Towards a Just Theory of Measurement
A Principled Social Measurement Assurance Program for Machine Learning

McKane Andrus
UC Berkeley
mckaneandrus@berkeley.edu

Thomas K. Gilbert
UC Berkeley
tg340@berkeley.edu

Abstract

While formal definitions of fairness in machine learning (ML) have been proposed, its place within a broader institutional model of fair decision-making remains ambiguous. In this paper we interpret ML as a tool for revealing when and how measures fail to capture purported constructs of interest, augmenting a given institutions understanding of its own interventions and priorities. Rather than codifying "fair" principles into ML models directly, the use of ML can thus be understood as a form of quality assurance for existing institutions, exposing the epistemic fault lines of their own measurement practices. Drawing from Friedler et al.s recent discussion of representational mappings and previous discussions on the ontology of measurement, we propose a *social measurement assurance program* (sMAP) in which ML encourages expert deliberation on a given decision-making procedure by examining unanticipated or previously unexamined covariates. As an example, we apply Rawlsian principles of fairness to sMAP and produce a provisional *just theory of measurement* that would guide the use of ML for achieving fairness in the case of child abuse in Allegheny County.

Introduction

Motivation

Machine learning (ML) is now widely deployed to shape life outcomes in high-risk social settings. Social scientists have criticized this deployment as needlessly automating tasks once performed by human-led bureaucracies, and unfair in its unequal treatment of vulnerable subpopulations and opaque classifications (Larson, 2016). While practitioners have responded by posing technical model fixes to deal with biased datasets, these discussions have not explored the problem of *measurement* itself as a core domain for fair ML research, neglecting important questions at the intersection of data collection, policy formulation, and what fairness even means in context.

What might an ML-informed measurement assurance program look like? And what choice of procedure is morally appropriate?

Contribution

We take a pragmatic, contextualized view of measurement, interpreting matters of fairness in terms of how well institutions can predict and reshape the social dynamics that motivate their own decision-making. ML serves as a tool for critical self-reflection by interrogating these dynamics, expanding an institution's horizons by uncovering limitations of its adopted variables of interest and suggesting the need for alternatives. While ML is morally neutral, its discoveries (e.g. how well predictions map to outcomes for different subpopulations) will clarify when assumptions about the value of any specific metric are in need of revision. This may compel a new interpretation of the lived experience of relevant communities, and in turn suggest new sampling procedures. ML thus helps ensure fairness through critical reflection on measurement, and helps reveal causal assumptions that either impede or cultivate fair outcomes. We argue ML must be deployed earlier in the decision-making process to interrogate measurement practices, and propose greater points of contact between algorithmic systems and domain experts, such as doctors and parole officers.

We present a *social measurement assurance program* (sMAP) that deploys ML to investigate how institutional policies are meant to resolve pertinent historical inequalities. sMAP rests on a representational mapping between observed human data, difficult-to-measure human qualities of causal significance, and ML-informed interventions. We then demonstrate this framework's value for resolving both conceptual and empirical puzzles for fair investigations of measured social phenomena. As an example, we apply Rawlsian principles of justice (Rawls, 2009) to sMAP to produce a *just theory of measurement*, by which an institution could falsify or bolster decision policies. This *just theory of measurement* examines suspect representations for possible bias, deploys ML to find morally counterintuitive covariates, uses sampling to probe the assumed preconditions for fair interventions, and consults with affected groups to falsify the institution's representation of them.

The Case of Allegheny County

Here we present our motivation through the example of the Child, Youth, and Family (CYF) division of the Allegheny County Department of Human Services, recently the subject of a case study by Virginia Eubanks. The Allegheny CYF

provides many services, such as protecting children from abuse and neglect, and has implemented the Allegheny Family Screening Tool (AFST) to predict risk using data for 287 variables scraped from CYFs database.

The AFST illustrates the difficulties faced by organizations that must implement fair policies in the face of meager data and funding shortages, considerable social distance between policymakers and subpopulations, and incomplete understandings of the causes of human suffering. In this context, the hypothetical social realities of "abuse" and "neglect" are difficult to confirm outside of fatalities or near-fatalities. As such, the AFST is limited to predicting community re-referral and child placement, two outcome measures that provide a much larger training and validation set than systematic child abuse. Ironically, the very families most in need of intervention may be invisible to CYF's policy interventions, compounding domestic abuse with institutional neglect.

