The Perils of Objectivity:
Towards a Normative Framework for Fair Judicial Decision-Making

Andi Peng
Adaptive Systems and Interaction Group
Microsoft Research
Redmond, WA, USA
andipeng@microsoft.com

Malina Simard-Halm
Institute of Criminology
University of Cambridge
Cambridge, UK
mjs296@cam.ac.uk

Abstract

Fair decision-making in criminal justice relies on the recognition and incorporation of infinite shades of grey. In this paper, we detail how algorithmic risk assessment tools are counteractive to fair legal proceedings in social institutions where desired states of the world are contested ethically and practically. We provide a normative framework for assessing fair judicial decision-making, one that does not seek the elimination of human bias from decision-making as algorithmic fairness efforts currently focus on, but instead centers on sophisticating the incorporation of *individualized* or *discretionary bias*—a process that is requisitely human. Through analysis of a case study on *social disadvantage*, we use this framework to provide an assessment of potential features of consideration, such as political disempowerment and demographic exclusion, that are irreconcilable by current algorithmic efforts and recommend their incorporation in future reform.

Introduction

The prison and its penumbra of control have become spectacles of social and economic inequality (Alexander 2012). The increased deployment of algorithmic risk assessment tools to aide judicial decision-making has been consistently found to reflect and exacerbate problematic stereotypes related to race, class, and gender in an increasingly unequal society (Green and Chen 2019; Angwin et al. 2016; O'Neil 2016). This indicates an issue and an irony—that the ostensible instruments of "justice" have in fact come to deepen the contours of historic disadvantage. Moreover, the technical foundations of algorithmic design necessitates the existence of an ideal world state for system builders to optimize for—one that does not and, we argue, should not exist in the real world. In this paper, we contend that current risk assessment efforts in criminal justice, including work on algorithmic fairness, are unproductive due to the misguided desire to eliminate human bias from decision-making processes. We provide a normative framework for assessing fair judicial decision-making, one that seeks to incorporate the value of *discretionary bias*, not its elimination, from decision-making processes.

We begin by reviewing the history of criminal risk assessment, including the failures of earlier systems designed to predict the "dangerousness" of potential criminals (Scott 1977). We detail *selective incapacitation theory* as a flawed ethical framework for assessing the social utility of risk assessment tools, including how the costs of false positive prediction errors were ultimately considered too high for system adoption (Harvard Law Review 1982). We overview the present day status of algorithmic tools in the justice system and argue that this current effort represents the next chapter in a long-running trend of risk mitigation which unwisely undermines the spirit, if not the letter, of individual liberty.

We then turn to an analysis of a young but rapidly growing body of work on algorithmic fairness, which details efforts to fix the unsavory yet increasingly prevalent consequences derived from algorithms' unequal treatment of different groups. We review challenges faced by the computer science community at addressing these disparities, including how fairness as a metric is one that is difficult and optimize for. In this section, we make the argument that irrespective of techniques developed, technical attempts at achieving algorithmic parity will remain normatively unproductive so long as there is no consensus on a desired world state.

Next, we contribute a normative framework for assessing what constitutes fair judicial decision-making—one that is derived from primary philosophical principles of justice systems, the role of procedural legitimacy, and the importance of judicial discretion. We contend that the incorporation of improved human *discretionary bias*, not its elimination, from judicial decision-making processes is paramount to effective criminal justice reform in systems where highly complex and constantly evolving aims exist. We use this framework to assess the case study of *social disadvantage* and highlight how important features of consideration, such as political disempowerment and demographic exclusion, that were previously irreconcilable by algorithmic optimization functions can now be assessed.

To conclude, we offer future directions of reform for the criminal justice system, including the role of algorithms within it, that seek to capture interdisciplinary efforts towards establishing a more comprehensive notion of fairness in judicial decision-making.

The Questionable Practice of Risk Assessment

Algorithmic risk assessment tools are utilized widely in the criminal justice system to assign bail, predict recidivism, and police cities (Angwin et al. 2016; Hvistendahl 2016). Although theories of risk mitigation are not new to the domain of criminal justice, the wide proliferation of machine learning advances in the past decade have led to an increase in the sophistication and subsequent deployment of automated systems. However, algorithms have been found to unfairly discriminate against different groups on a variety of judicial decision-making processes along racial, gender, and socioeconomic lines (Green and Chen 2019; Angwin et al. 2016). Consequently, a rising body of work on algorithmic fairness has focused on ensuring these systems are trained to be quantitatively fair and impartial.

