Towards a Just Theory of Measurement

A Principled Social Measurement Assurance Program for Machine Learning

McKane Andrus*
Thomas K. Gilbert*
mckaneandrus@berkeley.edu
tg340@berkeley.edu
UC Berkeley

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 institution's 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 recent discussion of representational mappings in the Fair Machine Learning literature 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.

KEYWORDS

measurement; machine learning; justice; fairness; representational measurement; measurement assurance; Rawls; institutional decision-making

ACM Reference Format:
McKane Andrus and Thomas K. Gilbert. 2019. Towards a Just Theory of Measurement: A Principled Social Measurement Assurance Program for Machine Learning. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES '19), January 27–28, 2019, Honolulu, HI, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3306618.3314275

INTRODUCTION

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 human-led tasks, and unfair in its unequal treatment of vulnerable subpopulations and opaque classifications [15]. 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.

We take a deeply pragmatic view of measurement, interpreting matters of fairness in terms of how well institutions can predict and reshape the context-specific 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 [20] 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

[*] Both authors contributed equally to this research.

AIES '19, January 27–28, 2019, Honolulu, HI, USA

ACM ISBN 978-1-4503-6324-2/19/01…$15.00
https://doi.org/10.1145/3306618.3314275

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 CYF's database.

The case of AFST illustrates the difficulties organizations face when they are pushed to address funding shortages with automation but are still devoted to implementing fair policies. In the context of child welfare, the underlying 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. Furthermore, 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 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.

Both qualitative and quantitative criteria for this case study have recently been proposed by the developers of AFST and the wider academic community. Brown et al. [4] suggest communication strategies between AFST and impacted families that might increase public trust and establish system accountability, while Chouldechova et al. [6] consider how the use of analytics can both create problems through excessive reliance on administrative data and also refine human judgment through its use. Currently missing from this work is a systematic exposition of measurement as the crucible for how interventions are morally justified and reinterpreted in context, as well as the epistemic stakes behind possible interventions.

FAIR ML IN INSTITUTIONAL CONTEXTS

Eubanks' case study raises a wider question for data scientists – what is the proper relationship between institutionally-ingrained measurement practices and the more recent epistemic affordances of machine learning, such as model selection, training, and optimization? The problem of "fair" machine learning sits right at this intersection, pursued by distinct research efforts such as formal "explainable" models or technical refinements to measurement procedures without analytically addressing the relationship itself.

Skepticism towards ML is also found within the technical literature, such as ongoing struggles against 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 [17] 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 tension governing the ML research agenda, as well as ML's ambiguity for the institutions whose data it processes and whose interventions it justifies. In any given empirical context, compensating for inadequate data (through guesswork, model overtuning, technical overcompensation, or imprecise terminology) is motivated by expert intuitions about unobserved phenomena behind the data, which the system may capture through further optimization. Yet ML can also reveal new contexts endogenously from diverse data sources, whose covariates suggest poorly understood political and social realities. This tension has motivated the search for explanatory models so that authorities can commit either to deploying it with confidence in service of their own assumptions, or deferring to it if its operational efficiency and technical innovations are sufficient to challenge their own assumptions – in effect, to determine whether the actual mechanisms of social reproduction are better established by domain experts or technical models.

Indeed, it is unclear if the findings or assumptions of ML models even require explanation if they could, in theory, guarantee robust predictions. Lipton and Steinhardt [17] lament 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 [9] 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 institution's job to address this incompleteness while weighing context-specific concerns about safety and ethics, not demand explainable models out of hand.

However, many critics reject this vision in favor of a more traditional human-in-the-loop approach, and questioned ML's value for predictive risk assessment. Barabas et al. [1] 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, Corbett-Davies and Goel [7] are critical of naive operational "solutions" to problems of fairness, and argue for aligning model representation with traditional principles of due process. However, neither of these positions presents explicit ontological assumptions that would firmly ground 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 a decision pipeline that supports due process might be augmented through the application of ML to measurement procedures, even if its history has been spotty.

We must ask a more radical question: could ML help reshape decision-making by formalizing the conditions under which interventions are perceived as fair? Mullainathan and Spiess [18] consider this possibility 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 et al. [8] 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

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 sort of mapping between them. We then present a general model of institutional decision-making, highlighting where measurement assurance is most relevant.

