Algorithmic Reparation

Abstract

Machine learning (ML) algorithms pervade contemporary society. They are integral to social institutions, inform processes of governance, and animate the mundane technologies of daily life. Consistently, the outcomes of ML reflect, reproduce, and amplify structural inequalities. The field of fair machine learning (FML) has emerged in response, developing mathematical techniques that increase fairness based on anti-classification, classification parity, and calibration standards. In practice, these computational correctives invariably fall short, operating from an *algorithmic idealism* that does not, and cannot, address systemic, intersectional stratifications. Taking present FML methods as our point of departure, we suggest instead the notion and practice of *algorithmic reparation.* Rooted in theories of Intersectionality, reparative algorithms name, unmask, and undo allocative and representational harms as they materialize in sociotechnical form. We propose algorithmic reparation as a foundation for building, evaluating, adjusting, and, when necessary, omitting and eradicating ML systems.

Introduction

In socially stratified societies, power concentrates but its mechanisms are diffuse. Power flows through governing bodies, social institutions, and micro-interactions, all of which entangle with technologies of the time. By default, technologies reflect and reinforce existing social orders, expressing and materializing hierarchical relations. However, technologies can also be tools of liberation. They can expose, undo, and reshape status quos. This latter project necessitates concerted and targeted efforts, underpinned by socially informed perspectives. In service of such efforts, we present *algorithmic reparation* as a concept and a scaffold for Intersectional[1] approaches to machine learning (ML) systems. Beyond improving code, a reparative approach uses computational tools for social intervention, while critically assessing when and where computation does not belong.

Algorithmic reparation is a transdisciplinary, sociotechnical proposal that converges theories of Intersectionality with acts of reparation, together applied to ML, with the goal of recognizing and rectifying structural inequality. Both Intersectionality and reparation have legal historical foundations, and each address systemic inequalities. Both have also now expanded beyond their legal origins via intellectual and activist movements. We continue these expansions, fusing Intersectionality and reparation into a cogent framework for critical algorithmic reform.

Algorithmic reform requires both social and technical expertise. Transdisciplinary collaboration is thus central to this proposal. Social theorists and computer scientists are equally vital for the design, production, and evaluation of equitable algorithmic systems, best achieved through tandem work. This does not mean perfunctory partnerships in which technicians work on one thing and theoreticians another, but meaningful collaboration and

[1] Throughout we capitalize "Intersectional" when referencing the theoretical paradigm, as is convention. We use a lower case "i" in all other circumstances.

cross-training (and cross-training *through* collaboration)[2] such that reforms emerge from the pools of multiple knowledge.

Our argument proceeds as follows: First, we review the problem of algorithmic inequality in ML—what it is, why it persists, and how technologists have attempted to address the issue. Next, we summarize key tenets of Intersectionality, link it to ML, and delineate how its pairing with reparation produces a critical orienting framework. With this foundation, we dig into the central techniques that drive the field of fair machine learning (FML), analyzing how and why these techniques are ineffective at combatting algorithmic inequality, and thus making the case for an alternative, reparative approach. Finally, we discuss methods for, and barriers to, implementing algorithmic reparation, addressing opportunities and constraints for a reparative algorithmic praxis.

Algorithmic Inequality in Machine Learning

An algorithm is simply a set of rules for completing a task. In computation, these are encoded mathematical directives which traditionally, have been written manually by computer programmers. ML uses a special type of algorithm developed via automated statistical inference procedures over large datasets (Barocas, Hardt and Narayanan, 2017; Kearns and Roth, 2019). ML is utilized by major institutions to guide criminal sentencing, welfare distributions, access to loans, hiring processes, and other resource allocations that shape opportunity structures for individuals and groups. ML also pervades everyday practices through search engines, dating applications, social media platforms, and entertainment streaming services. ML thus informs governance, shapes organizations, and weaves through the mundanities of daily life.

[2] We embody this call in the present work, co-authored by two sociologists and a computer scientist, with combined backgrounds in critical race theory, critical technology studies, communication theory, and machine learning.

The rationale for ML is pleasantly benevolent—to make institutional decisions fairer and to make tasks more convenient. However, the implementation of these systems consistently results in data driven outcomes that reflect and augment patterns of inequality (Benjamin, 2019; Costanza-Chock, 2020; Crawford et al. 2019; Noble, 2018; O'Neil, 2016). These patterns have been documented by journalists, academics, and activists over the past decade, exemplified by high profile cases of automation gone awry, such as Google Image results that conflate Black people with animals (Simonite, 2018), pricing algorithms that over-charge Asian communities for college test-prep services (Angwin, Mattu and Larson, 2015), and facial recognition tools that result in wrongful arrests due to poor fidelity with dark skin combined with racist patterns of over-policing (Hill, 2020). These harms are both allocative and representational, creating material divisions and reinforcing cultural stereotypes that devalue marginalized individuals and groups (Barocas, Crawford, Shapiro and Wallach (2017).

Why does Machine Learning Reproduce Inequality?

The fundamental reason that ML algorithms continue to reproduce inequality is because these technical systems are intrinsically and fundamentally social (Ames, 2018; Bucher, 2018; Kitchin, 2017; Seaver, 2017). Put simply, algorithms are animated by data, data come from people, people make up society, and society is unequal. Algorithms thus arc toward existing patterns of power and privilege, marginalization and disadvantage (Benjamin, 2016; Benjamin, 2019; Broussard, 2018; Browne, 2015; Costanza-Chock, 2020; Davis 2020).

Barocas, Hardt and Narayanan (2017) summarize the ML process as a pipeline that proceeds in four steps: capture and quantify what is (measure)→model generalizations from the training data (learn)→apply the model to novel inputs (action)→collect feedback and refine**.** Through the course of this pipeline, there are several specific, overlapping ways algorithmic inequalities materialize. They can be a product of unjust goals rooted in racist, sexist, heteronormative, ableist, nationalist, and/or colonialist priorities; they can derive from biased, nonrepresentational data; they can use biased proxies (e.g., arrest rates as an indicator of actual crime rates); and they can take real population differences that have been created through structural oppression and treat these differences as unproblematic and essential (e.g., health insurance pricing that penalizes Black men and rewards White women based on differential rates of chronic illness) (Caplan et al., 2018; Hoffmann, 2019).