More critically, AFST is primarily used to supplement case worker judgment, not determine which cases are worthy of closer investigation. This favors individual (and possibly parochial) judgments based on variable expertise over a more systematic perspective on county-wide child protection services. In extreme cases, it is possible for AFST to demand caseworker inspection, but AFST does not determine the style of intervention CYF might employ–program managers are free to ignore its predictions entirely or re-interpret them at will. Moreover, if the caseworker believes there is insufficient evidence of abuse to remove the child, they may refer benefits to the family instead. This ad hoc decision-making may be influenced by political commitments (to local schools, churches, family ties), cultural prejudice, or the limits of CYFs monthly finances.

In effect, while AFST relies on manifold statistical measures rather than caseworkers expert judgment, both model and humans are trying to decide whether or not to delve deeper into a case without the criteria for decision-making being clearly specified or shared between parties. At a minimum, AFST could be used to account for faulty intervention decisions (if not improve them by revealing new human contexts), serving as an instructive elaboration of design dilemmas and procedures for XAI.

Fair Machine Learning in Institutional Contexts

Eubanks case study raises a wider question for machine learning practitioners–what is the place of traditional expert judgment in the context of changing institutional priorities and the uncertain effects of well-intended interventions? The problem of "fair" machine learning cannot be narrowly defined in terms of formal models of fairness or technical refinements to existing approaches, but must be expanded to confront and remake the conditions under which institutional interventions are able to be seen as fair.

Illustrating these stakes, Madrigal (2019) considers the use of ML to identify Flints lead pipes to help diagnose the ongoing water crisis. Local officials determined this approach would be more efficient and systematic than guess-and-check pipe inspections, but this strategy was abandoned as politically toxic once community groups complained it left neighborhoods at the mercy of a poorly-understood (and widely distrusted) administrative tool that classifies certain pipes as not worth checking. Thus, MLs supposed fairness is only as good as the authorities through which it operates–a major problem within contexts of widespread institutional failure.

On a technical level, this skepticism towards ML can also flow from dataset bias. If ML's utility is the ability to discern implicit structure in reams of data, this structure does not necessarily scale with the accurate representation of diverse subpopulations. The data may lack adequate record-keeping, inaccurately portray disadvantaged social groups, or simplify social contexts in a way that omits key causal relations. Lipton and Steinhardt (2018) has diagnosed this tension in much ML scholarship, including a failure to distinguish explanation and speculation, failing to identify sources of empirical gains, mathiness, and misuse of language.

Each pitfall reflects a paradox governing the ML research agenda, as well as MLs ambiguity for the institutions whose data it processes and whose interventions it justifies. Compensating for inadequate data (through guesswork, model overtuning, technical overcompensation, or imprecise terminology) is motivated by intuitions about unobserved phenomena behind the data, which the system may capture through further optimization. Yet ML has also revealed new contexts endogenously from diverse data sources, whose covariates suggest poorly understood political and social realities. This tension begs the question of whether ML itself is in dire need of explanatory models so that authorities can deploy it with confidence, or if its operational efficiency and technical innovations are sufficient to challenge the expertise of authorities themselves–in effect, to automate how the mechanisms of social reproduction are discovered and explained.

Critics have rejected the latter in favor of a more traditional human-in-the-loop approach to using ML, and questioned MLs value for predictive risk assessment. Barabas et al. (2018) show how risk assessment itself has historically swayed between behavior predictions and justifying draconian sentencing policies. Consequently, they suggest that ML "should be used to surface covariates that are fed into a *causal model* for understanding the social, structural and psychological drivers of crime," rather than help shape penal policies directly. Likewise, cor are critical of naive operational "solutions" to problems of fairness, and argue for aligning model representation with traditional principles of due process. They dismiss formal fairness criteria, as "it is often preferable to treat similarly risky people similarly, based on the most statistically accurate estimates of risk that one can produce." However, neither presents explicit ontological assumptions that would justify this skepticism of ML and delineate how dataset bias, optimization, prediction, and interventions might be related procedurally. In other words, it is not clear how ML could be used to augment due process itself, its spotty history notwithstanding.

