: Lessons from Political Philosophy." In this case, the approach may be to identify analogues to the various existing fairness definitions in the philosophy, law, or social science literature and to tease out insights that may exist through such comparisons. The previous section may have very briefly introduced such ideas but furthering this approach to find a "philosophically best" definition for fairness in ML is far beyond the scope of this paper. For this, we would have to synthesize ages of debate from philosophers, legal scholars, and other literature on the nature of contested ideas such as fairness, justice, or discrimination and then, essentially, settle on a definition compatible with *all* applications of ML. Because these very notions are *still* debated, embarking on this approach might be too unwieldy for a single paper, and, some would argue, any amount of philosophizing.

On the other hand, we may critically evaluate the *methodology* of fairness in machine learning itself. Although the technical literature has been increasingly prolific in finding new mathematical formalizations, technical solutions, and algorithms to this ethical problem in machine learning, perhaps the process *itself* has been too hasty. Evaluating ML's methodology and questioning its assumptions brings us into the realm of philosophy. In this paper, I follow this second route instead, arguing that a closer look at some root assumptions that underscore fairness in machine learning, as it currently stands, have serious *methodological* issues that may prevent further progress towards truly fair machine learning. I conclude that fairness in machine learning, in its current form, is rife with *blind spots.*

In the second half of this paper, I first propose an intuitive way to think of fairness as a *context-sensitive value*, loosely related to value incommensurability in value theory (Hsieh). I argue that this allows us to adopt a view of fairness as dependent on context without committing to a staunch anti-realist or relativist stance. I also argue that it is safer to work under this assumption in the specific field of fairness in machine learning than to rely on an absolute notion of fairness. Then, with this assumption on fairness in hand, I propose three methodological "blind spots" in the ML process drawn from the philosophy of computer science and the philosophy of science: *abstraction, induction,* and *measurement*. I argue that, without addressing the issues each of these philosophical concepts pose to machine learning, the methodology towards obtaining some sort of "fairness in machine learning" is ultimately flawed. Finally, I propose some nascent solutions that may alleviate these issues by framing the methodology of fair-ML with *context-sensitivity* in mind. By reframing the methodology as such, we adopt a more holistic approach to fairness that keeps in mind necessary human and societal actors. This, in turn, might invite broader and more productive reflection in the effectiveness of certain attempts at implementing fairness in machine learning.

4. Fairness as a *Context-Sensitive Value*

Before exploring the methodological blind spots in applying fairness to machine learning, I first make an assumption on the nature of fairness as a value. This initial assumption stems from the intuition in the field of fair-ML that implementing a “fair” ML system must take into account its context and environment (Selbst et. al.). In order to make this assumption, I first briefly overview the theory of value. Then, I will introduce and motivate the term *context-sensitive value.* Finally, I will argue that, at least in the field of fair-ML, fairness is one of these *context-sensitive values.*

First, I provide background for the term *context-sensitive values* by briefly exploring the theory of value. The theory of value in moral philosophy is broadly concerned with “which things are good or bad, how good or bad are they, and what is it for a thing to be good or bad” (Hirose and Olson 1). However, the theory of value need not pertain to moral philosophy – all varieties of values can be said to fall into the domain of value theory. The objects of study in value theory simply need to be evaluative in nature (Schroeder). For instance, while moral philosophy may be concerned with the values of justice, goodness, and truth, the theory of aesthetics is concerned with the value of beauty. In most cases, a value *X* should conform to questioning how *X* something is (How beautiful is this painting? How just is this law?). To evaluate some object with respect to a value is to make a value judgment. One might make the value judgment “That painting is very beautiful,” with respect to the value of beauty, or “This law is unjust,” with respect to the value of justice.

With this conception of value in mind, it is intuitive, then, to discuss whether values can be compared or measured. In value theory, this is the notion of *commensurability,* where two values are said to be *incommensurable* if they cannot be reduced to a common measure (Hsieh).

In an example from Joseph Raz, a person faces the choice between two equally successful careers: a lawyer or a clarinetist. Here, neither appears to be better than the other, but the careers do not seem to be equally good. Raz argues that if they were of equal value, a slightly improved version of the legal career would be better than the musical one (Raz 342). But that judgment appears to be incorrect in this case (Hsieh). Thus, the incommensurable values here might be the value of serving one's fellow citizens through study of the law and the aesthetic value of musical performance. In this essay, however, I will focus on a related property – the sensitivity to different *contexts* for evaluating a certain value. This is the property of a *single* value (as opposed to two separate values, as in incommensurability) which we call *context-sensitivity.* We shall call the particular values that obtain this property *context-sensitive values.*

To motivate and exemplify this term, I draw attention to a couple values: usefulness and dangerousness. When we say something is *useful,* it seems intuitive that how useful that thing is depends on the surrounding context. For instance, suppose we compare the usefulness of a chair in a bare room by itself to the last open chair at a table during a holiday dinner. In the latter context, the chair is clearly more *useful*, and we seem to be talking about the same value in both contexts. The same can be said of a chef's knife. Suppose we compare a sheathed chef's knife in a kitchen to an unsheathed knife on the edge of a table. In the second case, all else equal, the knife seems to be more *dangerous.* One might object that this conception is missing a frame of reference to truly evaluate it. However, if we take the most intuitive frame of reference – oneself – and place oneself in each scenario, the same difference in usefulness or dangerousness obtains. For instance, if I compare being in a bare, empty room with a chair with being at a holiday dinner with my family where there is a single spare chair, the latter chair is clearly more useful from my perspective. However, there are values that might be more difficult for one to justify as *context-sensitive* without also committing to some version of relativism – justice or truth immediately come to mind. It is important to note here that deeming *some* values such as usefulness or dangerousness as *context-sensitive* does not necessitate a commitment to relativism. One can still be a realist (or a relativist, for that matter) while accepting *context-sensitivity.* Thus, I argue that some (not all) values can be said to have this intuitive property of *context-sensitivity.*