Fair Machine Learning and Algorithmic Idealism

The problem of algorithmic inequality is not lost on computer scientists and engineers. Indeed, a vibrant field of fair machine learning (FML) has emerged with the shared goal of rectifying biases in ML systems (e.g., Barocas, Hardt and Narayanan, 2017; Chouldechova and Roth, 2020; Corbett-Davies and Goel, 2018; Kearns and Roth, 2019; Suresh and Guttag, 2019). A recent review categorizes technical FML solutions into three categories, which map onto distinct definitions of fairness: anti-classification, classification parity, and calibration (Corbett-Davies and Goel, 2018). We define and discuss each of these in a subsequent section. For now, the relevant point is that each of these solutions proposes a computational path toward fair algorithmic outcomes. However, despite laudable aims, the proposed solutions consistently fall short.

FML approaches fall short because they stem from what we refer to as *algorithmic idealism*, enacting computation that assumes a meritocratic society and seeks to neutralize demographic disparities. Such an approach will always be inadequate in a context that is fundamentally unjust (Fazelpour and Lipton, 2020). Algorithmic idealism begins with a base belief in equal opportunity, defining the problem of stratification as one caused by fallible human biases on the one hand, and imperfect statistical procedures, on the other. This perspective derives from illusory cultural narratives that misalign with the world that is—a world in which discrimination is entrenched, elemental, and compounding at the intersections of multiple marginalization. From their current theoretical packaging, FML proposals emerge disinterested and objective; they seek optimal precision in order to apportion risks and rewards evenly across neatly bounded identity-based groups. Such proposals are consistently eluded by the fairness they mean to achieve.

We take FML's idealism as our point of departure, proposing instead *algorithmic reparation*, which re-conceives society through a critical Intersectional lens. This approach strives not for social equality, which treats everyone the same, but for social equity, which provides resources based on differential need, thus accounting for deficits as they track onto axes of historical (dis)advantage (Cook and Hegtvedt, 1983; Deutsch, 1975; Rawls 1971). This means doing away with fairness and instead, coursing resources to those who have been systematically denied. This approach pairs the logics of Intersectionality with the praxis of reparation.

Algorithmic Reparation

Intersectionality as a Lens on Machine Learning

Intersectionality is not a singular theory, but an approach and a prism with a set of orienting assertions, goals, and tools. It undergirds critical theories across subfields—critical race theory, critical feminist studies, queer theory—all of which share a fundamental focus on systemic power relations that privilege and penalize, centralize and silence (Cho, Crenshaw and McCall 2013, Collins, 2019; Crenshaw, 1990; Hooks, 2000; Rahman, 2010). An Intersectional orientation is premised on the notion that identities are multiple and interrelated, shaped by, and filtered through, societal structures and institutions. These structures and institutions concentrate and compound opportunities and constraints in ways that reflect and reinforce essentialized hierarchical arrangements. However, these hierarchical arrangements are not predetermined, and practitioners of Intersectionality task themselves with revealing and undoing, systems of injustice (Chepp and Collins, 2013; Collins, 2002; Collins and Bilge, 2020).

Intersectionality has taken on various meanings and been deployed toward varied ends, while sustaining a core set of tenets (Cho, Crenshaw and McCall, 2013; Collins, 2019; Ferree, 2020; McCall, 2005). The main tenets of Intersectionality are that inequalities are systemic and entangled, meaning that identities cannot be understood apart from their interrelation with each other and from their imbrication with socio-structural systems; objectivity is not neutral, meaning positionality always matters and marginal subjects provide a necessary but undervalued lens; that inequalities manifest through legal, personal and professional (dis)advantage; and that hierarchies of power and privilege hide behind essentialisms, rendering their mechanisms imperceptible by default. These tenets combine with imperatives to expose and negate essentialisms; empower the marginalized; and to name, highlight, and challenge agents and structures of domination (Carastathis, 2016; Collins and Bilge, 2020; Ferree 2018).

Although Intersectionality has become embedded in academic texts and activist movements, it originates in the legal sector. Intersectionality arose in response to legal codes that erased and ignored co-occurring identity axes (e.g., Black women), working to account for discriminatory policies and practices that affect doubly marginalized legal subjects. With these legal foundations, proponents of Intersectionality emphasize the approach as an active political project (Cho, Crenshaw and McCall, 2013; Collins and Bilge, 2020). Intersectionality is not just something to think with, but something to *do*. It is an intellectual method, but also, and in the first instance, a tool for empowering people and fostering social justice (Collins and Bilge, 2020). Thus, beyond identifying cases of systemic disadvantage, an Intersectional project also works to surge resources to those who are marginalized and deprived. This imperative to treat Intersectionality as a grounded, practical, material endeavor, can be served through the application of Intersectionality to ML evaluation and design.

As an approach to ML, our deployment of Intersectionality joins with and builds on a growing body of work centralizing socio-historical power relations within computational systems. These include proposals for critical race methodologies for algorithmic fairness (Hanna et al., 2020), critical race theories applied to human-computer interaction (Ogbonnaya-Ogburu et al., 2020), decolonial AI (Mohamed, Png, and Isaac, 2020), decolonial computer science (Birhane and Guest, 2020), and affirmative action in algorithmic policing and criminal sentencing (Humerick, 2020; Skeem and Lowenkamp, 2020). Inspired by, and combining elements from each of these projects, algorithmic reparation has a fundamental foundation in praxis, an emphasis on the multiplex of intersecting identities, and an explicit position of compensatory resource redistributions accomplished proactively through a reparative approach.

A Reparative Approach

Bringing Intersectionality to bear on ML, and bringing ML to bear on Intersectionality, grounds Intersectional politics in material conditions that interplay with contemporary lived experience through computational forms of governance and mundane technical engagements. That is, the *doing* (and undoing) that drives Intersectionality converges directly with issues of algorithmic inequality (Benjamin, 2019; Costanza-Chock, 2020; Mann and Matzner, 2019). We suggest animating Intersectional politics through practices of reparation.

"Reparation" is a historically grounded mechanism by which offending parties symbolically and materially mend wrongdoings enacted against individuals and groups (Torpey, 2006). Reparations have been assigned in the context of war (Lu, 2017; Young, 2010), in acknowledgement of and apology for acts of colonialism (Gunstone, 2016;

Lenzerini, 2008), and they remain a point of mobilization for Black civil rights activists in the United States, demanding material recompense for the multigenerational damages of slavery and segregation (Bittker, 2018 [1972]; Coates, 2014; Henry, 2009). Reparative acts are not just backward looking, but also proactive, aiming to address the way historical wrongdoings affect current and future opportunity structures by channeling resources to make up for and overcome existing deficits.