Indeed, it is unclear if the findings or assumptions of ML models even require explanation if they can, in theory,

guarantee robust predictions. Lipton and Steinhardt (2018) laments that ML papers often purport to explain model results by proposing highly intuitive theories that, while lacking "crisp formal representations," are still meant to rhetorically justify exploration. This begs the question of whether we are letting the administrative tail wag the algorithmic dog: perhaps we are better off trusting model robustness to augment the assumptions that motivated data collection and intervention in the first place. In fact, Doshi-Velez and Kim (2017) argue that we need interpretability only when there is *incompleteness* in the problem formalization, as this creates a barrier to optimization and evaluation. It would therefore be an institutions job to address this incompleteness while weighing context-specific concerns about safety and ethics, not demand explainable models out of hand.

We must therefore ask a more radical question: could ML help formalize possible interventions? Mullainathan and Spiess (2017) examine this question from an econometrics standpoint, arguing that ML should be used to probe social settings with strong verifiable assumptions (at which it excels) but relatively poor understanding of how or why social reproduction of inequality occurs. This interprets data exploration as a designed intervention on model assumptions rather than a sanitized approach to risk assessment. Dawes, Faust, and Meehl (1989) suggests this approach in a clinical context: "What is needed is the development of actuarial methods and a measurement assurance program that maintains control over both [clinical and actuarial] judgment strategies so that their operating characteristics in the field are known and an informed choice of procedure is possible."

Social Measurement Assurance Program (sMAP)

In this section we outline a *measurement ontology* for organizational interventions on the social world which could ensure an automated decision system fulfills its intended purpose. We propose a social Measurement Assurance Program (sMAP) whose measurement procedures are defined by representational mappings between *observed* human data, *difficult-to-measure* human qualities of *causal significance*, and *institutional interventions*. We will first review distinct theories of measurement in order to define these components and justify this mapping between them. We then present a general model of institutional decision-making, highlighting where measurement assurance is most relevant. We conclude this section with a description of what an sMAP might look like in practice.

Theories of Measurement

We draw from the general definition of measurement from Hand (1996) as an interpreted relation between what is real and what is observed, made possible by sampling interventions. Two such interpretations are relevant for social phenomena. First, representational measurement aligns some definition of what is real with a given empirical process. For example, an empirical process used to measure dire poverty can relate different measurable variables such as total accessible funds, average daily caloric intake, and risk of debilitating illness, but these variables serve only as proxies for some underlying, unobserved reality (the state of being poor). Second, observational measurement relies solely on an empirical system, with no baseline definition of what is real or proposed underlying constructs. Under this theory, total accessible funds, average daily caloric intake, and presence of fever are what is real – not funding sources, food supplies, disease, or poverty as such.

Model of Institutional Decision-Making

Making ML "fair" requires mapping our observed reality onto some space of possible actions that might push us towards a more equitable world state. As described in Friedler, Scheidegger, and Venkatasubramanian (2016), when institutions conduct a mapping from observed data to decisions, there is an implicit mapping onto some *construct* space that defines the context for the decision. For example, in college admissions, GPA scores hint at *general intelligence*, a difficult-to-define construct that the admissions officers are actually using as a basis for decisions.

We further note that the decision space also affects the *observation space*. As a more detailed example, consider a welfare program that processes short-term housing for the homeless. Those that have been given short-term housing in the past will start to appear different in the observation space, i.e. interactions with shelters and law-enforcement will decrease, as a direct result of past interventions. These outliers cannot be summarily removed from the system, as they may again "become" homeless in the near future. The program coordinators must thus consider how the decision space might map onto a new observation space, and even reconsider the underlying construct space, which comprises the context of underlying, policy-relevant inequalities.