Theories of Measurement

We draw from the general definition of measurement from Hand [12] 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. We introduce these theories of measurement because any institution that makes it their task to intervene on the real world must grapple with the question of how to represent and understand it. In the following sections we argue that in order for an institution to enact meaningful, systematic change, it must first espouse a representational system of measurement.

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 et al. [11], 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. Thus, there is a subsequent mapping of these constructs onto a space of possible decisions, or interventions, that the institution can make.

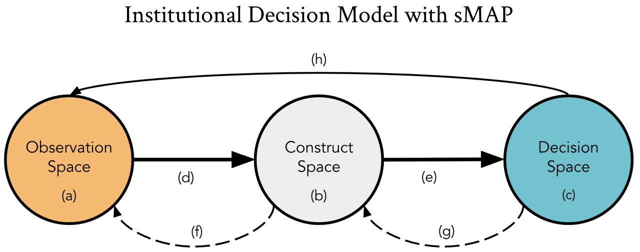

Figure 1: Enforcing measurement consistency and validity is a direct intervention on (a). Verifying the measure to construct correspondence is an intervention on (d) by way of (f). Decision-to-measurement feedback adjustment is an intervention on (a) by acknowledging (h). Finally, confirming the constructs is an intervention on (b) by way of (g).

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.

This mapping from observed reality onto constructs is best defined by a representational theory of measurement. Bartholomew [2] has suggested that social measurements are necessarily operational because they arbitrarily define a means to gather data that does not exhaustively map between observed and represented variables, as a perfectly representational system of measurement requires. We note, however, that there is a mismatch in the stated aims of social institutions and pure science which would seem to permit a more liberal interpretation of representational measurement. Social institutions strive to effectively intervene on relevant communities, based on construct spaces that reflect context priorities. While these representations are never perfect, they are still acted upon in principled, meaningful ways. For example, how admissions represent college applicants might differ based on school quality. Where 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.

Given this representational style of measurement and intervention, we warn that adopting tools of automated decision-making puts these composite measures with groundings in social theory at risk. Instead, institutions may come to rely on measures developed operationally by ML systems. As a means of both confronting this loss of context and ensuring that representational measurements remain grounded in social reality, we propose the use of a social Measurement Assurance Program.

Outlining sMAP

Measurement assurance programs (MAP) are common in engineering disciplines. As defined by Speitel [21], 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 [2, 19]. 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." [5] As measured objects and interventions are in a dynamic relationship, 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 first and foremost ensure that (1) the methods of measurement are consistent and correctly implemented. As we are considering representational measurements, an sMAP also needs to (2) ensure gathered measures map onto the institution's construct space. Inspired by Campbell's law, another task unique to an sMAP is (3) accounting for the changes made to the observation space by the presence of the decision space. However, an sMAP cannot stop here. The representational systems of measurement employed by institutions stem from specific beliefs about how the world works. These social ontologies make up an institution's construct space, but are crucially subject to change by the advent of new evidence. As such, an sMAP further requires (4) a feedback mechanism to the constructs, such that they might be adjusted to better align with intervention aims. Figure 1 illustrates the sMAP adjustment to our model of institutional decision-making.

While an sMAP could incorporate diverse tools and approaches, for this work we specifically consider the ways in which ML can be repositioned as a tool in support of an sMAP. Through this repurposing of ML, we aim to show how ML's moral content is not defined by the algorithms it employs, but by the role it plays in making decisions.

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.

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 to more effectively map onto the 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.

Decision-to-Measurement Feedback Adjustment. Once sMAP serves a dynamic system that updates a given decision 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 questioned, 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 observation 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 students' attributes develop 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 both surface and *contour covariates* for decision-making. Not only are important measures revealed, unimportant constructs may be dismissed as context-inappropriate. 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 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 [3], 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 [3] 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 [16] 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 help both the evaluation of causal effects of future interventions and causal mechanisms behind historical inequalities visible to the model.