Although traditionally applied in a legislative, often geo-political context, we use "reparation" in a broader sense, arguing for structural redress through algorithmic reform. This is more than the conceptual loosening of a legal term. Legal and political systems hinge reparation on identifiable culprits and victims along with demonstrable links between the wrongdoing of one party and the consequences of wrongful actions upon the aggrieved. However, this is rarely how structural, intersectional oppressions operate. What makes intersectional oppressions so pervasive and pernicious is their diffusion through institutional infrastructures, policies of governance, language, culture, individual attitudes, and interpersonal dynamics. The systematic, multifaceted, often subtle nature of intersectional inequality is at odds with linear relations of harm and blame. Algorithmic reparation thus incorporates redress into the assemblage of technologies that interweave macro institutions and micro interactions, embedding an equitable agenda into the material systems that govern daily life[3].

Our call for reparative algorithms is motivated by a broader mandate for equity and social justice, but it is also motivated by the specific conditions of automation that leave no

[3] Some theorists contest the use of reparation in a structural sense due to its uneasy fit with the specified relations of harm, and the adjudication of the specific wrongdoings that currently define reparative outcomes in legal settings (Young, 2010; Lu, 2017). We do not disagree with this point but depart from it, challenging the specified nature of reparation as a tool of redress when in practice, harms are often structural and diffuse, operating outside the scope of legal-political institutions.

neutral option (Broussard, 2018; Bucher, 2018; Noble, 2018; Mann and Matzner, 2019). In general, the distribution of resources can either reinforce inequalities, make them worse, or make them better. However, ML systems are intrinsically self-perpetuating in ways that ossify and intensify the outcomes they engender. This is because algorithms render decisions seemingly objective and divorced from human discretion; because they are opaque and inscrutable; and because their outcomes often have no technical means of undoing, even if circumstances call for correction (Bucher, 2018; Eubanks, 2018; Gillespie, 2014; Gillespie, 2018; Pasquale, 2015; Vaidhyanathan, 2018). Our proposal for algorithmic reparation assumes a moral duty to ameliorate, rather than aggravate, structural and historical stratifications as they manifest in computational code. This proposal sits in direct opposition to the prevailing logics of FML, which seek to de-bias algorithms and make them fairer. In contrast, a reparative approach assumes and leverages bias to make algorithms more equitable and just.

A Critical Read on FML: From Fair to Reparative

The field of fair machine learning is dedicated to making algorithms fairer for the people whom ML systems affect. In a review of the field, Corbett-Davies and Goel (2018) catalogue FML strategies, distinguishing between three definitions of fairness that underpin various computational solutions: *anti-classification, classification parity,* and *calibration*[4]. As lamented by the authors, these efforts have been largely unsuccessful, reconstituting the unjust social conditions they were designed to alleviate (Corbett-Davies and Goel, 2018:2).

[4] In addition to the fairness models reviewed here, causality-based notions of fairness are of burgeoning interest in FML. These models have not yet been implemented in mainstream ways and are thus not included in Corbett-Davis and Goel's (2018) review or in our paper. However, recent critiques of causal fairness in ML show similar vulnerabilities to existing approaches (see Hu and Kohler-Hausmann, 2020).

FML's troubles, we argue, stem from the field's foundation in *algorithmic idealism*— a meritocratic misconception of the world and a political ambivalence that this fallacy permits.

In this section, we describe existing FML solutions and the definitions of fairness to which they ascribe, highlight empirical instances in which these solutions proved lacking, and reimagine for each instance an alternative starting point derived from an Intersectional reparative approach. In doing so, we advance the case for algorithmic reparation in juxtaposition to idealist algorithmic reforms.

Anti-Classification

Anti-classification stipulates that algorithmic estimates not consider protected class attributes such as race, class, gender, or (dis)ability. This includes direct consideration of these characteristics as well as proxies for them. Corbett-Davis and Goel (2018) equate this to principles of equal protections under the law (Karst, 1977) and "taste-based" discrimination in economics (Becker, 2010 [1957]), by which advantages and disadvantages cannot be assigned based on demographic preference. Algorithmically, anti-classification systems strive to encode indifference to the identities of individuals who will be subject to automated outcomes.

Anti-classification principles underly automated employment programs that aim to circumvent managerial biases in candidate selection, avoiding the historical race-class-gender-age-nationality (dis)advantages that have historically shaped which candidates make it past initial screenings (Lahey and Oxley, 2018; Oreopoulos, 2011; Quillian et al., 2017). In practice, these algorithmic systems reproduce social hierarchies pervasive to the populations from which they select. Technology conglomerate Amazon's use of anti-classification algorithms exemplifies this point.

In 2014, Amazon developed a recruitment tool to aid in their own hiring processes. The tool used ML to sort applicants based on optimal fit for each position, removing social identity characteristics from consideration (Dastin, 2018). The trifold purpose was to increase efficiency, select the best candidates, and avoid implicit biases, especially against women, as this group has been (and remains) underrepresented in the technology sector (Beede et al., 2011; Harrison, 2019). However, by 2015 it became evident that the automated system was not operating as planned. Consistently, the recruitment algorithms assigned higher scores to men and lower scores to women. The reason for this is that the system was trained on the company's previous 10-years of employment data, which reflected a male-dominated sector. That is, Amazon's workforce, like the broader technology workforce, was populated disproportionately by men. Consequently, using existing data, the hiring system learned that men were the preferred candidates. This self-perpetuating cycle was so pronounced that any indicator of feminine gender identity in an application lowered the applicant's score. A degree from a women's college, participation in women-focused organizations, and feminized language patterns all reduced the evaluative outcome. Although Amazon attempted to adjust for these issues, the system continued to find proxies for gender and reward men at the expense of women. Amazon eventually retired the program (Dastin, 2018).

From an Intersectional perspective, anti-classification systems are intrinsically faulty. These systems are premised on erasure of difference, a flattening of demographic traits. Such an approach ideologically sidesteps the empirical reality of systemic inequality, but it cannot statistically or mathematically address it. The data that feeds these systems and the people who are subject to them operate from hierarchically differentiated positions. These distinctions are, and will continue to be, captured and reproduced through computation.

In contrast, a reparative approach would highlight, name, and encode hierarchical distinctions as they manifest across social identity categories. From this foundation, Amazon's algorithms would not invisiblize gender, but would instead define gender as a primary variable on which to optimize. This could mean weighting women, trans, and non-binary applicants in ways that mathematically bolster their candidacy, and potentially deflating scores that map onto stereotypical indicators of White cisgender masculinity, thus elevating women, trans, and non-binary folks in accordance with, and in rectification of, the social conditions that have gendered (and raced) the high-tech workforce. Moreover, it would not treat “woman” as a homogenous (binary) category, but would label and correct for intersections of age, race, ability, and other relevant variables that shape gendered experiences and opportunity structures.