We claim this mapping from observed reality onto constructs is best defined through a representational theory of measurement. While Bartholomew (1996) has suggested that social measurements are operational because they must arbitrarily define a means to gather data rather than provide an exhaustive map between observed and represented variables, we note that most institutions do not claim to perfectly model the social world. Rather, they strive to effectively intervene on relevant communities, based on construct spaces that reflect context priorities. For example, how admissions represent college applicants might differ based on school quality. Whereas a 3 on an AP exam signifies barely passing for a student from a prestigious private school, it might signify great self-motivation from an underserved rural or inner-city high school, empowering admissions committees to encode *drive* as more a desirable trait than *elite pedigree*. In this way, the institution relies on a principled, composite measure that stems from a contextual understanding of the underlying relations between attributes. While this approach is by no means revolutionary, by adopting tools of automated decision-making these composite measures with groundings in social theory become easier to ignore in favor of the operational measures already at-hand. As a means of confronting this loss of context, we propose the use of a social Measurement Assurance Program.

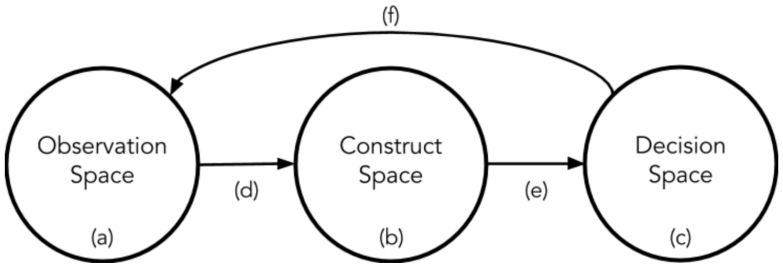

Figure 1: Representational Mappings for a Social Measurement Assurance Program

Outlining sMAP

Measurement assurance programs (MAP) are common in engineering disciplines. As defined by Speitel (1982), a MAP is "a program to establish, evaluate, and control the quality of measurement." Accurate measurements are requisite for complex systems that rely on both sensors and actuators, since slight deviations can rapidly lead to system-wide failures. MAPs are thus managerial tools that ensure the measurements being taken are the *correct ones to use*, that measurement systems *operate properly*, and that the broader system is *robust to the precision of measurement*.

We apply this intuition to measures used for institutional decision-making. Social measures, unlike physical properties, are often not consistent or generalizable (National Research Council, 2011; Bartholomew, 1996). Campbell's law provides the most instructive justification: "The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor." (Campbell, 1979) As the objects being measured are in a dynamic relation with the decisions being made, any attempt to directly implement an engineering MAP will be unsuccessful. Thus, a *social* measurement assurance program (sMAP) will require a new set of tools.

Like a physical MAP, an sMAP should ensure that: (1) gathered measures map onto the institution's construct space; (2) the methods of measurement are consistent and correctly implemented; and (3) decisions based on selected measures have the intended effect on the observation space. These components respectively reflect the standard statistical concepts of construct, content, and criterion validity. However, our treatment also entails interrogating and reinforcing the measurement procedure with respect to context-specific definitions of fairness, beyond representativeness and correctness. Given an sMAP's underlying ontology, it further requires (4) a feedback mechanism to the constructs, such that they might be adjusted to better align with intervention aims. Applying this to Fig. 1, (1) is an intervention on (d), (2) an intervention on (a), (3) an intervention on (f), and (4) an intervention on (b).

Measure to Construct Correspondence We propose that ML can validate the mapping of measured attributes to desired constructs. Even simple regression methods surface relations between measures and decisions that encourage critical reviews of how a given measure reflects the underlying construct. For instance, a college admissions office might discover that key covariates in admission decisions are AP scores with little dependence on school rankings. By incorporating ML into their sMAP, the college might investigate this discrepancy and decide to weight AP scores by high school rankings, effectively merging these two measures and remaking the construct for *academic effort*. While this new measure may not be useful for existing data, the college can now provide admissions officers with this measure directly, refining the admissions pipeline.