In the domain of fair-ML, I argue that fairness is also one of these *context-sensitive values* because of ML's main goal – prediction or classification of phenomena via mathematical formalization on large amounts of data. More precisely, the "contexts" we consider when implementing a fair ML algorithm include, but are not limited to, the societal context, the timeframe, the application domain, the data acted upon, the stakeholders involved, and the algorithms and implementations involved – that is, the entire pipeline of an ML task. Though the "context" here might be more difficult to parse than in our trivial chair and knife examples, an example from Section 2 might help. This example draws inspiration from a similar one from "Fairness and Abstraction in Sociotechnical Systems" by Andrew Selbst et. al. Suppose that we implement a fair-ML algorithm with the fairness definition of *Separation,* which, recall, constrains false positive or false negative rates between two groups to be equal. Then, suppose two contexts: *Separation*-constrained classification with equalized false-positive rates for hiring and *Separation*-constrained classification with equalized false-positive rates for criminal recidivism. In the former case, a false-positive (someone unfit for the job is employed) is less harmful than a false-negative (someone fit for the job is denied). An unfit but accepted applicant might just get fired in a couple months, while there is no "second chance" for a fit but unaccepted applicant. However, in the latter case, equalized false-positives might keep more of a minority population wrongly in jail, propagating an already existing injustice. Instead, we might use a different fairness definition or approach altogether for the latter case. This stems from the intuition that our single mathematical formalization fails to generalize to different social contexts (Selbst et. al. 6). Fairness might also change over the context of time. Our intuitive notions of what is fair drastically evolve over time periods or eras in a society. Thus, we see that fairness in the specific domain of fair-ML is a *context-sensitive value*, and ignoring the context surrounding the implementation of a certain approach to fairness in ML is a serious methodological error.

Finally, to be precise, I am committed that fairness *in the domain* of fair-ML is *context-sensitive.* I *do not* make assumptions of fairness as a value in general, irrespective of its domain application. There are two reason to make this assumption clear. First, it seems methodologically safer and more intuitive for our later exploration to not work under the assumption that there is a notion of fairness that is absolute in all contexts. Certainly, this is a sentiment that is increasingly reflected in the fair-ML research community as more researchers begin to realize that writing yet another mathematical definition of fairness will not be sufficient in all cases (Selbst et. al. 6). Second, accepting the existence of *context-sensitive values* does not necessitate us to accept the strong commitments of moral relativism or realism. The existence of some *context-sensitive values* does not necessitate that *all* values are *context-sensitive,* and we leave trickier values such as truth, justice, or fairness in the general domain (unconstrained by ML) on the table.

Motivated by this characterization of fairness as a *context-sensitive value*, I will provide three methodological blind spots (Sections 5, 6, 7) present in the current state of ML from the domains of philosophy of computer science and philosophy of science worth critically engaging with if we desire establishing fairness in ML.

5. Abstraction

In the practice and philosophy of computer science, abstraction is a fundamental concept underlying the field (Turner). It is basic knowledge to any computer scientist that abstraction drives the process beginning with physical hardware to electrical signals to 0's and 1's to a fully specified computer program. However, this does not begin and end with hardware and programming – mathematical abstractions and formalizations drive the practice of machine learning. For instance, an ML algorithm in the business of processing natural language might represent a word as a "word-embedding," a geometric representation of words as vectors in high-dimensional space that have a level of "closeness" to other words in a vocabulary. In this section, I argue that current computer science and ML paradigms regarding abstraction have tension with the notion that fairness is a *context-sensitive value.* In particular, the paradigms of specifications and black-box "portability" conflict with *context-sensitivity.* Thus, abstraction is the first "blind spot" methodological assumption in developing fair-ML systems.

Before the very beginning of the abstraction process, *specification* is crucial to designing any algorithm or program in computer science (Turner), but I argue that current notions of specification in applications of fair-ML are too vague and context agnostic to truly fit the bill of *context-sensitive fairness.* In order to design any algorithm or program, the computer scientist begins with a *specification,* a description describing how inputs to an algorithm should result in certain outputs. A specification describes what an algorithm does but not how it does it. For instance, the specification "take the square root of input *x*" might yield an algorithm that first checks if *x* is a nonnegative number, attempts multiplying multiple instances of an arbitrary number by itself before converging to a correct one, and then outputs that number (Turner). Nowhere in the specification are the considerations of checking for nonnegativity (a precondition for a real square root) or the process of running through possible square roots before arriving at a suitable output. Specification merely concerns input and output, and, thus, allows us to package our algorithm in a black-box and label it with its specification designating what it accomplishes.

However, in the case of fair-ML, the notion of specification should be reevaluated if a fair-ML algorithm is to meet the *context-sensitivity* of fairness. Though the history of computer science has developed from pure vernacular specifications to more rigorous ones, the field of fair-ML calls not just for mathematical rigor in specifications, but for normative values. Suppose we create a vanilla ML algorithm with some initial specification *S,* typically maximizing predictive accuracy on some dataset. Then, suppose that we also add the constraint that the ML algorithm be fair, constraint *F.* This *F* might be one of our mathematical formalizations of fairness. Regardless, the specification $S + F$ should carry the assurance that a certain program does *S and* is fair – if it is specified as so, then its inputs must produce fair outputs.