This reparative system would do more than create fair conditions for underrepresented applicants. It would also, literally, value the contributions these applicants bring to the company while normalizing intersectional gender diversity in tech (Albert and Delano, 2021), such that high level positions and the pathways to them, are recast as plausible and expected across gender groups. The technical solution (women, trans, and non-binary individuals get a statistical boost) would thus have direct effects on the company’s work environment (more women, trans, and non-binary employees are hired at Amazon) and broader social effects on intersectional gendered social relations (women, trans, and non-binary folks in technology are normalized and the pathways to technology careers more seamless for these individuals to pursue). If these ends remain untenable, a reparative approach would indicate that ML ought not be used in hiring decisions.

Classification Parity

Classification parity is defined in terms of equal errors in classification across social identity groups. This aims to achieve parity in the error rates of predictive performance measures. Corbett-Davies and Goel (2018) identify several measures of classification error: false positive rates, false negative rates, precision, recall, the proportion of decisions that are positive, and the area under the ROC curve (AUC) (see Berk et al., 2018; Skeem and Lowenkamp, 2016). They focus in particular on false positives and proportion of decisions that are positive, as these are the error metrics that FML researchers have given the most attention (Corbett-Davies and Goel, 2018). We also focus on those metrics here, along with false negatives, as these are relevant to high profile cases of algorithmic inequality.

False positives and false negatives are errors in predicting how likely it is that something will (or will not) happen. Proportion of positive decisions, also known as "demographic parity" (Feldman et al., 2015), means that a given outcome distributes equally across social identity groups. These measures—false positives, false negatives, and demographic parity— have been central to debates about (and critiques of) ML in criminal sentencing, the most notable case of which is the COMPAS recidivism risk assessment tool.

COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is a widely used and commercially available tool designed to predict the likelihood that a criminal defendant will re-offend. In 2016, a ProPublica report analyzed pre-sentencing data from Broward County, Florida, a large jurisdiction using the COMPAS system. The report found that Black defendants were systematically assigned higher risk scores than White defendants, and that risk was *over*predicted for Black defendants and *under*predicted for White defendants (i.e., Black defendants recidivated at a lower rate than what the algorithm predicted, and White defendants recidivated at a higher rate than what the algorithm predicted) (Angwin et al., 2016a; Angwin et al. 2016b). The overall disparity in Black-White risk assessments represents an error of demographic parity, while the over-and-under prediction of Black-White recidivism represent errors of false positives (Black defendants) and false negatives (White defendants).

Classification parity is rooted in the assumption that if errors distribute evenly, then decisions will be fair, and that if data are accurate and representative enough, fair distributions can be achieved. Intersectionality indicates that these assumptions are misguided. They are misguided because the *effects* of risk are not equivalent across groups and because the data that feed into the ML system are imbricated with social policies and practices that shape statistical inputs and outputs. In terms of sentencing decisions, a criminal record and time in prison are undesirable for anyone. However, the negative effects of conviction and incarceration compound for individual Black defendants, flow on to their families and communities, and reinforce racial disparities in wealth, health, family stability, and mental wellbeing (Pettit and Western, 2004; Travis, Western, and Redburn, 2014; Western and Pettit, 2010; Western and Sirois, 2019). Moreover, the likelihood of contact with police and of conviction is significantly higher for poor Black men than any other group (Alexander, 2010). Indeed, criminal justice data are self- perpetuating, such that those groups defined as criminally "risky" are in fact, at disproportionate risk of ensnarement by the criminal justice system (Brayne, 2017; Brayne, Rosenblat and Boyd, 2015; Ferguson, 2017; Richardson, Schultz and Crawford, 2019; Skeem and Lowenkamp, 2020).

A reparative approach would supplant the goal of "parity" with, instead, systemic redress, beginning with the social facts of disproportionate risk between racial groups and the history of race-class dynamics that inform training data. From this, reparative decision aids would work to actively protect poor communities of color, especially poor Black men, over and above other subpopulations. This means the production and deployment of algorithms that keep Black men out of prison and keep police out of Black communities, defending against the criminalization of Blackness and rectifying racialized prison pipelines.

Calibration

Calibration specifies that "outcomes should be independent of protected attributes conditional on risk scores" (Corbett-Davies and Goel, 2018:6). Calibration can be thought of as a more nuanced take on anti-classification. The calibration approach to fairness is such that identity characteristics should only be considered by an algorithmic equation if those characteristics have demonstrable, empirical effects on the outcome under consideration.

That is, the system calibrates to differential risk levels between groups and assigns scores according to those base level differences.

To illustrate calibration, we remain with the COMPAS example. We do so because Northpointe, the company behind COMPAS, has responded to critics by claiming that in fact, their algorithms are fair because they satisfy calibration. What they mean is that a Black defendant classified as high risk by COMPAS is equally likely to recidivate as a White defendant classified as high risk. In an open letter to ProPublica, the company states:

> ProPublica focused on classification statistics that did not take into account the *different base rates of recidivism for blacks and whites*. Their use of these statistics resulted in false assertions in their article that were repeated subsequently in interviews and in articles in the national media (Dietrich, Mendoza and Brennan, 2016:1) (emphasis added).

Defending itself, Northpointe justifies its product based on calibration standards. Their defense is inadequate on both technical and social grounds.

On a technical level, although errors calibrate for base differences between groups, the *kinds* of errors are inconsistent. As detailed by ProPublica, Black defendants remain subject to disproportionate false positives and White defendants are rewarded with disproportionate false negatives (Angwin et al., 2016a). Black people are mis-assessed with overly strong risk scores and White people are mis-assessed with overly lenient risk scores. Concretely, this means that more Black people end up in jail and more White people remain free.

There are also non-technical reasons to be dissatisfied with Northpointe's response and in turn, dissatisfied with calibration as an algorithmic fairness standard. In particular, the data used by Northpointe to train their algorithms reflect racist policing tendencies in the United States that over-indict Black men, creating (not just reflecting) different base rates between raced, classed, and gendered groups (Brayne, 2017; Brayne, Rosenblat and Boyd, 2015; Ferguson, 2017; Richardson, Schultz and Crawford, 2019). Moreover, the carceral system not only responds to criminality, but through a constellation of mechanisms, also begets further violations (Alexander, 2010). Thus, Northpointe's reliance on calibration as a technical justification affirms and entrenches a system in which existing injustices act as the basis for their own amplified reproduction.