Measurement Consistency and Validity Measurements could also be rendered meaningless through inconsistencies in sampling or the loss of key contextual details. Extending the welfare program example, different caseworkers' ability to garner trust with their charges might produce measurement inconsistencies. The validity of the total annual income measure could also be compromised by omitting illicit sources of income, such as drug sales. Using ML to model the decision-making process could address the observation space's limitations. For example, indirect measurements that figure strongly in the construct space but not the decision space (e.g. total annual income as a proxy for housing stability) indicate that the institution's interventions may not be contextually valid. Furthermore, if the measure carries predictive weight only for specific groups within the sampled population, it is likely that the measure needs to be adjusted to better reflect group idiosyncrasies. Income from unconventional or illicit professions like prostitution is likely to play a differential role for distinct subgroups, such that its omission from the measurement procedure without considering treatment effects would be inappropriate.

Decision Measurement Feedback Adjustment Once sMAP serves a dynamic system that updates a given decision-making process, we can produce a history of predictive ML models. Comparing these models can determine the impact of a decision-making procedure on the observation space, and how a new procedure may reshape future observations. For example, if certain measures lose predictive value, the measures may have fallen prey to Campbell's law, or the affected populations may have exogenously changed. If instead predictive outcomes are consistent regardless of interventions, it implies the construct space should be ques-

tioned, altered, or abandoned. Either way, an important feature of an sMAP is to keep the measures in constant correspondence with the institution's construct commitments, a procedure that must be iteratively carried out and maintained.

The measurement space can also be assessed by maintaining close contact with subjects. This will help gauge real impacts, revealing how certain measures might be adjusted or how the measurement ontology might need realignment. ML practitioners should solicit input from ethnographers, survey methodologists, social policy planners, and other qualitative experts to better capture contexts that remain invisible to their models. In our previous example of college success based on high school performance, ML practitioners might partner with guidance counselors at both public and private high schools to better identify the contexts within which covariates germinate and how their own model assumptions (e.g. the weighting of SAT vs. AP scores) create distinct incentives for socially unequal college-bound students.

Construct Verification There must be explicit record-keeping of how constructs relate to interventions, and what observed outcomes would falsify them. If the measures that represent the construct do not carry much weight, the construct may be (a) not as important as the institution believes it to be, or (b) not as central to decision-making as it is meant to be. As an example, consider that after consulting with guidance counselors, admissions experts learn that *drive to succeed* germinates differently among first vs. second generation college-bound students, where for the former it tends to manifest in extracurricular activities and for the latter in higher final GPA. This both challenges the underlying reality of the construct and implies that existing policies do not capture its complexities, which may suggest alternative constructs that are better understood and easier to measure.

Conclusion We have suggested that ML can not merely surface but also *contour covariates* for decision-making. Not only are important measures revealed, unimportant constructs may be dismissed as contextually inappropriate or non-existent. Depending on the method used, connections may also be drawn between measures, hinting at deeper relations between constructs and potentially altering existing legal or scientific understandings of relevant social variables. Entirely new measures could be codified if ML helps identify and codify covariates for which we have no prior construct-based intuition. ML thus constitutes a *technical* intervention in how *policy* interventions are made, shifting the grounds by which decisions are justified by altering the conditions under which they can be envisioned, enacted, and evaluated. In these ways, use of ML can hint at where a given institution might dedicate future resources to improve understanding of the context of its own decisions.

Applying Rawlsian Principles to sMAP

Here we develop a *just theory of measurement* by applying a specific fairness ontology to the representational mappings within sMAP. As an example, we deploy Rawlsian principles of justice due to their specific relevance for the Allegheny County case as well as their more general influence on scholarly debates surrounding fairness. While sMAP requires such an ontology in order for it to justify fair interventions, other philosophical theories of fairness could be used instead of Rawls. We hope this section serves as an early example to be refined by the wider community of ML researchers and social activists interested in combining technical models with notions of procedural justice.

Our general approach is influenced by (Binns, 2017), who discusses various moral and political-philosophical approaches to ML fairness, with two key elaborations. First, because we interpret measurement *representationally* rather than *operationally*, ML can be used as a tool to test existing representations of the social world for unacceptable forms of bias, rather than merely surface covariates for existing causal models. Second, we align our Rawlsian sMAP with what Binns (2017) calls *deontic justice*: "the sense in which egalitarianism can be...not concerned with an unequal state of affairs per se, but rather with the way in which that state of affairs was produced". Deontic justice defines how the world would need to be observed in order for abstract moral principles to hold, rather than how we should model features in order to uphold specific fairness classifications, such as equal parity. In other words, ML fairness is not simply a matter of ensuring that measures capture the construct of interest, but that such constructs need to be a reasonable and reliable basis upon which an institution can pursue its goals.