This section outlines the limitations of these efforts by:

1. Overviewing failed historic attempts at risk assessment
2. Paralleling those attempts to present day trends
3. Describing how the technical frameworks for training these systems presupposes an ideal world state—one that currently does not and should not exist in the real world

Risk Assessment's Spotty Record

Although the past decade has witnessed an explosion in the deployment of "algorithmic decision aides" or "risk assessment tools" in criminal justice, the notion of a third-party system being used to aid legal decision-making originates from the 1920s, when actuarial tools to assess risk were developed and implemented in correctional settings to predict future criminal behavior (Mathiesen 1998). Early proponents of these systems asserted that offloading decision-making to an automated model would promote a more efficient justice system by identifying an element of "dangerousness" in offenders—namely, an individual's capacity to commit future crime (Scott 1977). Known as *selective incapacitation theory*, this logic presupposed that by identifying a subset of individuals who are particularly prone to violence or recidivism (colloquially known as "career criminals") and keeping them incapacitated in prison, society would experience an overall reduction in crime (Lewin 1982).

This concept of punishing individuals not for what they've done in the past but rather for what they may do in the future represented a drastic shift in theories of sentencing (Cohen 1983). In 1982, a report released by the RAND Corporation challenged this assumption when it conducted a longitudinal survey on risk assessment by surveying inmates in California, Texas, and Michigan over a six-year period (Greenwood and Abrahamse 1982). The report found that while its predictive scale was "reasonably" accurate at identifying repeat low-risk offenders (76%), it was highly inaccurate at identifying high-rate offenders (45%) and resulted in a false-positive prediction rate of 55%. In other words, more than half of supposed high-risk offenders had been incorrectly labeled by the model as being likely to commit a crime when they never did.

Although proponents of *selective incapacitation theory* maintained that risk assessment remained an improvement over existing judicial processes due to the elimination of human judgement and subjectivity from decision-making, public debate raged over concerns of predictive accuracy and individual fairness (Wright 1994; Mathiesen 1998). In false negative cases, individuals mistakenly predicted as unlikely to recommit but subsequently did were allowed back into a society that had attempted to screen for them. False positives, more dangerously, represented an existential threat to individual liberty and fundamental American rights like the presupposition of innocence (The Economist 2015). The costs of such mistakes, however small, were considered too high to warrant broad legal adoption, and the theory ultimately faded from mainstream considertion (Desmarais and Singh 2013).

Risk Assessment: Renewed but Not Redefined

The rise of machine learning advances of the past decade have resulted in a proliferation of newer, more sophisticated criminal risk assessment tools. Incorporation of more complex feature analysis, higher-dimensionality datasets, and improved training techniques have resulted in higher than ever seen before accuracy rates on prediction tasks, renewing public interest in these tools once more (Desmarais and Singh 2013; O'Neil 2016). Although the specific systems deployed by jurisdictions differ by state or even county, most flavors have been adapted from three main systems: Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), Public Safety Assessment (PSA), and Level of Service Inventory Revised (LSI-R). COMPAS, created by the for-profit company Northpointe, assesses features related to criminal involvement, individual lifestyle, personality, family, and social exclusion (Northpointe 2012). LSI-R, developed by Canadian company Multi-Health Systems, also utilizes features from criminal history and personality. PSA, developed by the Laura and John Arnold Foundation, only utilizes criminal history and age (O'Neil 2016).

Today, COMPAS is the most popular tool and is deployed by dozens of courts and correctional facilities across the country to inform a wide-array of judicial decision-making processes regarding pretrial release, probation, institutional programming, reentry, parole, and sentencing (Angwin et al. 2016). The recidivism prediction instrument (RPI) metric of COMPAS is called the Recidivism Risk Scale and is computed from 137 questions, which are either personally answered by a defendant or are determined automatically from a defendant's criminal record. Because COMPAS is proprietary software, there is little to no transparency regarding its training and prediction techniques, with federal oversight virtually nonexistent (Diakopoulos 2016; O'Neil 2016). Its widespread usage resulted in a 2016 report published by ProPublica detailing concerns that ring eerily reminiscent of those from the 1982 RAND report. Using COMPAS, ProPublica found that not only did prediction errors exist, the classifiers also demonstrated *disparate outcomes* towards the treatment of different races, with black defendants who did not recidivate almost twice as likely to be classified as recidivists compared to white defendants who did not recidivate (Angwin et al. 2016). Of the black defendants who did not go on to recidivate, 44.9% of them were misclassified as recidivists whereas of the white defendants, only 23.5% were misclassified. In other words, ProPublica found that the false positive error rate for black defendants was almost twice as high as the false positive error rate for white defendants, contributing an added layer of racial complexity to already existing concerns of false prediction.