Like anti-classification, fairness through calibration seeks to remove identity from of the decision equation (though in qualified form). Like classification parity, fairness through calibration works to achieve equivalence between groups (by adjusting for differential base risk). As detailed in the subsections above, both of these objectives are ineffective for reducing inequality and may intensify inequitable social arrangements. In contrast, algorithmic reparation rejects the notion of identity erasure, even on the grounds of empirically distinct risk rates. It instead takes stock of social disparities as they map along axes of identity, exposing the underlying causes of differential risk and undoing the stratification that those differences both represent and produce. In practice, this could mean higher recidivism risk thresholds for Black defendants, lower thresholds for bail and parole, and weighted statistical adjustments that account for over policing in poor communities of color (see Skeem and Lowenkamp, 2020). Evaluated through a reparative Intersectional lens, any algorithm that did not address these base inequities would be deemed inadequate.

Methods and Barriers

The technical means of algorithmic reparation are already computationally viable, but its social effects can only take hold through meaningful implementation. It is thus to implementation that we now turn. Rather than reinvent the wheel, we select two recently proposed methods of algorithmic praxis that serve as possible tools of application for the reparative strategies discussed herein: archivist data curation and distributed AI power. Both of these methods are founded in trans-disciplinarity and require mutual collaborations between academic and non-academic actors. We also identify and discuss three challenges to implementing algorithmic reparation, including social, legal, and institutional barriers. Together, these methods and barriers ground algorithmic reparation within a context of both possibility and constraint.

Methods of Implementation

Archivist curation is one promising approach to implementing algorithmic reparation. This draws on the professional expertise of archival practice, honed by librarians and museum curators, applying these skills to ML data (Donovan, 2020; Jo and Gebru, 2020). Unjust algorithmic outputs are inextricable from problems with source data. These problems can be a function of representation in datasets and/or social factors that crystalize in data form. Managing these data issues can be prohibitively complex. However, professionals trained in collection and curation have skillsets that are transferrable to the ML sector, with Jo and Gebru (2020) noting *consent, inclusivity, power, transparency, and ethics & privacy* as data-relevant issues that have been well addressed in library sciences.

Drawing on their extant skillsets, curation professionals are capable of managing, collecting, arranging, and auditing data in ways that not only avoid re-entrenched inequalities, but optimize for marginal elevation, enacting targeted precision unachievable by those who are not professionally trained in curatorial methods. This includes the capacity to account for complex identity configurations in which advantages and disadvantages are in simultaneous operation, and the insight to determine which pieces of data are relevant to collect and, more importantly, what data ought not be collected. Such skills and practices are well suited to the problems discussed above, such as hiring and criminal sentencing, in which the complexity of the data and its entanglement with a multitude of confounding and compounding variables has proven intractable for data practitioners alone.

Distributed AI power is a second potential method. This method is premised on undoing standard power asymmetries between those who make, and those who are affected by, ML systems. The approach argues for tools that are legible to, and co-created with, impacted communities, especially those communities with histories of vulnerability prior to, and re-entrenched with, automation (Kalluri, 2020). Distributed AI power tactics rely on reciprocal engagement between developers and community stakeholders, with reverse pedagogies by which community stakeholders serve as experts in their lived experiences (Mohamed, Png and Isaac, 2020).

This method is exemplified by academic-activist collaborative projects undertaken by groups such as Data for Black Lives and the Algorithmic Justice League, which both challenge and partner with, commercial, governance, and regulatory bodies to enact technical, social, and policy changes. For example, Data for Black Lives, which coordinates thousands of engineers, mathematicians, and activists, is reportedly training former inmates in data science so that this directly affected population can actively participate in data-based reformation of the criminal justice system (Heaven, 2020). In turn, the Algorithmic Justice League has performed audits of race and gender in facial recognition technologies, leading several companies to revamp their programming in ways that improve classification accuracy for dark skinned women in image-search tasks (Raji and Buolamwini, 2019). Both organizations have also joined with others to activate against the use of facial recognition at all in policing, demonstrating the fundamental incongruity between these tools and racial justice, and resulting in a cascade of corporate and legislative moratoria (Heilweil, 2020; Flynn 2020). These projects begin with, are led by, and develop through, affected communities, with a record that demonstrates the capacity to enact reparative approaches to ML evaluation and design.

Barriers

Grounding algorithmic reparation means identifying both opportunities and challenges. The methods just discussed represent encouraging prospects, but there are empirical reasons that ML keeps reproducing inequality, and these realities are robust and obdurate. Enacting algorithmic reform requires unvarnished realism about the conditions under which any sociotechnical intervention will go into effect. For algorithmic reparation, implementation will face interrelated social, legal, and institutional barriers. Although addressing each barrier is beyond the scope of the present work, we lay them out to set clear terms for the path ahead.

Socially, reparation relies on a base logic that diverges from normative conceptions of fairness, opting instead for uneven resource allocations, targeted at the margins. As evidenced by the backlash against affirmative action policies and recent U.S. government-led opposition to critical race-based diversity trainings (Vought, 2020), the intentional reallocation of resources will, undoubtedly, come up against social resistance. Rectificatory tactics will be difficult to swallow for those who begin with the base (mis)assumption that society is functionally meritocratic, and this assumption is indeed deep-seated.

There will also be legal and institutional challenges. Reparation calls for centralized knowledge about and action based upon, protected class attributes. This is difficult under legal conditions that prohibit the collection of such data and/or its consideration in consequential decisions like employment, lending, school admissions, and criminal sentencing (Lieberwitz, 2008; Long and Batemen, 2020; Skeem and Lowenkamp, 2020). Similar prohibitions written into institutional policies will create blockades against algorithmic reparation within organizational settings.

There are also real challenges to the kinds of interdisciplinary and socially engaged collaborations necessary for reparative algorithmic projects. Power and compensation disparities persist between computer scientists and social scientists, and between academic and non-academic organizations (Carrigan and Bardini, 2021; Hackett and Rhoten, 2011; Stavrianakis, 2015; Viseu, 2015) along with epistemological schisms that are difficult to reconcile (Bauer 1990; Richter and Paretti, 2009). These impediments to meaningful inter/trans/non-disciplinary collaboration are exacerbated by academic incentive structures that reward traditional intra-disciplinary outputs over and above hybrid and expansively defined research products (Woelert and Millar, 2011), despite widespread statements about the value of disciplinary blending and community engaged science (Hackett and Rhoten, 2011, Viseu, 2015). Contending with these institutional challenges means considering not only who will do the work of algorithmic reparation, but also how it can be done across sectors, with the support of leadership, mechanisms of accountability, democratic oversight, and equitable returns for practitioners' labor.