However, whereas specifications such as “take the square root of the input *x*” might be easily fulfilled with the algorithm above for all *x*, our specification $S + F$ can only be achieved in the same way *syntactically* with respect to *F.* That is, our algorithm might successfully model all the constraints of fairness that definition *F* imposes, but that output is only “fair” with respect to the “mathematical language” of *F.* It does not have a notion of *actual* normativity, as any notion of its social or political context is independent of this specification. It might fulfill equalized false positive rates brilliantly while also fulfilling *S,* but I argue that this is missing a crucial dimension. This relates to Searle’s famous Chinese Room Argument – while machines might use syntactic rules to manipulate symbols, they have no understanding of the meaning or semantics of those symbols (Searle 418). So long as *F* is properly defined, we might assume our machine fits the constraint *F* perfectly, but we cannot say it is truly “fair” until we take the context in mind. Essentially, the machine is just “symbol-crunching” the notion of fairness encoded in *F.*

I am not arguing that we need to have conscious machines or machines capable of actual *understanding* of our semantics of fairness, as one might think because of the reference to Searle. Instead, I am arguing for better specifications so *humans* in the ML process are able to understand the semantics of fairness in a given context. Until we incorporate normative assumptions based on our context, we cannot, in any meaningful sense, call an algorithm from an $S + F$ specification truly “fair.” Thus, in order to embed *context-sensitivity* to specifications of fairness, I argue that all normative assumptions $N_1, N_2, \ldots N_n$ must somehow be incorporated in a *non-technical manner* for *humans* in the process, resulting in a specification $S + F + N_1 + N_2 + \ldots + N_n$ to better incorporate our intended meaning of “fair” into the specification. When we say we want an algorithm to be “fair,” we do not mean we want it to “follow a fairness definition” – we want it to actually be *fair* in the sense that it complies to whatever normative assumptions on fairness are specific to the context of our application. For instance, an $S + F$ specification might be, “Recommend whether to hire or not hire applicants with constraint of equalized group rates between group *a* and group *b”* while an $S + F + N$ specification might add, “assuming that false positive rates are more acceptable *due to* the moral motivation to equalize long-run diversity between group *a* and *b.*” Though this kind of value-laden, normative specification might be less amenable to strict computational checks or tests, I argue that interdisciplinary checks from humans in the loop (possibly from domain experts in each context to check for fairness based on assumption *N*) would better achieve the shared goal of *actual* fairness, not just “following the math” of a specification.

The second concern with the abstraction methodology in fair-ML is the notion of portability, which also runs counter to our intuitive notion of *context-sensitivity.* In "Fairness and Abstraction in Sociotechnical Systems," Sebst et. al call this the *Portability Trap.* Abstraction demands algorithms to be *portable* – it is core to any computer scientist's skillset to be able to create modularized functions with well-defined inputs and outputs that can be imported and exported to other applications. ML is no exception to this paradigm. In ML, problems are categorized by their nature of learning task (e.g. classification, clustering, reinforcement learning, regression), which is then applied to all kinds of different real-world problems (Sebst et. al. 4). For instance, we might apply binary classification to medicine to determine whether a patient has a disease, or to images, to predict whether an image is a cat or a dog. Regardless of instance, the same algorithms for binary classification get optimized (SVM, nearest neighbors, perceptron), though the *actual* task from a human perspective is remarkably different. This portability then transfers to machine learning in practice, where actual coding tools or platforms (scikit-learn, PyTorch, tensorflow, etc.) are designed precisely *for* the means of portability (Sebst et. al. 4). Though the current state of fair-ML literature has moved beyond a purely algorithmic focus, it still embraces portability as a core value: some fix definitions of fairness as portable modules and then optimize, while others focus on building a "fair wrapper" around a classifier to make all outputs fair (Sebst et. al 4). The preoccupation with the value of portability still underscores the current ML paradigm, as well as the fair-ML research community.

Although portability and the abstraction that comes from portability are not inherently harmful, I argue that it is counter to the notion of *context-sensitive* fairness. Suppose we were to build an ML system that incorporated all notions of fairness in a certain context satisfactorily. Upon an audit from relevant ethicists and domain experts, the system is specified correctly, and the abstractions do not abstract away relevant normative assumptions. Then, upon completion, the system relies necessarily on its context, but, if so, loses the value of portability. We cannot simply “ship” this system to another context and expect that it will also be fair. That is, by ensuring fairness with respect to all of the normative assumptions and specifications in one context, it loses the ability to be fair in another. To draw on a previous example, by settling on a fair system that equalizes false positive rates for hiring by fulfilling all the normative assumptions for fairness in hiring, we cannot simply “export” this system to the domain of criminal justice, as the normative assumptions for criminal justice must then be fulfilled as well. Pure portability and *context-sensitive* fairness are therefore mutually exclusive.

Thus, I conclude that the core paradigm of abstraction is a methodological blind spot when it comes to the field of fair-ML. Specifically, the notions of specification and portability are in tension with the *context-sensitivity* of fairness. This is not an attack against the enterprise of abstraction in fair-ML, but, rather, a call to be wary of the dangers of following an abstraction paradigm in developing fair ML systems. I propose that deeper reflection on the nature of abstraction and its relevant notions of specification and portability is needed by ML practitioners in the current state of fair-ML, with more concrete suggestions in Section 8.