Although more sophisticated, risk assessment tools today suffer from the same moral quandaries of those in the past. Machine learning techniques must still look to the past to predict the future—a practice that violates the spirit, if not the legality, of the presupposition of innocence. Although much like historic proponents of selective incapacitation, advocates for the widespread use of risk assessments today appear to be doing so out of a genuine desire to improve upon untethered human judgement, but the alarm bells that were raised in the 1980s regarding threats to individual liberty from false positive errors ring ever more today.

Algorithmic Fairness' Misguided Efforts

With this reinvigoration of algorithmic-aided decision-making in the public sphere, it has now become more important than ever for society to quantify and understand the biases in machine learning models that reinforce the disadvantaged status of different groups (Kleinberg et al. 2019; Barocas and Selbst 2016). Risk assessment tools are increasingly utilized in the criminal justice system because they are perceived to be less biased and more accurate than human predictions, but the numbers suggest otherwise (O'Neil 2016). A parity in classification accuracy between different racial and gender classes, whether due to unrepresentative training data or reinforced systemic world biases, remains highly concerning for system designers (Bolukbasi et al. 2016). Moreover, not only are these instances of unfair treatment harmful for one-off decision-making scenarios, the repeated deployment of such algorithms have been found to actually compound imbalances and lead to exacerbated inequalities over time—a phenomenon known as the "leaky pipeline" (Romanov et al. 2019).

Spurred by such concerns, the computer science community has responded with a rapidly growing body of work on algorithmic fairness, which strives to eliminate quantitative bias from model prediction. A brief overview of these efforts ranges from foundational approaches that seek to mitigate bias by training models that remain "unaware" of protected attributes like race and gender (Hardt, Price, and Srebro 2016) to more sophisticated techniques that seek to impose fairness as a "constraint", defined by the prevalence of protected attributes, to limit undesirable correlations found already in the data (Dwork and Ilvento 2018; Romanov et al. 2019). However, in almost all cases, trade-offs are made between predictive accuracy vs. fairness in outcome, with both technical and legal scholars disagreeing on where exactly the correct balance is.

In practice, these efforts face two fundamental implementation flaws. First, any technique that relies on protected attributes for model training stands at odds with Title VII of the Civil Rights Act of 1984, which forbids the usage of protected attributes in model prediction, even if the purpose of such an approach is to mitigate bias (Stone 1990). Second, current state-of-the-art models that claim to be *de-biased* without relying on protected attributes have also been challenged as merely disguising the bias (Gonen and Goldberg 2019). Third, irrespective of the techniques developed, algorithmic fairness efforts rely on the assumption that an ideal world state exists for which fairness as a metric can be technically defined and optimized for—an assumption that is grounded in normative, not technical, evaluation. Even within legal frameworks describing the role of incarceration, defining fairness remains a perennial "criminological puzzle" for academics and practitioners alike, one that system designers are not immune from.

Towards a Normative Framework for Fair Judicial Decision-Making

Fairness in criminology hinges on fundamentally subjective judgements—it is at once a study of psychology, power, sociology, ethics and politics. Appraising the criminal justice system and its manifestations—whether it be policing, pre-sentencing, sentencing, prison conditions, or parole and probation—requires consideration of endless moral quagmires (Bonilla-Silva 2006): What is it that people deserve? What are the limits of human agency? How best do we hold people accountable? Why are certain demographics chronically over-represented in punitive systems? Is the punitive paradigm inevitable? The process of punishment uniquely enlists constant engagement with these questions and necessitates the incorporation of human values into its analysis.