Summary and Conclusions

Technologies reflect and create the societies from which they stem and in which they proliferate. By default, technologies will embody the values of the powerful and reconstitute the stratified hierarchies those values represent (Benjamin, 2016; Benjamin, 2019; Broussard, 2018; Browne, 2015; Costanza-Chock, 2020; Davis, 2020). These patterns of reflection, reconstitution, and in turn, amplification of structural inequality have borne out in spectacular fashion with the integration of ML systems into personal and institutional life (Broussard, 2018; Crawford et al., 2019; Eubanks, 2018; Noble, 2018).

The field of FML has emerged in response, with computer scientists and engineers proposing myriad technical fixes to the injustices of automation. Yet, algorithmic inequalities persist. In their efforts to hide, distribute evenly between, and calibrate social identity traits, FML practitioners operate with a goal of fairness and equality when instead, equity and reparation are required. We make this case in the body of the text above, suggesting instead algorithmic reparation as an anti-oppressive, Intersectional approach. We intend for this approach to guide algorithmic design and to act as an evaluative standard by which existing algorithmic systems are judged, adjusted, and where necessary, omitted or dismantled. Our proposal is thus geared toward building better systems and holding existing ones to account.

We highlight two possible methods of implementation—professionalized archival data curation and distributed AI power. Both methods are consonant with the base assumptions and objectives of algorithmic reparation and they both show promise as practical means for algorithmic reform. We also take stock of social, legal, and institutional barriers to implementation, providing a realistic perspective on the work ahead.

Continuing this focus on the work ahead, we conclude by considering next steps in the ongoing project toward social and technical reform. Here we emphasize the need for context-specific attention, more and multiple tools, and multi-pronged approaches that converge technical, social, and institutional efforts.

Instruments of social change—technical or otherwise—never operate in a vacuum. In the final substantive section of this paper, we selected two newly introduced mechanisms by which algorithmic reparation might be implemented. Testing these in diverse empirical settings will reveal how they function, where they fall short, and what kinds of infrastructural conditions will be required for these methods to take meaningful effect.

It will also be vital to explore and create a cache of methods and tools, addressing specific needs, specific conditions, and creating interoperability between social and technical systems. The acute need for a constellation of methods and tools becomes clear when we consider the varied and engrained structural reasons why inequalities continue to manifest in algorithmic form. Algorithmic reparation will, necessarily, run against the grain of multiple status quos, requiring numerous iterations, agile applications, and persistent adjustments for this uphill endeavor.

In service of creating a robust toolbox, [third author name removed for peer-review] is currently leading a project to devise technical instruments that audit and optimize for inequality reduction in decision systems. This is a computational mechanism that centers impact estimations that most reduce inequality in automated decision outputs. These auditing tools are intended specifically for institutional decision aids, such as those used in hiring processes, loan allocations, and admission decisions, calibrated to the particular inequalities of the communities affected. Projects such as this, which are currently in development, portend a new and critically informed landscape of sociotechnical relations.

We also note that “next steps” cannot be exclusively technical, but must enjoin a multi-pronged effort that will result necessarily, from the work of many hands (Abebe et al. 2020, D’Ignazio and Klein 2020). Any algorithmic solution to social problems is necessarily partial and incomplete, requiring complementary social, legal, and institutional evolutions. Concretely, this means rethinking discrimination policies that erase and thus ignore identity attributes; reworking institutional incentive structures and power arrangements that silo academic disciplines from each other and from the public sector; introducing regulatory implements that capture and censure discriminatory algorithmic outputs; and forming organizational bodies dedicated to auditing technical systems and assuring their allocative and representational ends.

In practice, the problems of algorithmic systems are the problems of social systems, and meaningful solutions will be technical *and* social in nature. These solutions will not come easy, nor fast, nor with absolute finitude. Just as undoing racism, sexism, classism and colonialism are continuous, evolving, and nonlinear projects, so too is the journey to unmake inequality with algorithms and code. Algorithmic reparation is not a final or encompassing answer, but a critical, equitable, Intersectional foundation.

References

Abebe R, Barocas S, Kleinberg J, et al. (2020) Roles for computing in social change. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.* 252-260.

Albert K and Delano M (2021) "This whole thing smacks of gender": Algorithmic exclusion in bioimpedance-based body composition Analysis. arXiv preprint arXiv:2101.08325.

Alexander M (2010) *The new Jim Crow: Mass incarceration in the age of colorblindness.* NY: The New Press.

Ames MG (2018) Deconstructing the algorithmic sublime. *Big Data & Society* 5(1): 1-4.

Angwin J, Larson J, Mattu S, et al. (2016b) Machine bias: There's a software used across the country to predict future criminals. And it's biased against Blacks. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Angwin J, Larson J, Mattu S, et al. (2016a) How we analyzed the compas recidivism algorithm (2016). *ProPublica*. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Angwin J, Mattu S and Larson J (2015) The tiger mom tax: Asians are nearly twice as Likely to get a higher price from Princeton Review. *ProPublica*. https://www.propublica.org/article/asians-nearly-twice-as-likely-to-get-higher-price-from-princeton-review

Barocas S, Crawford K, Shapiro A, et al. (2017) The problem with bias: Allocative versus representational harms in machine learning. *9th Annual Conference of the Special Interest Group for Computing, Information and Society.*

Barocas S, Hardt M and Narayanan A (2017b) Fairness in machine learning:Limitations and Opportunities. *NIPS Tutorial*, 1.

Bauer HH (1990) Barriers against interdisciplinarity: implications for studies of science, technology, and society (STS. *Science, Technology, & Human Values* 15(1): 105-119.

Becker GS (2010 [1957]) *The economics of discrimination.* Chicago, IL: University of Chicago press.

Beede DN, Julian TA, Langdon D, et al. (2011) Women in STEM: A gender gap to innovation. *Economics and Statistics Administration Issue Brief*.(04-11).

Benjamin R (2016) Catching our breath: Critical race STS and the carceral imagination. *Engaging Science, Technology, and Society* 2: 145-156.

Benjamin R (2019) *Race after technology: Abolitionist tools for the new jim code.* Cambridge, UK: Polity.

Berk R, Heidari H, Jabbari S, et al. (2018) Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*. 0049124118782533.

Birhane A and Guest O (2020) Towards decolonising computational sciences. *arXiv preprint arXiv:2009.14258*

Bittker B (2018 [1972]) *The case for Black reparations.* Boston, MA: Beacon Press.

Brayne S (2017) Big data surveillance: The case of policing. *American Sociological Review* 82(5): 977-1008.