6. Induction

In the philosophy of science, the principle of induction – while crucial to the enterprise of science – is a fundamental problem. To quote Hume, induction is the principle that allows us to predict that "instances of which we have had no experience resemble those of which we have had experience" (Hume 1.3.6.10). For instance, suppose letting go of some ball has reliably led to that ball falling downward and hitting the ground. We predict future instances of letting go of that ball will also lead to the ball falling and hitting the ground. However, in one of philosophy's most famous arguments, David Hume posed the *problem of induction* – put simply, that inductive inferences are no more than "habits of the mind," and there is no necessary connection from observations to unseen events, despite how strong our inductive intuition might be. In this section, I first focus on Nelson Goodman's "new riddle of induction," the successor to Hume's problem of induction, to explain the learning-theoretic conception of induction that drives the theory behind machine learning applications today. Then, I show that the current state of ML ignores the problem of learning-theoretic induction in applications to *human contexts*, giving three instances: inductive patterns from historical trends in data, nonsensical trends in data, and fed-back data. Finally, I argue that faulty inductive patterns in ML data often avoid detection, hearkening to Hume's epistemological problem of induction – how do we tell good inductions from bad ones? Thus, induction poses a second methodological blind spot in ML that conflicts seriously with *context-sensitive* fairness.

First, I outline Goodman's "new riddle of induction" in order to provide insight into the learning-theoretic conception of induction relevant to ML (Goodman 74). In the "new riddle of induction," we suppose that, up to time *t,* we have observed many emeralds to be green and no emeralds to be any color other than green. We might have a series of observations of the form

"Emerald *x* at time *a* was green," where *a* < *t*. Then, at *t*, our hypothesis "all emeralds are green" is supported by inductive reasoning. Goodman then introduces the predicate "grue," which applies to all things that are observed to be green before a future time *t* but applies to all things observed to be blue at or after a future time *t.* So, with the same set of observations that allowed us the hypothesis "all emeralds are green," there are a series of equivalent observations "Emerald *x* at time *a* was grue" where *a* < *t*. Thus, we can also take these observations together to make the general hypothesis "all emeralds are grue." Further, the two statements (1) "The next emerald observed at or after time *t* will be green" and (2) "The next emerald observed at or after time *t* will be grue" both are confirmed to the same degree under inductive reasoning. However, they are mutually incompatible, for the emerald denoted in (1) will be the color green and the emerald denoted in (2) will be the color blue. Therein lies the problem. For this essay, this problem fundamentally concerns the crucial distinction that Goodman argues Hume misses in his original problem of induction – how do we distinguish good ("lawlike," in Nelson's words) from bad inductions (Cohnitz and Rossberg 5.3)? Intuitively, it seems that "all emeralds are green" is a better inductive generalization than "all emeralds are grue," but, by merely observing the data and hypotheses at hand, there is nothing to show that is the case.

The field of algorithmic learning theory (the theory *behind* ML) answers that the correct hypothesis corresponds to the *simplest* one: "all emeralds are green." One might recognize this as Occam's Razor, the principle that one should choose the simplest explanation of a phenomenon compatible with one's experiences (Harizanov et. al.16). Then, because algorithmic learning theory (and, thus, ML's applications in general) is the study of "computational strategies for converging to the truth," (Harizanov et. al. 1) the assumption of Occam's Razor *necessitates* that the most truth-conducive hypothesis is the simplest one. In fact, "no strategy that violates

Occam's Razor is optimally truth-conducive" (Harizanov et. al. 16). This embeds a crucial philosophical assumption deep into the theoretical underpinnings of ML – optimization for the truth *must also* optimize for simplicity of explanation. Then, despite the data we feed into an ML system, we can rest assured that the output hypothesis on the data is the *simplest* and *most efficient* convergence to some "truth" about the data. However, in the field of fair-ML, we do not merely want some simple "truth" about the data – depending on our context and the data at hand, we also want a "fair" hypothesis. In a wide range of applications, this learning-theoretic assumption of Occam's Razor conflicts with our notions of *context-sensitive* fairness.

The first instance where Occam's Razor might fail us is when ML systems generalize historical trends in data that may have a sociologically or normatively charged context. As explained in Section 2, the old adage of ML goes, "Garbage in, garbage out" – an ML system is only as good as the data you feed into it. The "truth" obtained from learning-theoretic induction on historical data might run counter to our notions of fairness if the historical data is biased in some significantly unfair way. For instance, consider a binary resume screening application in ML, where we simply classify resumes as "give interview" or "deny interview." Suppose our training data is a historical collection of resumes from people at our company with indication of good or bad performance over the past twenty years. However, if the past twenty years at this company was wrought with gender discrimination against females in the workplace, the training data might indicate that gender is an expressive signal of performance. Then, following our principle of learning-theoretic induction with Occam's Razor, our ML algorithm should *also* learn the hypothesis (put loosely): "females perform worse in this workplace." We might even extrapolate and assume that these historical trends in data are society-wide, not simply company-wide. This may be the case for attributes such as race or gender, and, thus, our learning algorithm would learn society-wide biases in data. These clearly conflict with our notion of *context-sensitive* fairness, where the context might be a society where we value equal treatment on the basis of gender. To compare to Goodman's new riddle of induction, we might actually want the "grue" hypothesis (a more complicated *violation* of Occam's Razor) over the "green" hypothesis. However, the methodology of ML denies us that hypothesis.