Lippert-Rasmussen (2014) support this intuition: "Statistical facts are often facts about how we choose to act. Since we can morally evaluate how we choose to act, we cannot simply take statistical facts for granted when justifying policies: we need to ask the prior question of whether it can be justified that we make these statistical facts obtain." For deontic justice, ML should be oriented not just for evaluating the causal effects of future interventions, but also the causal mechanisms behind historical inequalities that are visible to the model. This might be reflected early in feature selection or later in the choice of model fairness criteria, either of which can produce "moral tensions" through dilemmas of incompatible assumptions, which the example of COMPAS strikingly demonstrated Larson (2016). Direct engagement and application of moral principles from ethics and political philosophy is necessary if fair ML research is to be placed on systematic, meaningful, and internally consistent foundations.

Rawls' framework is perhaps the most influential and systematic theory of deontic justice currently available. Its foundation is that humans are rational beings capable of articulating abstract moral maxims that hold universally. Rawls acknowledges that circumstances hinder this capability, and he presents a thought experiment to account for this. In the Original Position, the members of a society are made unaware of their material wealth, political power, and mental/physical aptitude, and deliberate on operating principles for the society they are about to enter. From this removed position, individuals' rational faculties will permit a convergence on two universal principles of justice. Rawls describes these as follows:

"First Principle: Each person is to have an equal right to the most extensive total system of equal basic liberties com-

patible with a similar system of liberty for all. Second Principle: Social and economic inequalities are to be arranged so that they are both: (a) to the greatest benefit of the least advantaged, consistent with the just savings principle, and (b) attached to offices and positions open to all under conditions of fair equality of opportunity." (Rawls, 2009)

The first principle implies that others' liberties cannot be violated. Part (a) of the second principle requires that any material inequality is to directly benefit the least well-off, such that an increase in wealth for the most well-off must be proportionally matched for the least well off. Part (b) of the second principle establishes that there must be an equality of opportunity in acquiring goods or in obtaining public or corporate office. Rawls argues that these principles should be implemented in society in stages, where at each stage the principles are applied at a finer grain to political and social life. Even if a resulting equilibrium between social reality and fair principles is feasible, the specific staging from the real world to decisions procured from these principles has been a source of controversy (Nozick, 1974; Roemer, 2009).

To apply Rawls to sMAP, we interpret these two principles of justice as constitutive elements of the *construct space* that must correspond to an institution's observed social measures as well as its acceptable decision space. Any proposed modifications to the sMAP would be *just* so long as they aim to better align its measures, constructs, or decisions *with Rawlsian principles*. In this manner, sMAP can give form to specific fairness ontologies, allowing specific moral intuitions to be put to work by informing the range of observations analyzed and decision interventions proposed. In the next section we apply this intuition to a recent case study.

An sMAP for Allegheny County

Returning to our original motivating case, here we consider how a Rawlsian sMAP might be applied by the Allegheny County Department of Human Services to resolve its intervention dilemmas. The Allegheny CYFs services and goals appear to fall under the Rawlsian principles that define a just institution, making this is a natural setting to apply the just theory of measurement previously described.

Recall that "abuse" and "neglect" serve as constructs that are difficult to confirm outside of fatalities or near-fatalities. From here, we consider each requirement of an sMAP and outline a more Rawlsian instantiation of CYF to satisfy those requirements.