Brayne S, Rosenblat A and boyd d (2015) Predictive policing. *Data & Civil Rights: A New Era Of Policing And Justice*. https://datacivilrights.org/pubs/2015-1027/Predictive_Policing.pdf

Broussard M (2018) *Artificial unintelligence: How computers misunderstand the world.* Cambridge, MA: MIT Press.
Browne S (2015) *Dark matters: On the surveillance of blackness.* Durham, NC: Duke University Press.
Bucher T (2018) *If... Then: Algorithmic Power and Politics.* NY: Oxford University Press.
Caplan R, Donovan J, Hanson L, et al. (2018) Algorithmic accountability: A primer. NY: Data & Society. https://datasociety.net/library/algorithmic-accountability-a-primer/
Carastathis A (2016) *Intersectionality: Origins, contestations, horizons.* Lincoln, NE: University of Nebraska Press.
Carrigan C and Bardini M (2021) Majorism: Neoliberalism in student culture. *Anthropology & Education Quarterly*. In press https://doi.org/10.1111/aeq.12361.
Chepp V and Collins PH (2013) Intersectionality. In: Celis K, Kantola J, Waylen G, et al. (eds) *The Oxford Handbook of Gender and Politics*. New York: Oxford University Press, pp.57-87.
Cho S, Crenshaw KW and McCall L (2013) Toward a field of intersectionality studies: Theory, applications, and praxis. *Signs: Journal of Women and Culture in Society* 38(4): 785-810.
Chouldechova A and Roth A (2020) A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM* 63(5): 82-89.
Christin A, Rosenblat A and boyd d (2015) Courts and predictive algorithms. NY: Data & Society. https://datasociety.net/library/data-civil-rights-courts-and-predictive-algorithms/
Coates T-N (2014) The case for reparations. *The Atlantic*, 313. https://www.theatlantic.com/magazine/archive/2014/06/the-case-for-reparations/361631/
Collins PH (2002) *Black feminist thought: Knowledge, consciousness, and the politics of empowerment.* London, UK: routledge.
Collins PH (2019) *Intersectionality as critical social theory.* Durham, NC: Duke University Press.
Collins PH and Bilge S (2020) *Intersectionality.* Medford, MA: Polity Press.
Cook KS and Hegtvedt KA (1983) Distributive justice, equity, and equality. *Annual Review of Sociology* 9(1): 217-241.
Corbett-Davies S and Goel S (2018) The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*. https://arxiv.org/abs/1808.00023
Costanza-Chock S (2020) *Design justice: Community-led practices to build the worlds we need.* MIT Press.
Crawford K, Roel Dobbe, Theodora Dryer, et al. (2019) AI NOW 2019 Report. NY: AI NOW Institute. https://ainowinstitute.org/AI_Now_2019_Report.pdf
Crenshaw K (1990) Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review* 43(6): 1241-1299.
Dastin J (2018) Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*, 9. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G
Davis A (2000) Masked racism: reflections on the prison industrial complex.[Article reprinted from Colorlines]. *Indigenous Law Bulletin* 4(27): 4.
Davis JL (2020) *How artifacts afford: The power and politics of everyday things.* Cambridge, MA: MIT Press.
Deutsch M (1975) Equity, equality, and need: What determines which value will be used as the basis of distributive justice? *Journal of Social issues* 31(3): 137-149.

Dieterich W, Mendoza C and Brennan T (2016) COMPAS risk scales: Demonstrating accuracy equity and predictive parity. Northpointe Inc. http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf.
Donovan J (2020) You purged racists from your website? Great, now get to work. . *Wired*. https://www.wired.com/story/you-purged-racists-from-your-website-great-now-get-to-work/
Eubanks V (2018) *Automating inequality: How high-tech tools profile, police, and punish the poor*. New York: St. Martin's Press.
Fazelpour S and Lipton ZC (2020) Algorithmic fairness from a non-ideal perspective. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, NY: pp.57–63. Association for Computing Machinery.
Feldman M, Friedler SA, Moeller J, et al. (2015) Certifying and Removing Disparate Impact. *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining.* Sydney, Australia, 259-268.
Ferguson AG (2017) Illuminating black data policing. *Ohio State Journal of Criminal Law* 15: 503-525.
Ferree MM (2018) Intersectionality as theory and practice. *Contemporary Sociology: A Journal of Reviews* 47(2): 127-132.
Flynn S (2020) 13 Cities where police are banned from using facial recognition tech. *Innovation and Tech Today* https://innotechtoday.com/13-cities-where-police-are-banned-from-using-facial-recognition-tech/
Gillespie T (2014) The relevance of algorithms. In: Gillespie T, Boczkowski PJ and Foot KA (eds) *Media technologies: Essays on communication, materiality, and society*. Cambridge, MA: MIT Press, pp.167-193.
Gillespie T (2018) *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media.* New Haven, CT: Yale University Press.
Gunstone A (2016) Reconciliation, reparations and rights. In: Short CLaD (ed) *Handbook of Indigenous Peoples' Rights*. London, UK: Routledge.
Hackett EJ and Rhoten DR (2011) Engaged, embedded, enjoined: Science and technology studies in The National Science Foundation. *Science and engineering ethics* 17(4): 823-838.
Hanna A, Denton E, Smart A, et al. (2020) Towards a critical race methodology in algorithmic fairness. *Proceedings of the 2020 conference on fairness, accountability, and transparency.* 501-512.
Harrison S (2019) Five years of tech diversity reports—and little progress. *Wired*. https://www.wired.com/story/five-years-tech-diversity-reports-little-progress/
Heaven WD (2020) Predictive policing algorithms are racist. They need to be dismantled. *MIT Technology Review*. *https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/*
Heilweil R (2020) Big tech companies back away from selling facial recognition to police. That's progress. *Vox* https://www.vox.com/recode/2020/6/10/21287194/amazon-microsoft-ibm-facial-recognition-moratorium-police
Henry CP (2009) *Long overdue: The politics of racial reparations.* New York: NYU Press.
Hill K (2020) Another arrest, and jail time, due to a bad facial recognition match. *The New York Times*. https://www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail.html
Hoffmann AL (2019) Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society* 22(7): 900-915.