In light of the challenges and limitations expressed in algorithmic system design, we outline a normative approach to studying the theoretical contours of fair judicial decision-making by:

1. Paying heed to broad philosophical frameworks of fairness
2. Understanding the role of legitimacy and procedural justice in criminology
3. Incorporating the value of human *discretionary bias*, not its elimination, from decision-making processes

Philosophical Frameworks of Justice

The terms which underwrite just outcomes in the justice system are constantly revised and often competing—experts in legal and criminological fields have delineated the aims of the criminal justice system broadly, but the actualisation of such aims is actively contentious (Duff 2003). Governments, as well as legal theorists, have identified five traditional goals of punishment: retribution, deterrence, rehabilitation, restoration, and incapacitation (Cole, Smith, and DeJong 2017). The US Sentencing Commission along with the UK's Criminal Justice Act (CJA) of 2003 have codified these five philosophical goals into law. Described as a "smorgasbord" approach of sorts (von Hirsch and Roberts 2004), codes like the CJA notably neglect how frameworks of punishment often compete and contradict each other, further ambiguating the fundamental rationale of punishment for practitioners to use.

Retribution or *desert* has become a centerpiece of academic and legal reasoning for exacting punishment in criminology. Such a framework relies on a ”proportional” response, or the notion that individuals ought be punished in accordance with what they deserve (von Hirsch 1976). A proportional system is not only commensurate with the gravity of harm done, but also mandates that punishments be individualized in a manner that reflects the culpability and circumstances of the individual involved (von Hirsch and Ashworth 2015). While these *just-deserts* principles underpin the dominant academic doctrine on sentencing, the role of public protection and risk management as a *deterrence* strategy has also increasingly become a legislative and political priority in modern society (Garland 2002). In the last several decades, risk assessment has come to define probation practices as well as sentencing more broadly—to the dismay of those who think individuals should be treated in accordance to their individual culpability rather than perceived risk.

Moreover, lawmakers have also questioned whether imprisonment effectively redresses the social ills of society at all. Consideration of ”problem-solving” justice solutions, which involve highly individualized treatment of defendants with the ultimate goal of *rehabilitation*, has risen in legal adoption (Mathiesen 1998). Ergo, many judges have been granted extraordinary discretion in regards to sentencing decisions through drug and mental health courts (Cole, Smith, and DeJong 2017). While the implementation of such an approach has led to disparate outcomes in sentencing between minority groups, the philosophical foundation of *rehabilitation* remains one that is broadly accepted today (United States Sentencing Commission 2018).

Others legal theorists have advocated that the *punitive paradigm* itself is fraught, and that prisons and policing as we know them today are not inevitable (McLeod 2015; Davis 2011). Such an *”abolitionist ethic”* nullifies the need for risk assessment models and instead values gradual decarceration and the substitution of prison with social services and alternative forms of accountability (McLeod 2015; Garland 2002). Future reforms should view the past and current legacy of criminal justice as one of domination and instead seek sophistication in how to balance many normative frameworks of justice systems that can run counter or parallel to each other.

Perceptions of Legitimacy and Fairness

In addition to the inevitable philosophical quandaries faced when selecting an appropriately “moral” conception of fairness, the processes that themselves govern the outcomes of justice systems must also be treated with care. While fair outcomes are certainly important to assessing fair decision-making, the *perception* of individual fairness may be equally valuable and is more often tied to factors associated with implementation of a *legitimate* decision-making process that is *procedurally fair* rather than the final outcome itself.

Theories of legitimacy assert that individuals abide by, and trust decision-makers in legal proceedings because they believe in its normative value and structural implementation (von Hirsch and Ashworth 2015). Others have theorized that perceptions of procedural fairness rely deeply on a process that is transparent, dynamic, respectful to all parties and involves dialogue between both stakeholders and powerholders (Garland 2002). For those already estranged by legal systems, algorithms do little to mend breaches of trust in this process by being frustratingly opaque as well as fundamentally impervious to dialogue (Chouldechova 2016).

Tom Tyler of Yale Law School cites several factors as potentially influencing an individual’s perception of justice: (1) voice (that one’s side of the story has been heard); (2) respect (that the system treats everyone with dignity and respect); (3) neutrality (that the process is trustworthy); (4) understanding (that transparency exists in how decisions are made); and (5) helpfulness (that system players are interested in one’s personal situation) (Tyler 2006). The balancing of these interacting, and often mutually antagonistic, factors requires the decision-maker to know as much as they can about the individual. This demand alone cannot be met by algorithms, and yet many prominent legal theorists have argued that even this conception of procedural justice is too narrow, asserting that notions of legitimacy need be fundamentally *dialogical* between powerholders and the audience of that power (Bottoms and Tankebe 2013). This conferencing in the justice system—between police, judges, prison officers, and community members—is one that is impossible to achieve through algorithmic means alone.