Another instance where learning-theoretic induction might fail is finding trends in the data where they simply do not exist. In a recent paper, an ML system showed that "faces contain much more information about sexual orientation than can be perceived and interpreted by the human brain" (Kosinski and Wang 1). Using a labeled dataset of images of heterosexual and homosexual male and female faces, researchers developed an ML system that correctly distinguished between homosexual and heterosexual males in 81% of cases and females in 74% of cases. Skepticism and controversy arose from this study, as it was claimed to be a new form of physiognomy (the pseudoscientific practice of distinguishing character traits from facial features or shape). In a rebuttal, researchers showed that the system actually picked up on patterns concerning glasses, makeup, eyeshadow, and image angle – contingent features of faces clearly not inherent in one's facial structure (Arcas y, Blaise Aguera et. al.). In a similar case, an ML system deemed capable of "classifying criminals with high accuracy through facial features" just picked up on a pattern that criminals in the dataset frowned more often in their portraits (Wu and Zhang 1). In both these cases, it is no longer about historical trends in data. Instead, Occam's Razor forces us to reach a potentially unfair ("criminal classifier" facial recognition software might be implemented in policing, for instance) hypothesis through nonsensical trends in data. Not every inductive pattern is a meaningful one. Further understanding the context and intricacies of the training data prevents us from folding to the simplest hypothesis.

A third instance where learning-theoretic induction betrays notions of *context-sensitive* fairness is the case of emergent bias or feedback loops. In one of the earlier works on bias in computer systems, Friedman describes emergent bias as bias in computer systems arising "…only in a context of use by real users […] as result of a change in societal knowledge, user population, or cultural values" (Friedman and Nissenbaum 335). Here, we refer to cases in which the ML system acts on the environment through "decisions, control actions, or interventions" (Dobbe, et. al. 3). In a well-known example, predictive policing might use discovered crime data (e.g. arrest records) to predict the location of new crimes and determine police deployment, causing increased surveillance of neighborhoods based on the data. But because arrest records correlate with increased surveillance, feeding the data *back* into the system might then cause even *more* surveillance and more arrests (Barocas et. al. 23). These "closed loop" ML systems can then easily validate their own learning-theoretic hypotheses by actually impacting the *real world,* enforcing the hypotheses on every subsequent iteration. In these cases, the hypothesis is generated not on a controlled set of observations, but a set that increases in fairness-violating measures *because* of the hypotheses themselves. Thus, feedback loops may force ML systems to pick up on the simplest, feedback-amplified hypothesis *H* from the data, but the more complex hypothesis, *C* (that there is a feedback loop) avoids detection until *human* thought on the context.

I conclude that the Occam's Razor assumption to learning-theoretic induction -- while conducive to the enormous progress in ML – is at tension with the methodology of fair-ML and *context-sensitive* fairness. Further, unfair inductive patterns like the ones above cannot be distinguished by a machine – the "simplest" pattern is always the best. Thus, I argue in Section 8 that human domain experts should be wary of the Occam's Razor assumption and *manually* look for hypotheses other than the simplest one.

7. Measurement

In the philosophy of science and epistemology, the nature of measurement and the various issues that arise from measuring the empirical world are rich grounds for philosophical inquiry. Though questions of measurement might evade immediate thought on how to make an ML system fair, it is obvious that data is the lifeblood of an ML system. The crucial question, then, is: how did we get this data? I suggest that the philosophy of measurement and discussions of the epistemology and methodology of measurement show that the question of *context-sensitive* fairness begins much before an ML system is even designed for the data. First, I explain the most influential framework of measurement – the representational theory of measurement – to formalize the concept. I argue that all measurements already incorporate some assumption about the empirical world that we cannot always "stasticize." Then, I show how this understanding of measurement factors into the data of an ML system, which takes these measurements and liberally plays with them as both *features* and (in most applications) *labels* of a model. I propose that measurement is a third methodological blind spot in ML that conflicts with *context-sensitive* fairness.

I first attempt to outline the representational theory of measurement (RTM), the most influential theory of measurement that formalizes measurement as a concept. In RTM, there is a distinction between *empirical relationship structures* (empirical objects to be measured) and *formal relationship structures* (the quantitative and mathematical relations for these empirical objects) (Tal 3.4). Put simply, measurement is then the construction of mappings from the *empirical relationship structures* to the *formal relationship structures.* A certain mapping is a *scale*, which specifies a certain many-to-one mapping (a homomorphism) from the empirical to the formal (Tal 3.4). For instance, I might try to measure a couple of wooden rods, *A* and *B.* To construct a rudimentary *scale*, I might say that rod *A* is *longer* than rod *B* if, when the ends on one side of each rod are aligned with each other, the other end of rod *A* extends past the other end of rod *B*. This is the empirical relationship of *longer*. Then, by assigning real numbers to the rods and assigning the symbol ">" to the relationship *longer* we can write *A* > *B* as the formal relationship in this scale. Here, our assumption in the empirical mapping is a relatively harmless procedure to distinguish the relationship of *longer* – namely, putting the rods flush to each other – but it is an empirical assumption, nonetheless. In RTM, these assumptions are necessary.

With RTM in mind, I argue that the first methodological tension is that not all scales can be liberally "statisticized" and, ultimately, some common ML operations become category errors on the measured data at hand. Because a scale is just a mapping from the empirical to the formal, different empirical relationships admit different formal ones. Types of scales include nominal, ordinal, interval, and ratio. In many cases, however, the scales themselves do not admit of common statistical techniques (Hardt and Barocas). For instance, suppose we measure a variety of restaurant reviews on an ordinal scale from 1 to 5. In an ordinal scale, the numbers have meaning in that a 2 is better than a 1, a 3 is better than a 2, and so forth, but just *how much* better is an undefined quantity. All the information we receive from an ordinal scale is the *ordering* of objects. Because of this, taking basic statistical operations such as the mean or standard deviation of an ordinally-scaled set is fundamentally a category error (Hardt and Barocas), as they work with more than just ordering. Despite this, data in all kinds of ML applications might be on an ordinal scale; restaurant reviews, movie ratings, or clinical surveys of "how much pain are you feeling" all come immediately to mind. I introduce this issue to argue that, by imposing a scale on empirical phenomena, assumptions are made that do not immediately lend themselves to machine learning compatible applications, and, by applying invalid operations on these already biased assumptions, fair-ML systems face a serious methodological issue. Here, the context to be sensitive to is at the very inception of how we get our data, but it is important context to *context-sensitive* fairness, nonetheless.