Rawls' framework is perhaps the most influential and systematic theory of deontic justice currently available. Put succinctly, Rawlsian justice advocates for (1) equal right to extensive basic liberties and (2) inequalities being permissible only insofar as they (a) work to the greatest benefit of the least advantaged, and (b) arise from processes that assure the equality of opportunity. 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. 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 Allegheny CYF and the AFST, CYF's services and goals map well onto the Rawlsian principles that define a just institution, comprising a natural setting for a just theory of measurement. Recall that *abuse* and *neglect* serve as constructs that are difficult to confirm outside of fatalities or near-fatalities. From here, we explore possible recommendations that stem from the components of a Rawlsian sMAP. These recommendations should be taken lightly, however, as the authors have not been directly involved with the Allegheny County CYF or its populace. Rather, this section provides a first glimpse of what an sMAP might require in practice.

Measurement Consistency and Validity

The bulk of [10] details how 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 recommends a disproportionate number of mandatory inspections of poor families. 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 – they are obligated by law to report children that show signs of *abuse* or *neglect*. Neglect, however, is easy to conflate with the effects of poverty. Furthermore, referrals largely come from professionals working with the poor and working classes, so the measurement *procedure* is inconsistent for the population as a whole. To render referral measurements more consistent with Rawlsian justice (holding that inequalities should work to the advantage of the worst-off), a just sMAP would suggest selective auditing referrals instead of processing them all uniformly. Another issue surrounding referrals that Eubanks [10] hints at is 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" [14]. This can result in a screening bias that almost ensures residual unfairness in AFST predictions [13]

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 support 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 such uncertainties 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.

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 agencies) onto *abuse* and *neglect*. The second maps from proxies of community re-referral and child placement onto *demonstrated risk*. In the first case, when AFST was being built, scraped variables were chosen based on correlation with risk predictions. The number of scraped variables in use has since been whittled down from a bit more grounding in social context [6], but this approach remains indicative of adopting an operational theory of measurement and prediction. Caseworkers, on the other hand, still implement a 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 directly 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, 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. When these steps are not taken, the liberty of an individual to be treated independent from their community is violated.

On the other side of the model, 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 are not likely to 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.

Decision-to-Measurement Feedback Adjustment

Families are often dramatically changed by support from CYF. Eubanks describes how CYF support is conditional upon families subjecting themselves to scrutiny and enrolling in time-consuming programs on work readiness and parenting skills. Thus, 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 lack 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 [10] 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 *neglect* in the decision-making procedure, beyond what is observed. Thus, an sMAP would suggest that where the *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.

CONCLUSION

We have pointed to the capability of ML to 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 programs whose existing deployment strategies are 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

[1] Chelsea Barabas, Madars Virza, Karthik Dinakar, Joichi Ito, and Jonathan Zittrain. 2018. Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment. In *Conference on Fairness, Accountability and Transparency*. 62–76.

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

[3] Reuben Binns. 2017. Fairness in Machine Learning: Lessons from Political Philosophy. *arXiv preprint arXiv:1712.03586* (2017).

[4] Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. 2019. Toward Algorithmic Accountability in Public Services. (2019).

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

[6] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*. 134–148.

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

[8] Robyn M Dawes, David Faust, and Paul E Meehl. 1989. Clinical versus actuarial judgment. *Science* 243, 4899 (1989), 1668–1674.

[9] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

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

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

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

[13] Nathan Kallus and Angela Zhou. 2018. Residual unfairness in fair machine learning from prejudiced data. *arXiv preprint arXiv:1806.02887* (2018).

[14] Annette Lareau. 2011. *Unequal childhoods: Class, race, and family life.* Univ of California Press.

[15] Jeff Larson. 2016. How We Analyzed the COMPAS Recidivism Algorithm. *Propublica* (2016). https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[16] Kasper Lippert-Rasmussen. 2014. *Born free and equal?: A philosophical inquiry into the nature of discrimination.* Oxford University Press.

[17] Zachary C Lipton and Jacob Steinhardt. 2018. Troubling Trends in Machine Learning Scholarship. *arXiv preprint arXiv:1807.03341* (2018).

[18] Sendhil Mullainathan and Jann Spiess. 2017. Machine learning: an applied econometric approach. *Journal of Economic Perspectives* 31, 2 (2017), 87–106.

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

[20] John Rawls. 2009. *A theory of justice.* Harvard university press.

[21] K.F. Speitel. 1982. Measurement Assurance. In *The Oxford Handbook of Innovation*, Gavriel Salvendy (Ed.). John Wiley and Sons, San Francisco.