Hooks B (2000) *Feminist theory: From margin to center.* London, UK: Pluto Press.
Hu L and Kohler-Hausmann I (2020) What's sex got to do with machine learning. *arXiv preprint arXiv:2006.01770*.
Humerick JD (2019) Reprogramming fairness: Affirmative action in algorithmic criminal sentencing. *HRLR Online* 4(2): 213-244.
Jo ES and Gebru T (2020) Lessons from archives: Strategies for collecting sociocultural data in machine learning. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.* 306-316.
Kalluri P (2020) Don't ask if AI is good or fair, ask how it shifts power. *Nature World View*. https://media.nature.com/original/magazine-assets/d41586-020-02003-2/d41586-020-02003-2.pdf
Karst KL (1977) Foreword: Equal citizenship under the fourteenth amendment. *Harvard Law Review* 91: 1-301.
Kaufman RL (2010) *Race, gender, and the labor market: Inequalities at work.* Boulder, CO: Lynne Rienner Publishers
Kayser-Bril N (2020) Google apologizes after its Vision AI produced racist results. *Algorithm Watch. https://algorithmwatch.org/en/story/google-vision-racism/*
Kearns M and Roth A (2019) *The ethical algorithm: The science of socially aware algorithm design.* Oxford, UK: Oxford University Press.
Kendi IX (2019) *How to be an antiracist.* New York: One world.
Kitchin R (2017) Thinking critically about and researching algorithms. *Information, Communication & Society* 20(1): 14-29.
Lahey JN and Oxley DR (2018) Discrimination at the intersection of age, race, and gender: Evidence from a lab-in-the-field experiment. Reportno. Report Number|, Date. Place Published|: Institution|.
Learned-Miller E, Ordóñez V, Morgenstern J, et al. (2020) Facial recognition technologies in the wild: A call for a federal office. Algorithmic Justice League https://global-uploads.webflow.com/5e027ca188c99e3515b404b7/5ed1145952bc185203f3d009_FRTsFederalOfficeMay2020.pdf

Lenzerini F (2008) *Reparations for indigenous peoples: international and comparative perspectives.* Oxford, UK: Oxford University Press.
Lieberwitz R (2008) Employment discrimination law in the United States: on the road to equality? *New developments in Employment Discrimination Law, Bulletin of Comaparative Labour Relations*. *New Developments in Employment Discrimination Law Conference.* Tokyo, Japan. https://www.jil.go.jp/event/ro_forum/resume/080220/USA_.pdf
Lu C (2017) *Justice and reconciliation in world politics.* Cambridge, UK: Cambridge University Press.
Maheshwari S and Stevenson A (2017) Google and Facebook face criticism for ads targeting racist sentiments. *The New York Times*. https://www.nytimes.com/2017/09/15/business/facebook-advertising-anti-semitism.html
Mann M and Matzner T (2019) Challenging algorithmic profiling: The limits of data protection and anti-discrimination in responding to emergent discrimination. *Big Data & Society* 6(2): 1-11.
McCall L (2005) The complexity of intersectionality. *Signs: Journal of Women in Culture and Society* 30(3): 1771-1800.

Mohamed S, Png M-T and Isaac W (2020) Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology* 33(4): 659-684.

Noble SU (2018) *Algorithms of oppression: How search engines reinforce racism.* NY: NYU Press.

Ogbonnaya-Ogburu IF, Smith AD, To A, et al. (2020) Critical race theory for HCI. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* 1-16.

O'Neil, C (2016) *Weapons of math destruction: How big data increases inequality and threatens democracy.* NY: Broadway Books.

Oreopoulos P (2011) Why do skilled immigrants struggle in the labor market? A field experiment with thirteen thousand resumes. *American Economic Journal: Economic Policy* 3(4): 148-171.

Pasquale F (2015) *The black box society: The secret algorithms that control money and information.* Cambrdge, MA: Harvard University Press.

Pettit B and Western B (2004) Mass imprisonment and the life course: Race and class inequality in US incarceration. *American sociological review* 69(2): 151-169.

Quillian L, Pager D, Hexel O, et al. (2017) Meta-analysis of field experiments shows no change in racial discrimination in hiring over time. *Proceedings of the National Academy of Sciences* 114(41): 10870-10875.

Rahman M (2010) Queer as intersectionality: Theorizing gay Muslim identities. *Sociology* 44(5): 944-961.

Raji ID and Buolamwini J (2019) Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society.* 429-435.

Rawls J (1971) *A theory of justice.* Cambridge, MA: Harvard university press.

Richardson R, Schultz JM and Crawford K (2019) Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *NYUL Rev. Online* 94: 15-55.

Richter DM and Paretti MC (2009) Identifying barriers to and outcomes of interdisciplinarity in the engineering classroom. *European Journal of Engineering Education* 34(1): 29-45.

Seaver N (2017) Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society* 4(2): 1-12.

Simonite T (2018) When it comes to gorillas, Google Photos remains blind. *Wired.* https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/

Skeem JL and Lowenkamp CT (2016) Risk, race, and recidivism: Predictive bias and disparate impact. *Criminology* 54(4): 680-712.

Skeem JL and Lowenkamp CT (2020). Using algorithms to address trade-offs inherent in predictive recidivism. *Social Sciences & the Law* (Forthcoming, Available at SSRN: https://ssrn.com/abstract=3578591 or http://dx.doi.org/10.2139/ssrn.3578591).

Stavrianakis A (2015) From anthropologist to actant (and back to anthropology): Position, impasse, and observation in sociotechnical collaboration. *Cultural Anthropology* 30(1): 169-189.

Suresh H and Guttag JV (2019) A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*. https://arxiv.org/abs/1901.10002

Torpey JC (2006) *Making whole what has been smashed: On reparations politics.* Cambridge, MA: Harvard University Press.

Travis J, Western B and Redburn FS (2014) The growth of incarceration in the United States: Exploring causes and consequences. Reportno. Washington, DC: National Research Council. https://www.nap.edu/catalog/18613/the-growth-of-incarceration-in-the-united-states-exploring-causes
Vaidhyanathan S (2018) *Antisocial media: How Facebook disconnects us and undermines democracy.* New York: Oxford University Press.
Viseu A (2015) Integration of social science into research is crucial. *Nature News* 525(7569): 291.
Vought R (2020) Memorandum for the heads of executive departments and agencies: Training in the federal government. *United States Office of Management and Budget* https://www.whitehouse.gov/wp-content/uploads/2020/09/M-20-34.pdf
Western B and Pettit B (2010) Incarceration & social inequality. *Daedalus* 139(3): 8-19.
Western B and Sirois C (2019) Racialized re-entry: labor market inequality after incarceration. *Social Forces* 97(4): 1517-1542.
Woelert P and Millar V (2013) The 'paradox of interdisciplinarity' in Australian research governance. *Higher Education* 66(6): 755-767.
Young IM (2010) *Responsibility for justice.* Oxford, UK: Oxford University Press.