Second, I argue that methodological error in measuring data *features* is another obstacle to fair-ML because of the measurement of often elusive empirical concepts. Recall that the *features* of a dataset describe the various attributes given by each sample in the set. If we are designing an ML system to classify risk for heart disease, for instance, we might have the features: weight, age, past diseases, and whether heart disease runs in the family. As RTM claims, the *empirical relationship structures* are distinct from *formal relationship structures,* and only through certain mappings can we obtain a formal measurement of empirical things. In the case of the two wooden rods, this was simple, but ML is often concerned with measuring relatively elusive concepts. Further, when we work in the context of fairness, these concepts are typically related to *humans*. For instance, an ML system for hiring may want to measure features such as intelligence or communicability, but the only measurements for these empirical qualities might be IQ test and a scored writing sample. But who decides, and how do we know if IQ test is a good measurement of intelligence? And how do we score a writing sample without allowing some sort of reader bias? Thus, in many cases, the very data features that we use to make predictions in the first place *already* incorporate context dependent, sometimes subjective assumptions, before any sort of system is even built. This, of course, reinforces the idea of fairness as *context-sensitive* in fair-ML – without better knowledge of the measurements used to form our data, how can we ever truly assure fairness? As Moritz Hardt brilliantly put it in his NIPS 2017 talk on “Fairness in Machine Learning,” “every feature is a model.” That is, features incorporate important normative and subjective context-dependent assumptions about the measurements, but all features initially look the same to an ML system.

Third, in supervised learning (the most prevalent form of ML), the *labels* of data incorporate even a greater degree of subjectivity than the features (Barocas et. al. 16). Recall that supervised learning involves taking labelled examples of *training data*, where labels represent some "ground truth" about the samples in the training data. In a training set of images for cancerous tissue, each image has either label 0 for non-cancerous or 1 for cancerous. A panel of certified doctors or domain experts might provide these labels. In any supervised learning problem, these training data are then fed into the ML system for the system to learn and generalize to predict on new, unlabeled examples. However, oftentimes, these labels are even more nebulous than our features, as they are often constructs created for the purpose at hand (Barocas et. al. 16). Whereas features might be real, empirical properties in the world, the labels are oftentimes entirely subjective measurements. For example, one might need to choose a label for job performance, but the only existing proxies are historical performance reviews (Barocas et. al. 17). However, historical performance reviews may admit all sorts of biases such as those of past managers or of different teams in the workplace. In another well-cited example in the fair-ML literature, we might be in the business of predicting crime, but instead of measuring which training samples have actually committed crimes, we might only have access to arrest records (Barocas et. al. 17). Yet, arrests are certainly not equivalent to *actually* committing crimes, and using arrests as a proxy admits possible misinformation or misrepresentation about the samples. In both these cases, context is crucial to understanding the measurements involved, but, above all else, the training data is taken as given "ground truth" for our ML systems to learn from. There is no encoding of the context surrounding the choice of proxy, only the proxy itself.

More importantly, there is no encoding of the context of subjective *choices* that went into choosing some label-construct. Thus, measuring the labels of training data are context insensitive and violate the *context-sensitivity* of fairness.

I conclude that the measurement of empirical phenomena in order to generate training data for an ML system is, in its current form, context insensitive, disagreeing with our characterization of *context-sensitive* fairness. In the case of features, the translation from *empirical relationship structures* to *formal relationship structures* often includes a normative judgment or assumption on how to measure an empirical quality. In the case of labels, *proxies* for the actual empirical object are often measured instead of the object itself. Indeed, Cartwright and Bradburn argue that some concepts might be "too multifaceted to be measured on a single metric without loss of meaning and must be represented as a matrix of indices or by several different measures of what goals or values are at play." This characterization of measured concepts suggests a possible start of a solution – somehow encoding normative or subjective assumptions into the data upon measurement along with the *actual* measurement. Following this intuition, I argue in Section 8 that a greater focus on the sources and measurement of data *before* its acceptance as a given, "ground truth" in ML systems is crucial to fit a *context-sensitive* methodology to fairness.

8. Suggestions

At this point, we have shown three methodological obstacles to implementing fairness in ML from the philosophy of computer science and the philosophy of science: abstraction, induction, and measurement. These objections stem from the realization that fairness is a *context-sensitive* value, and, as such, avoids capture through methods such as simply specifying a "one-size-fits-all" definition of fairness (abstraction) or assuming we have well-measured data (measurement). However, this is not to say that fair-ML is a lost cause. Instead, the purpose of this paper is to provide a closer look, from a philosophical lens, at these three fundamental components of the ML process to reframe current research to embrace broader solutions. In one sense, this paper advises against "losing the forest for the trees." Because of this, I provide three suggestions in this section for ways in which ML researchers, philosophers, and domain experts alike might better approach fair-ML in a way that embraces the *context-sensitivity* of fairness and begins to assuage these deeper-rooted, philosophical issues. Though these suggestions are by no means technical solutions, I intend for them to be guidelines for researchers to avoid methodological blunders that betray the understanding that fairness is *context-sensitive.*