Measure to Construct Correspondence

Within the CYF's intervention model, there are two measurement-construct mappings. The first maps from measures (e.g. community referrals, past interaction with welfare institutions) onto *parenting quality*, used to gauge risk. The second are the proxies of community re-referral and child placement onto *demonstrated risk*. In the first case, when AFST was being built, 156 of the 287 scraped variables were chosen based on their correlation with risk predictions. This unprincipled variable selection ignores the construct space by adopting an operational theory of measurement and prediction. However, caseworkers still implement the representational mapping in their own risk assessments.

sMAP suggests several strategies for aligning caseworkers' representational mappings with the AFST as a validation tool. For example, caseworkers might find that they are called to certain communities more than others, and consulting AFST reveals that it is using zipcode in its predictions. The caseworkers infer that the predictive power of zipcodes is only acceptable if it maps onto the constructs of *abuse* or *neglect*, as it should not have an impact on more individualized constructs like *parenting quality*. Unable to disentangle these mappings in the AFST model, and armed with Rawlsian principles, the caseworkers should advocate for the removal of the zipcode feature from the model and instead solicit geographic measures more meaningfully or historically connected to *abuse* or *neglect* (e.g. concentration of cultural groups that condone abuse) in order to better establish the contours of these specific constructs. This is because these contours may unfairly conflate relevant constructs in a manner that violates Rawls' liberty principle.

On the other hand, we have the mapping of re-referrals and child placement onto the construct of *demonstrated risk*. Eubanks discusses how re-referrals are especially fraught because of reporters' racial biases against possible abusers. A Rawlsian sMAP might suggest surfacing other possible covariates to see if these biases are rooted in more specific social categories besides race (e.g. rival church affiliations, clan marriages) so that re-referrals are not held up to an invisible, empirically-unverified construct, which would also violate the liberty principle. Furthermore, as the caseworkers themselves do not share these biases, their insight into contextual differences between prejudiced and earnest referrals could also prove useful in finding alternative measures that more accurately map onto the construct of *demonstrated risk* beyond racial classification.

Measurement Consistency and Validity

Eubanks further notes that the poor are disproportionately subjected to automated institutional processing. In the case of Allegheny County's CYF, this asymmetrical treatment is reflected in how the ASFT produces far more mandatory inspections of poor families than of others. Eubanks shows that this difference can be largely attributed to the measures used to predict risk. The most important measure, referrals, is a product of many social factors, including majority perceptions of what "good" parenting looks like. Eubanks describes how employees from welfare institutions are often mandatory reporters, i.e. they are obligated by law to report children that show signs of *abuse* or *neglect*. As previously discussed, neglect is easy to conflate with the effects of poverty. Furthermore, the relative dearth of processed middle and upper class families speaks to measurement inconsistencies and possibly invalidity. As referrals largely come from professionals employed at institutions that interact mainly with the poor and working classes, the measurement *procedure* is inconsistent, as the middle and upper classes are not subject to the same modes of scrutiny. To remedy possible measurement mistakes and render measurements more consistent with Rawls' Difference Principle

(holding that inequalities should work to the advantage of the worst-off), a just sMAP would require selectively auditing referrals instead of processing them all uniformly.

On top of this, Eubanks (2018) hints at professional bias against poor and working class "natural growth" parenting styles. This bias might render the measure invalid, as it unduly esteems middle class "concerted cultivation" (Lareau, 2011). According to Eubanks, CYF employees are aware of the inconsistencies in these measurement procedures, but they believe that they have no means of addressing them. Wealthy families resist the forms of surveillance and interference that poorer families must accept. Employees from supportive institutions (therapy practices, Alcoholics Anonymous, other rehabilitation centers) are often not mandatory reporters, and changing this would be a political challenge. In the face of these difficulties, it is unclear what steps the CYF might take to adjust measurement strategies in a way that better instantiates the Difference Principle. As a result, the sMAP would be best supported by extended participant observation to uncover new ways of assessing risk. If the measures must remain largely inconsistent between populations, separate models and strategies will need to be employed.

Decision Feedback Adjustment

Families that receive support from CYF are often dramatically changed. Eubanks describes how CYF support is conditional upon families subjecting themselves to greater scrutiny and enrolling in time-consuming programs on work readiness and parenting skills. In this way, time spent on parenting (one of CYF's observational measures) will not only decrease, but CYF is also given more access to the home to observe this decline. If implementing an sMAP, CYF should incorporate the impacts of their interventions into their decision-making procedure, such as tailoring a method or model for predicting risk of previously processed families.

Construct Verification

As already noted, poor families often do not have all necessary child care resources. Might this interact with the construct of neglect? And how might ML, combined with a more just theory of measurement, better elucidate this construct for CYF? Firstly, Eubanks (2018) points out that poorer families routinely receive higher risk scores from the AFST. If the model permits isolating or omitting certain features, the CYF could very well find that the economic status of a family has an impact on risk score beyond just the downstream effects of income on other variables, such as school attendance and living situation. If this is the case, then it is likely that the economic difficulties a family faces are indirectly mapping onto the construct of neglect in the decision-making procedure, beyond what is observed. Thus, an sMAP would suggest that where the construct of neglect is meant to be used in determining CYF interventions, a construct distinction needs to be made between *intentional neglect* and *means-induced neglect*. In the case of *means-induced neglect*, the role of CYF should be to ameliorate the situation with increased family support. This might require a distinct verification system, as community re-referrals and child placement are not likely to be accurate indicators of means-induced neglect.

Conclusions

ML's potential for reinventing institutional measurement is already palpable. For one, it *challenges the timescales* for common social variable measures and construct assumptions. Recidivism is infamously difficult-to-measure, and can be approximated only through an arbitrary timescale (two years in the case of COMPAS (Larson, 2016)) that may vary widely for distinct genres of crime as well as different types of criminals (petty crooks, gangs, serial killers, etc.). However, we can deploy ML to help reveal the arbitrariness of such measures and suggest, through alternate modeling assumptions, a wider range of timescales against which such constructs retain semantic meaning and predictive currency.

Second, ML can also *expand the spaces* of relevant observations, possible interventions, and imaginable constructs whose correspondences determine the efficacy of interventions. Greater tolerance of diverse data sources, a willingness to rethink pre-ML policies through what these sources reveal, and deliberation over narrative constructs can only be a good thing for social programs whose existing deployment strategies are repeatedly threatened by budget cuts and lack of political support. ML ultimately supplies a broader means of measuring the inequalities that define us, and will help lay the groundwork for rethinking principles of justice and social democracy in the coming decades.

References

Barabas, C.; Virza, M.; Dinakar, K.; Ito, J.; and Zittrain, J. 2018. Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. In *Conference on Fairness, Accountability and Transparency*, 62–76.

Bartholomew, D. 1996. Response to statistics and the theory of measurement. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 473–474.

Binns, R. 2017. Fairness in machine learning: Lessons from political philosophy. *arXiv preprint arXiv:1712.03586*.

Campbell, D. T. 1979. Assessing the impact of planned social change. *Evaluation and program planning* 2(1):67–90.

Dawes, R. M.; Faust, D.; and Meehl, P. E. 1989. Clinical versus actuarial judgment. *Science* 243(4899):1668–1674.

Doshi-Velez, F., and Kim, B. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Eubanks, V. 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

Friedler, S. A.; Scheidegger, C.; and Venkatasubramanian, S. 2016. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236*.

Hand, D. J. 1996. Statistics and the theory of measurement. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 445–471.

Lareau, A. 2011. *Unequal childhoods: Class, race, and family life*. Univ of California Press.

Larson, J. 2016. How we analyzed the compas recidivism algorithm. *Propublica*.

Lippert-Rasmussen, K. 2014. *Born free and equal?: A philosophical inquiry into the nature of discrimination*. Oxford University Press.

Lipton, Z. C., and Steinhardt, J. 2018. Troubling trends in machine learning scholarship. *arXiv preprint arXiv:1807.03341*.

Madrigal, A. C. 2019. How a feel-good ai story went wrong in flint.

Mullainathan, S., and Spiess, J. 2017. Machine learning: an applied econometric approach. *Journal of Economic Perspectives* 31(2):87–106.

National Research Council. 2011. *The importance of common metrics for advancing social science theory and research: A workshop summary*. National Academies Press.

Nozick, R. 1974. State, anarchy, and utopia. *Malden, Mass: Basic Books*.

Rawls, J. 2009. *A theory of justice*. Harvard university press.

Roemer, J. E. 2009. *Equality of opportunity*. Harvard University Press.

Speitel, K. 1982. Measurement assurance. In Salvendy, G., ed., *The Oxford Handbook of Innovation*. San Francisco: John Wiley and Sons.