The Moral Role of Human Discretionary Bias

There is a tremendous moral cost to neglecting certain characteristics of the individual. Risk assessment tools that strive for identity-neutrality often ignore features that are crucial sources of information both within the realm of the penal system and of the individual and her culpability (Chouldechova 2016). The distinction between discretion and bias is tenuous—the lack of bias does not coincide with justice; in fact, it can often foster the opposite. Though judicial discretion is certainly predicated on human biases, some of which are bigoted, the removal of human input generates an anemic and abstract sense of the individual. In the words of Hegel, simplification of how we consider criminals is to ”annul all other human essence in him with this simple quality [of criminality]” (Hegel 1808).

There are various forms of discrimination that can take place in a courtroom, not all of which are undesirable. Principled frameworks of judicial decision-making consistently indicate that while there are constraints and consistencies in the way decision-making occurs, adjudications of a ”fair” sentence or bail often pull from variegated sources that may or may not be relevant, on paper, to a case. In her extensive research interviewing judges in the United Kingdom, Joanna Shapland noted 876 different potential factors of mitigation mentioned in speeches made by the defense (Shapland 2015). When such considerations for moral evaluation are innumerate, decisions need be influenced by intuition (Spohn 2008). Features related to race, gender, and class may appear to be objectively quantifiable at first, but often manifest and compound in nuanced intricacies that are difficult to untangle later. While these individual characteristics should contribute a requisite complexity to the moral evaluation of criminality and ergo, fair decision making, they should never be penalized for the purpose of additional punitiveness, as algorithmic risk assessment tools largely do today (Northpointe 2012). Human decision-makers are uniquely positioned to grapple with the multifarious considerations and traits which may or may not be relevant on an individualized basis in criminal justice.

Ethical Irreconcilability

Implementing the framework we propose, we see that fair principles of crime and punishment can be excavated exclusively from the realm of human deliberation and ethicism. Machine learning deployments in criminal justice are subject to the highest of scrutiny. First, the stakes associated with prediction errors could not be higher, and the implementation of inaccurate models amplifies such harms for millions of stakeholders (O'Neil 2016). Second, and perhaps more importantly, risk assessment tools that claim to promote fairness are guaranteed to be inaccurate, because there is no technically accurate or objective rendering of fairness by which ground truth can be optimized for. While we recognize the importance of *principled decision making*, *procedural fairness*, and *individualization* as being critical to fair judicial decision-making, there exists no algorithmic technique to incorporate those aspects into an individual's moral worth and their individualized evaluation in the courtroom. With such matters, replacing human discretion with an automated program only calcifies an already problematic and simplified understanding of the goals of justice.

This new wave of algorithmic risk assessment efforts represents simply yet another reiteration of *social incapacitation theory*, one that unwisely undermines the spirit of individual liberty. More importantly, current efforts at algorithmically addressing fairness miss the point—machine learning systems that use the past to predict the future represent a fundamental threat to the legal right to presupposition of innocence. Truly fair judicial decision-making relies on a whole host of individual factors, including many features that cannot be measured by an algorithm.

We cannot morally quantify "an ideal state of the world." As such, the pursuit of fairness through algorithmic means is fundamentally at odds with the unanswerable normative questions of the criminal justice system. False positives and negatives become obsolete in the adjudication of fair practices, where even the terms that define "accuracy" become morally fraught. Algorithmic risk assessment tools are ill-equipped to mediate the competing philosophical aims of a system where the complexities of individuals and their identities compound, and where human discretionary bias is helpful. Algorithms can provide information that may assist judges in information distillation and evidence gathering, but they critically ignore that fair decision-making arises from a complex and inconsistent scaffolding of individual and general factors.

A Case Study of Social Disadvantage

Perhaps the most enduring trend of incarceration is that it is not cross-sectional: the prison has historically been a congregating space for certain demographics of populations—namely the most disadvantaged. Despite such glaring inequity, there has been historic neglect in the legal system and now the computer science community of the realities of *social disadvantage*, and its factors related to political disempowerment, civil disenfranchisement, physical and mental disability, and psychological oppression. This is in part because such phenomena cannot be quantified by objective measures (Clear 2009).

Even if risk assessment tools could account for social disadvantage in a way that captures all the relevant features necessary for evaluation, standardizing such an analysis would prove impossible. There are three primary features of social disadvantage which consistently reappear in criminological and sociological literature: 1) relation to power, 2) multidimensionality, and 3) often a result of institutional structures (Wacquant 2009). As highlighted in the section discussing legitimacy, disadvantage has been understood as a phenomenon that extends beyond purely individualized features. Socio-structural concerns have been as, if not more, important for scholars of disadvantage than individual or behavioral concerns. The deprivation of socioeconomic status and political freedom limits individuals' choices, potentially leading to psychological and social pressures that affect decision-making (Thaler and Sunstein 2008). This damage is further amplified by the larger subtleties of economic exclusion, which perpetuate low-wage labour, unemployment and educational exclusion (Wacquant 2009). Similarly, though hate crimes and slurs plainly convey racism, one need not experience such explicit treatment to be a subject of institutional racism. Such a nuanced analysis of these compounding factors requires a sophisticated human moral evaluation to understand the compounding impacts of state and social power, one that algorithms are ill-equipped to handle.

The limitations of risk assessment tools like COMPAS come into further focus when a social, and therefore subjective, phenomenon such as disadvantage is considered in the scope of risk mitigation. When implemented into judicial decision-making processes, risk assessment tools promote certain contested ideologies—ones that have historically led to the over-incarceration of historically marginalized communities (Hudson 2002). The use of such models reinforces that "risk" prevention is a proper framework to exacting punitive justice and should be weighted over values like deservingness and opportunity (Clear 2009). We see this in historic risk calculations, where social disadvantage often overlaps with higher rates of risk and re-offending. For this reason, dynamic and static risk factors that coincide with adverse personal situations often lead to increased time of incarceration, decreased chance of parole, or higher bail suggestion (Boswell, Davies, and Wright 1993). By utilizing such a framework for risk mitigation that ignores the relative impacts of features like social disadvantage, risk assessment tools obstruct the principle that an individual should be evaluated in accordance with the harms they have caused and their individual blameworthiness—and not for "prospective" crime.

Moreover, quantitative attempts at studying social advantage have resulted in questionable criminological work. Social disadvantage should be understood as a phenomenon *emergent* from the fabric of power and its racial, classist and sexist threads (Boswell, Davies, and Wright 1993). Quantitative accounts of disadvantage, as those seeking reforms through algorithmic fairness efforts also hope to incorporate, inevitably erode a normative framework's ability to understand and address these nuances, often due to an overweighting of a predefined but unrepresentative set of features. Criminological research's attempts to disambiguate social advantage have reflected this. For example, a recent attempt at comprehensively quantifying social disadvantage took great care to define all aspects related to family and neighborhood such as household income, parents' education level, and neighborhood mortgage price (Wikström and Treiber 2016). However, such a focused analysis invariably left out simple features of interest such as gender, race, and sexual orientation—ultimately resulting in a fundamentally flawed conception of disadvantage. Risk assessment techniques suffer from the same reasoning errors—there is at once a multi-directional, and often antagonistic, balancing act of feature weighting that must apply to individualized cases. Algorithms do poorly on such tasks—human discretion is needed.

Recommendations for Reform

While we criticize both the current deployment of algorithmic risk assessment tools and algorithmic fairness efforts towards addressing these concerns, this paper does not ultimately argue that there is *no role whatsoever* for machine learning and algorithmic risk assessment techniques in the criminal justice system. Machine learning systems may indeed be helpful in parallel processes such as aiding investigators in gathering forensic evidence, detecting patterns of sophisticated misconduct such as drug smuggling and human trafficking, and analyzing an over-abundance of messy data. However, the deployment of risk assessment tools for predicting the propensity of an individual to predict future crimes or the future ambition to model "fair sentences" represents a requisite human evaluation of moral character: one that violates our normative framework for fair judicial decision-making. We therefore recommend that an over-reliance on algorithmic efforts to impart fair and impartial predictions of future crime be weaned from existing judicial decision-making processes.

Moreover, this paper highlights how the theoretical foundations and goals of punishment are often taken for granted in contemporary penal systems as well as in algorithmic systems. So often we skip to the numbers without considering that risk assessment tools, "dangerousness" metrics, and other quantifications of human moral worth are ideologically encoded enterprises that often masquerade as objectivities (Burchardt and Hick 2017). Appraising fairness in full demands consideration of the complexities of individual desert, the drivers of crime, the ontology of moral judgement, and steadfast critique from all stakeholders. Therefore, we must not surmise a panacea conception of fairness; instead, we would do well to heed the advice of scholars like Foucault before us and default to philosophical humility, admitting that "justice must always question itself." A normative approach to integrating these philosophical principles represents a good start.

Regardless of the path forward that the American criminal justice system chooses to take with respect to how it wishes to utilize risk assessment tools in the sentencing process, it is clear that cross-pollination across technical and legal fields must occur. The technical research on bias in machine learning and AI algorithms is still in its infancy. Questions of bias and systemic errors in algorithms demand a different kind of wisdom from algorithm designers and data scientists. These practitioners are often engineers and scientists with less-than-ideal exposure to legal or policy processes. The demographics of algorithm designers are also less than diverse. Algorithmic transparency requires a more educated public capable of understanding algorithms. Diversity in the ranks of algorithm developers and technical education of the general public could help improve sensitivity to potential disparate impact problems.

Conclusion

In this paper, we set out to prove the irreconcilability of current algorithmic risk assessment in criminal justice by using a normative framework for assessing fair judicial decision-making. First, we detailed how algorithmic risk assessment tools are counteractive to fair legal proceedings in social institutions where desired states of the world are contested ethically and practically. We then provided a normative framework for fair decision-making assessment—one that included the sophistication of human *discretionary bias*, not its elimination, from judicial decision-making processes and the consideration of complex social phenomenon. Finally, we used this framework to consider the case study of social disadvantage in the United States and provided an assessment of crucial features of interest in the search for fairness, such as political disempowerment and demographic exclusion—phenomenon that could not be captured by machine optimization functions.

The running trend of algorithmic risk deployment accentuates the most damaging of failures in our criminal justice system: ones where the punitive enterprises of the state have come to inordinately prioritize values of risk and deterrence over desert and rehabilitation. Over the past several decades, legislators, police departments, prosecutors, and judges have been directed to incarcerate and punish more, without care to the livelihoods and communities destroyed in their wake (McLeod 2015). The cultural and economic cleavages of our society often *matter the most* when advancing social institutions, and they cannot and should not be encapsulated in their full by data. Though human judgement will always be subject to reasoning flaws, the way to confront such social phenomenon is not to lean into feckless "objectivity", but instead strive for the sophistication of how we consider fairness—a process that enlists human *discretionary bias*. Algorithmic risk assessment must recognize that when it comes to the moral quandaries underwriting criminal justice systems, there is no objective function to achieve. If there was, its discovery would constitute perhaps the final frontier of computer science: a frontier now only assailable from the fabric of humanity, however uncertain and unreliable.

References

Alexander, M. 2012. *The new Jim Crow: mass incarceration in the age of colorblindness*. The New Press.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine bias: there's software used across the country to predict future criminals, and it's biased against blacks. *ProPublica*.

Barocas, S., and Selbst, A. 2016. Big data's disparate impact. *California Law Review* 671.

Bolukbasi, T.; Chang, K.-W.; Zou, J.; Saligrama, V.; and Kalai, A. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 2016 Conference on Neural Information Processing Systems (NIPS 2016)*. NIPS.

Bonilla-Silva, E. 2006. *Racism without racists: Color-blind racism and the persistence of racial inequality in the United States*. Rowman and Littlefield Publishers.

Boswell, G.; Davies, M.; and Wright, A. 1993. *Contemporary probation practice*. Avebury.

Bottoms, A., and Tankebe, J. 2013. Beyond procedural justice: a dialogic approach to legitimacy in criminal justice. *Journal of Criminal Law and Criminology*.

Burchardt, T., and Hick, R. 2017. Inequality, advantage, and the capability approach. *Journal of Human Development and Capabilities*.

Chouldechova, A. 2016. Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. In *Proceedings of the 2016 Conference on Fairness, Accountability, and Transparency (FAT* 2016)*. ACM.

Clear, T. 2009. *Imprisoning communities: how mass incarceration makes disadvantaged neighborhoods worse*. Oxford University Press.

Cohen, J. 1983. Incapacitation as a strategy for crime control: possibilities and pitfalls. *Crime and Justice*.

Cole, G.; Smith, C.; and DeJong, C. 2017. *Criminal Justice in America*. Cengage Learning.

Davis, A. 2011. *Are prisons obsolete?* Seven Stories Press.

Desmarais, S., and Singh, J. 2013. Instruments for assessing recidivism risk: a review of validation studies conducted in the u.s. Technical report, Council of State Governments Justice Center.

Diakopoulos, N. 2016. We need to know the algorithms the government uses to make important decisions about us. *The Conversation*.

Duff, A. 2003. *Punishment, Communication, and Community*. Oxford University Press.

Dwork, C., and Ilvento, C. 2018. Group fairness under composition. In *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency (FAT* 2018)*. ACM.

Garland, D. 2002. *The culture of control: crime and social order in contemporary society*. University of Chicago Press.

Gonen, H., and Goldberg, Y. 2019. Lipstick on a pig: debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2019)*. ACM.

Green, B., and Chen, Y. 2019. Disparate interactions: an algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (FAT* 2019)*. ACM.

Greenwood, P., and Abrahamse, A. 1982. Selective incapacitation. Technical report, RAND Corporation.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. In *Proceedings of the 2016 Conference on Neural Information Processing Systems (NeurIPS 2016)*. NeurIPS.

Harvard Law Review, H. 1982. Selective incapacitation: reducing crime through predictions of recidivism. *The Harvard Law Review Association*.

Hegel, G. W. F. 1808. Who thinks abstractly? *Hegel Texts and Commentary*.

Hudson, B. 2002. *Justice in the risk society: challenging and reaffirming justice in late modernity*. SAGE Publications.

Hvistendahl, M. 2016. Can 'predictive policing' prevent crime before it happens? *Science Magazine*.

Kleinberg, J.; Ludwig, J.; Mullainathan, S.; and Sunstein, C. R. 2019. Discrimination in the age of algorithms. *SSRN*.

Lewin, T. 1982. Making punishment fit future crimes. *The New York Times*.

Mathiesen, T. 1998. Selective incapacitation revisited. *Law and Human Behavior*.

McLeod, A. 2015. Prison abolition and grounded justice.

Northpointe. 2012. Practitioners guide to compas. Technical report, Northpointe.

O'Neil, C. 2016. *Weapons of math destruction: how big data increases inequality and threatens democracy*. Crown Books.

Romanov, A.; De-Arteaga, M.; Wallach, H.; Chayes, J.; Borgs, C.; Chouldechova, A.; Geyik, S.; Kenthapadi, K.; Rumshisky, A.; and Kalai, A. 2019. What's in a name? reducing bias in bios without access to protected attributes. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2019)*. NAACL.

Scott, P. 1977. Assessing dangerousness in criminals. *British Journal of Psychiatry*.

Shapland, J. 2015. *Between conviction and sentence: process of mitigation*. Routledge and Kegan Paul Books.

Spohn, C. 2008. *How do judges decide? The search for fairness and justice in punishment*. SAGE Publications.

Stone, R. 1990. *The Civil Rights Act of 1984: overview*. Public Law.

Thaler, R., and Sunstein, C. 2008. *Nudge: improving decisions about health, wealth, and happiness*. Penguin Books.

The Economist, T. 2015. The moral failures of america's prison-industrial complex. *The Economist*.

Tyler, T. 2006. *Why people obey the law*. Princeton University Press.

United States Sentencing Commission, U. 2018. Demographic differences in sentencing. Technical report, Federal Sentencing Reporter.

von Hirsch, A., and Ashworth, A. 2015. *Proportionate sentencing: exploring the principles*. Oxford University Press.

von Hirsch, A., and Roberts, J. 2004. Legislating sentencing principles: the provisions of the criminal justice act of 2003 relating to sentencing purposes and the role of previous convictions. *Criminal Law Review*.

von Hirsch, A. 1976. Doing justice: the choice of punishments. *National Criminal Justice Reference Service*.

Wacquant, L. 2009. *Punishing the poor: the neoliberal government of social insecurity*. Duke University Press.

Wikström, P.-O., and Treiber, K. 2016. Social disadvantage and crime: a criminological puzzle. *American Behavioral Scientist*.

Wright, R. A. 1994. *In defense of prisons*. Greenwood Press.