First, ***abstraction*** *of a fair-ML system should necessarily contain consideration of the values and context that surround it as fundamental to the iterative process.* Design in machine learning and computer science is oftentimes iterative. However, I argue that design methodology must change if we want to align with the *context-sensitivity* of fairness. Our ML systems cannot iterate simply on the technical portion of the process, driving up accuracy as a specification until we notice some societally unfavorable result. In this view, each single iteration of design focuses on the technical, and fairness is added later, usually also in a technical sense. Instead, each single iteration should include, as pointed out in "Value Sensitive Design and Information Systems," three parts: conceptual investigations, empirical investigations, and technical investigations (Friedman et. al.). Instead of checking for normative assumptions of fairness at the *end* of the pipeline, domain experts and ML practitioners must incorporate context-dependent normative assumptions into the initial specifications of the system. This more interdisciplinary approach integrates domain experts who truly understand the context of the system, ensuring that values are not lost in the process before abstraction takes place.

Second, *when fairness is involved in* ***induction****, the simplest inductive patterns are not always the best.* As seen in Section 6, learning-theoretic systems are fundamentally inductive, using the principle of Occam's Razor to efficiently converge on the "simplest" hypothesis. However, this may lead to inductive hypotheses on nonsensical or historically biased trends in data. By better understanding that the simplest induction is not always the fairest induction when we review the context of the case, ML practitioners and domain experts alike must engage in dialogue about possible *alternative hypotheses* on the data. This requires an open-mindedness to the wide universe of "not-so-simple" hypotheses and cases where (as proposed in Section 6) emeralds are actually "grue" instead of "green."

Third, *take a closer look at the* ***measurement*** *of data, the representation of those measurements, and question "ground truth."* A huge blind spot in the current ML pipeline exists even before data is fed into an ML system – the inception of that data. Based on the context, measurements might be wrongly scaled, misused, or proxies for irrelevant or unrepresentative qualities. A better understanding of measurement as a subjective mapping from the empirical to the formal might inspire deeper inquisition into the *representativeness* of certain labels or features in data. In the domain of fair-ML, this is especially important, as, oftentimes, the traits we must measure are *human* or *social* traits that elude simple schemas. Again, open dialogue is needed between domain experts, ML practitioners, and, very importantly, *social scientists* (who have better understanding of how human qualities can be quantified) to take a closer look at the representativeness or abuse of how we get our data and come up with ways to improve.

9. Conclusion

In conclusion, I argue that, to ensure that the ML community truly develops fair ML systems, we must first address three main philosophical issues: abstraction, induction, and measurement. These exist as blind spots in the current design of ML systems, and knowledge of their existence might spur future research into how to find concrete solutions to each.

To reach this conclusion, I first provided an overview of the basic process of ML and a broad survey of the current main definitions of fairness in the fair-ML literature. Then, I argued the importance of focusing on *methodological* objections to the fair-ML field, revealing three blind spots from the philosophy of computer science and philosophy of science. In order to motivate these blind spots and show that they are, indeed, problems, I presented an intuitive assumption – recognizing that fairness is a *context-sensitive* value in the specific field of fair-ML. Importantly, the question of whether this claim holds in the general case with fairness is left on the table. With this characterization in hand, I moved on to each of the three blind spots. *Abstraction*, from the philosophy of computer science, conflicts with fairness as *context-sensitive* because it requires portability and concrete specifications. *Induction*, from the philosophy of science, might fail to provide the fairest hypothesis because of the principle of Occam's Razor in learning-theoretic systems; oftentimes the simplest hypothesis is not the fairest when the context of fairness is involved. *Measurement*, also from the philosophy of science, often provides *overlooked context* that is hidden in formal representations of the data, such as assumptions that certain human traits can be easily quantified, or proxies of complex empirical phenomena are sufficiently representative.

Although the fair-ML field has grown tremendously in recent years as ML technologies become ubiquitous and increasingly consequential in our daily lives, perspectives from philosophy have been sparse. My hope is that this paper, at the very least, teases out serious philosophical issues in fair-ML to inspire further reflection from those in the field of ML and further philosophical investigation from those in the field of philosophy. While technical and concrete solutions await, the advantage of philosophy is that it allows us to revisit our assumptions, see our blind spots, and recognize when we might be “losing the forest for the trees.” I can only hope that, moving forward, philosophy is a more widely used tool in machine learning to tease out and question its most basic assumptions.

Works Cited

Angwin, Julia et. al. “Machine Bias.” *ProPublica.* www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. 23 May 2016. Accessed 1 April 2019.

Arcas y, Blaise Aguera et. al. “Do algorithms reveal sexual orientation or just expose our stereotypes?” *Medium.* medium.com/@blaisea/do-algorithms-reveal-sexual-orientation-or-just-expose-our-stereotypes-d998fafdf477. 11 January 2018. Accessed 1 April 2019.

Arneson, Richard. "Equality of Opportunity." *The Stanford Encyclopedia of Philosophy* (Summer 2015 Edition), Edward N. Zalta (ed.). plato.stanford.edu/archives/sum2015/entries/equal-opportunity. 25 March 2015. Accessed 1 April 2019.

Barocas, Solon et. al. 2018. *Fairness and Machine Learning.* fairmlbook.org

Binns, Reuben. *Fairness in Machine Learning: Lessons from Political Philosophy*. Dec. 2017. *arxiv.org*, https://arxiv.org/abs/1712.03586v2.

Cartwright, Nancy, and Norman Bradburn. “A Theory of Measurement.” *The Importance of Common Metrics for Advancing Social Science Theory and Research: Proceedings of the National Research Council Committee on Common Metrics*, Washington: National Academies Press.

Cohnitz, Daniel and Rossberg, Marcus. "Nelson Goodman.” *The Stanford Encyclopedia of Philosophy* (Summer 2019 Edition), Edward N. Zalta (ed.). plato.stanford.edu/archives/sum2019/entries/goodman

Cole, David. "Nelson Goodman.” *The Stanford Encyclopedia of Philosophy* (Summer 2019 Edition), Edward N. Zalta (ed.). plato.stanford.edu/enties/chinese-room

Courtland, Rachel. “Bias Detectives: The Researchers Striving to Make Algorithms Fair.” *Nature*, vol. 558, June 2018, p. 357. *www.nature.com*, doi:10.1038/d41586-018-05469-3.

Dobbe, Roel, et al. *A Broader View on Bias in Automated Decision-Making: Reflecting on Epistemology and Dynamics*. July 2018. *arxiv.org*, https://arxiv.org/abs/1807.00553v2.

Dwork, Cynthia, et al. “Fairness Through Awareness.” *ArXiv:1104.3913 [Cs]*, Apr. 2011. *arXiv.org*, http://arxiv.org/abs/1104.3913.

Friedman, Batya, and Helen Nissenbaum. “Bias in Computer Systems.” *ACM Transactions on Information Systems*, vol. 14, no. 3, July 1996, pp. 330–47. *Crossref*, doi:10.1145/230538.230561.

Friedman, B. , Kahn, P. H. and Borning, A. (2009). Value Sensitive Design and Information Systems. In The Handbook of Information and Computer Ethics (eds K. E. Himma and H. T. Tavani). doi:10.1002/9780470281819.ch4

Dwork, Cynthia, et al. “Fairness Through Awareness.” *ArXiv:1104.3913 [Cs]*, Apr. 2011. *arXiv.org*, http://arxiv.org/abs/1104.3913.

Goodman, Nelson. *Fact, Fiction, and Forecast*. 4th ed, Harvard University Press, 1983.

Hardt, Moritz and Barocas, Solon. “Fairness in Machine Learning.” NeurIPS 2017. Long Beach Convention Center, Long Beach. 4 December 2017. Lecture.

Harizanov, Valentina S, et al. “Introduction to the Philosophy and Mathematics of Algorithmic Learning Theory.” Springer, 2017. *Induction, Algorithmic Learning Theory, and Philosophy*.

Hsieh, Nien-hê. “Incommensurable Values.” *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2016, Metaphysics Research Lab, Stanford University, 2016. *Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/archives/spr2016/entries/value-incommensurable/.

Hume, David. *A Treatise of Human Nature: Being an Attempt to Introduce the Experimental Method of Reasoning into Moral Subjects*. Floating Press, 2009. *Open WorldCat*, http://public.eblib.com/choice/publicfullrecord.aspx?p=435863.

Hirose, Iwao, and Jonas Olson. "Introduction to Value Theory." *The Oxford Handbook of Value Theory.:* Oxford University Press, April 07, 2015. *Oxford Handbooks Online*. Date Accessed 1 Apr. 2019 <http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199959303.001.0001/oxfordhb-9780199959303-e-1>.

Kleinberg, Jon, et al. *Inherent Trade-Offs in the Fair Determination of Risk Scores*. Sept. 2016. *arxiv.org*, https://arxiv.org/abs/1609.05807v2.

Wang, Y., & Kosinski, M. (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology, 114*(2), 246-257.

http://dx.doi.org/10.1037/pspa0000098

How Social Networking Sites May Discriminate Against Women | Columbia News. https://news.columbia.edu/news/how-social-networking-sites-may-discriminate-against-women. Accessed 1 Apr. 2019.

Raz, Joseph. *The Morality of Freedom*. Reprinted, Clarendon Press, 2009.

Russell, Chris, et. al. "When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness." *Neural Information Processing Systems 2017, Volume: 30.* NIPS 2017.

Russell, Stuart J., et al. *Artificial Intelligence: A Modern Approach*. 3rd ed, Prentice Hall, 2010.

Schroeder, Mark. "Value Theory." *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2016, Metaphysics Research Lab, Stanford University, 2016. *Stanford*

Encyclopedia of Philosophy, https://plato.stanford.edu/archives/fall2016/entries/value-theory/.

Searle, John R. (1980). Minds, brains and programs. *Behavioral and Brain Sciences* 3 (3):417-57.

Selbst, Andrew D., et al. "Fairness and Abstraction in Sociotechnical Systems." *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*, ACM Press, 2019, pp. 59–68. *Crossref*, doi:10.1145/3287560.3287598.

Tal, Eran. "Measurement in Science." *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2017, Metaphysics Research Lab, Stanford University, 2017. *Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/archives/fall2017/entries/measurement-science/.

Turner, Raymond, and Nicola Angius. "The Philosophy of Computer Science." *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2019, Metaphysics Research Lab, Stanford University, 2019. *Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/archives/spr2019/entries/computer-science/.

Whittaker, Meredith, et. al. 2018. *AI Now Report 2018*.

Wu, Xiaolin and Xi Zhang. "Automated Inference on Criminality using Face Images." *CoRR* abs/1611.04135 (2016): n. pag.

Zhong, Ziyuan. "A Tutorial on Fairness in Machine Learning." *Towards Data Science*, 22 Oct. 2018, https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb.