Uninventing Carceral Technology: Four Experiments in Imagining the World More Rigorously

by

Chelsea Marie Barabas

B.A., Stanford University (2009)
S.M., Massachusetts Institute of Technology (2015)

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning, in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Media Arts and Sciences

at the
Massachusetts Institute of Technology
September 2023

Authored by: Chelsea Barabas
Program in Media Arts and Sciences
August 8, 2023

Certified by: Justin Steil
Associate Professor of Law and Urban Planning
Thesis supervisor

Accepted by: Joe Paradiso
Academic Head, Program in Media Arts and Sciences

Uninventing Carceral Technology: Four Experiments in Imagining the World More Rigorously

by

Chelsea Barabas

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning, on August 8, 2023 in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Abstract:
How do we advance social justice in an increasingly datafied world? Against a backdrop of burgeoning social movements, data-driven technologies have become an important terrain of struggle. That's because the design and implementation of technology is not simply about the creation of software and hardware objects, but the negotiation of practices and possibilities for living life together differently (Suchman 2007). In this dissertation, I examine the role of technology in shaping our collective imaginations of what is possible in the context of the carceral state. By carceral state, I mean the expansive system of state-sanctioned capture, confinement, and control that underpins our current unjust social order. Drawing on rich intellectual traditions such as feminist, Indigenous, and Black studies, I interrogate the default assumptions underlying the design and implementation of data-intensive systems, in order to fundamentally reimagine the role of technology in larger struggles for justice. Ultimately, the aim of this work is to experiment with ways of "un-inventing" (MacKenzie 1993) carceral technology, by reconfiguring the structural, interpersonal, and personal aspects of computation to the point that harmful algorithms are no longer created.

Thesis advisor: Justin Steil
Associate Professor of Law and Urban Planning

Uninventing Carceral Technology: Four Experiments in Imagining the World More Rigorously

by

Chelsea Barabas

This dissertation has been reviewed and approved by the following committee members

Justin Steil
Associate Professor of Law and Urban Planning
Massachusetts Institute of Technology

Roderic Crooks
Assistant Professor in the Department of Informatics
University of California, Irvine

Danielle. Woods
Assistant Professor in the Program in Media Arts and Sciences
Massachusetts Institute of Technology

Ethan Zuckerman
Associate Professor of Public Policy, Communication, and Information
University of Massachusetts at Amherst

Acknowledgements

This dissertation would not have been possible without the guidance, friendship, and support of so many people. I'd like to thank my advisor, Justin Steil, for not only agreeing to take me on as a student during an incredibly uncertain and tumultuous time, but for also nurturing me as a scholar by creating a thoughtful and caring environment for me to think through half-formed ideas. Justin, you restore my faith in the academy as a place where kindness, generosity, and rigorous intellectual debate can live side by side. I also want to express my deep gratitude to the other members of my committee: Roderic Crooks, Danielle Wood, and Ethan Zuckerman. I could not have asked for a more engaged and thoughtful group of readers for my dissertation. Thank you for taking my work seriously and for being such attentive mentors throughout this process.

Over the past six years, I have had the privilege of working with an incredible set of collaborators and thought partners, including Rebekah Agwunobi, Audrey Beard, Shreya Chowdhury, Chinelo Dike, Theodora Dryer, NM Amadeo, Colin Doyle, Karthik Dinakar, Nasser Eledroos, Sarah Hamid, Chris Gilliard, Erhardt Graeff, Os Keyes, James Kilgore, Puck Lo, Rodrigo Ochigame, Beth Semel, Sonja Solomun, Sameeha Rizvi, Pilar Weiss, Madars Virza, Atara Rich-Shea, JB Rubinobvitz, and Jonathan Zong. Thanks so much to each of you for sharing your minds and hearts with me as we try to make sense of this messy world together.

I also could not have completed this PhD without the very practical guidance and friendship of other MAS students, especially Alexis Hope, Ufuoma Ovienmhada, Anastasia Ostrowski, and Pedro Reynolds-Cuellar. Thanks for your friendship and solidarity during the tough times, as well as the numerous small ways (i.e. sharing your reading lists, brainstorming proposal ideas, etc.) that you helped me to achieve each major milestone of the PhD process.

I also want to thank the senior scholars who gave my work their time and careful consideration at various stages of the writing process. Thanks especially to Martha Minow, Lucy Suchman, and Anna Lauren Hoffmann for your generous engagement with my work.

Finally, I'd like to thank my friends and family for their never-ending support and love throughout this process. I'd especially like to thank my parents, Bill Barabas and Lyn Few, who have always been my greatest role models and biggest fans. And thanks to my brilliant and kind partner, John Kidenda. Johnny, you make this life such a joy to live. Thanks for being by my side at every step of this journey.

Table of Contents

Chapter 1

Imagining Otherwise: Computation During Times of Crisis

Introduction

How do we advance social justice in an increasingly data-fied world? Government officials have embraced data-intensive systems as a means of addressing demands for social change across a number of high-stakes realms. Yet, a growing body of scholarship and activist thought reveals the ways that such technologies play an active role in reproducing our unequal and racialized social order (Benjamin 2019a; Eubanks 2018; Stop LAPD Spying Coalition 2020; Dixon-Román, Nichols, and Nyame-Mensah 2020; Community Justice Exchange 2022b). Against a backdrop of burgeoning social movements, data-driven technologies have become an important terrain of struggle. That's because the design and implementation of technology is not simply about the creation of software and hardware objects, but the negotiation of practices and possibilities for living life together differently (Suchman 2007). Data-driven technologies are increasingly the target of analysis and intervention by social movement actors, because they embody and make tangible the often invisible aspects and contradictions of our shared social imaginaries (Jasanoff and Kim 2015).

In this dissertation, I consider the imagination as an important site of political intervention, by interrogating the role of technology in shaping our collective imaginations of what is possible. I do this by examining the role that data-driven technologies play in expanding and reasserting the legitimacy of the U.S. racial order through one of its foundational institutions — the carceral state. By carceral state, I mean the expansive system of state-sanctioned capture, confinement, and control that underpins our current unjust social order. The carceral state is a fundamentally relational phenomenon that extends far beyond brick and mortar prison walls —

as Ruth Wilson Gilmore (2007, 242) explains, “prison is not a building ‘over there’ but a set of relationships that undermine rather than stabilize everyday lives everywhere.” These relationships are grounded in punitive systems of control that undermine the capacity for people to thrive and protect one another, while simultaneously buttressing existing relations of power and domination.

By anchoring my work in an expansive understanding of the carceral state, I aim to build a nuanced analysis of the context in which harmful technologies are built and implemented, as well as create frameworks for understanding how such technologies can be refused and fundamentally re-imagined. This dissertation contributes to a growing body of scholarship that pushes beyond technical and managerialist considerations of how predictive models are developed, in order to interrogate the deeper normative and epistemological assumptions of these tools.

By drawing on rich intellectual traditions such as feminist, Indigenous, and Black studies, we can interrogate the default assumptions underlying the design and implementation of data-intensive systems, so that we can “begin to count it all out differently” (McKittrick 2014). Ultimately, the aim of this work is to experiment with ways of “un-inventing” (MacKenzie 1993; Mackenzie 1996) carceral technology, by reconfiguring the structural, interpersonal, and personal aspects of computation to the point that harmful algorithms are no longer created (Keyes and Austin 2022). This requires the development of a robust critical vocabulary for naming and responding to the potential harms of data-driven regimes, as well as the development of practices that increase our collective capacity to envision and enact more just futures through and beyond technology.

There are three key research questions guiding this work, which I will explore at greater length in the following section.

1. What are the mechanisms through which data-driven technologies and research are used as a means of entrenching the carceral state during times of crisis?

2. What is the critical vocabulary that social movement leaders and critical scholars have identified for naming and responding to the harms that arise from data-driven tech?

3. In what ways do computational practitioners forge solidarities with transformative social movements in order to engage with data regimes in ways that posit radical alternatives to the status quo?

Where do we begin…

This dissertation was completed against a backdrop of multiple, interlocking crises. I started my doctoral program in the fall of 2019, just a few months before the COVID-19 pandemic brought the world to a standstill. It was a weird time to be doing a PhD, especially for someone whose research is usually out "in the field." Prior to enrolling in my program, I'd spent two years researching the contentious role of actuarial risk assessment in struggles to reduce U.S. jail populations. That work had involved many hours of convening with local community advocates, observing courtroom proceedings, and meeting with government officials to understand the ways that data-driven reforms like risk assessment had become a battleground for social change.

I'd intended to continue this kind of hands-on research in the spring of 2020 as I outlined my dissertation project. There was a retreat in Philadelphia called "Organizing, Power, Justice and Algorithms," which I planned to attend in March. The gathering was an opportunity for community organizers and scholars to come together to map the broader landscape of how algorithms were shaping the lives of marginalized and vulnerable groups. My hope was that the

convening would serve as a springboard into my next research project, a place to identify questions that were relevant and interesting, as well as connect with potential collaborators. But as the date drew near, it became clear that such a gathering was no longer possible. Airports were shutting down. Schools had been closed. People were advised to shelter in place. So I sat at home in my sweatpants, scrolling through my newsfeed as I tried to wrap my head around our new reality — trying to survive a deadly pandemic in a deeply unequal world.

At first, it seemed like the virus might serve as a great unifying force, bringing together people from all walks of life in a shared struggle for survival. But it soon became apparent that some people bore the brunt of the pandemic more than others. Low-income "essential" workers, such as public transit operators and grocery store clerks, were dying in large numbers across the country (Chen et al. 2022). Alarming statistics were released, showing that Black, Indigenous, and Latinx people were getting sicker and dying at much higher rates than their White counterparts (Pilkington 2020). Competing narratives began to emerge to make sense of such disparities.

For example, during a White House press briefing in April 2020, U.S. Surgeon General Jerome Adams advised African Americans and Latinx people to avoid using alcohol and other drugs, explaining that such factors might be the reason why they were experiencing higher death rates (Perez 2020). Public health advocates were quick to challenge this assertion, arguing that there was a wide range of other social factors — such as high rates of unemployment, mass incarceration, substandard and overcrowded housing, homelessness and lack of access to quality health care — which contributed to the increased vulnerability of low-income communities of color (Khazanchi, Evans, and Marcelin 2020). They argued that it was America's pathology of racism that was making the pathology of the coronavirus so much worse for these groups, not

their supposed bad habits (Patton 2020). The pandemic shed light on the ways that people had been living in various states of crisis for a long time. Yet those protracted crises were all too often ignored or explained away by the powers that be, through narratives that recast vulnerability as pathology and bad behavior (Benjamin 2020).

Amid these debates, my research on the role of technology in struggles for decarceration took on new urgency. The bloated U.S. penal system had become a major vector for the spread of the virus, both behind and beyond prison walls. For months, detention facilities were at the top of national lists for the largest outbreaks around the country, with transmission rates in prisons and jails skyrocketing to more than four times the rate of the general population (Ciaramella 2020; Sims, Foltz, and Skidmore 2021). These trends had significant spillover effects into mainstream society, in both urban and rural areas (American Civil Liberties Union 2020; Chappell 2020; Hooks 2020). Some researchers estimate that mass incarceration contributed more than a half million additional cases in the United States over just three months, which helped to explain why the country had one of the highest reported infection rates in the world (Hooks and Sawyer 2020).

As transmission and death rates rose, incarcerated people organized for their survival at unprecedented levels. Between March and June 2020, people detained in prisons, jails, juvenile facilities, and immigration detention centers organized over one hundred acts of resistance related to COVID-19. This upwelling of collective action represented "one of the most massive waves of prisoner resistance in the past decade" (Perilous Chronicle 2020). Incarcerated people also coordinated with loved ones on the outside to bring attention to the unsanitary living conditions and inadequate medical care that they were experiencing behind bars. In response to these collective efforts, a wide range of community advocates, policy makers, and public health

officials rallied together to call for significant decreases in prison and jail populations (United Nations 2020; Williams et al. 2020; Williams and Bertsch 2020).

At the same time, abolitionist social movements were gaining unprecedented traction in U.S. political discourse, as more people fundamentally challenged racist policing tactics, state surveillance, and the use of lethal force by law enforcement. The murders of George Floyd, Breonna Taylor and Ahmaud Arbery at the hands of White vigilantes and police officers revealed the ways that Black people were caught in a double-bind during the pandemic – trying to minimize the risks of COVID-19 while also navigating the everyday racism of a society whose foundational sickness is rooted in anti-Black violence (Benjamin 2022; Howard 2022). Against this backdrop, high-profile social movements such as the Movement for Black Lives explicitly aligned their cause with an abolitionist theory of change, foregrounding the need for lasting alternatives to punishment, while also pushing for the removal of tools which perpetuate violence against marginalized communities (McLeod 2018; Khan-Cullors and Bandele 2017). A growing number of state and local advocacy groups also called for the wholesale elimination of the punitive practices which fuel mass incarceration, such as the use of cash bail (Courtwatch MA 2020) and the presence of law enforcement in schools (Chavez 2020). Amid the breakdown of life-as-usual, people began to envision and experiment with new ways of living. The pandemic was reframed as an opportunity to imagine the world differently, in order to make it more livable for everyone long after the virus had run its course (Arundhati Roy 2020).

As I hunkered down in my Boston apartment, I began to read more about the role of the imagination in abolitionist struggles for a more just society. Abolition is both a positive and a negative project. As Ruth Wilson Gilmore (2018) explains, "[A]bolition isn't just absence…[it] is a fleshy and material presence of social life lived differently... It's about making things."

Abolition is about building our collective capacity to live our way into a world where everyone can thrive, while simultaneously undoing the institutions and policies which perpetuate violence and organized abandonment (Gabriel 2022). This often requires pushing past commonsense notions of what is possible or practical, in order to imagine how life might be lived otherwise.

As Gumbs (2013, 13) argues, such work hinges on our ability to imagine the world more rigorously, in order to identify "those patterns in our own lives that do not need to be tweaked or revised, but actually need to be abandoned completely… in order to earn the world that we deserve, which is beyond what all of us who have lived our lives in a police state can even imagine." Abolitionist world-making requires an immense amount of creativity and experimentation. Such work hinges on our ability to rework the default settings that structure society, so that we can expand our vision of what's possible (Benjamin 2016a, 1). As such, we might understand the cultivation of a rigorous imagination as an essential technology of abolitionist work.

During the early days of the pandemic, I began to fully consider the imagination as a site of political struggle, as well as the role of technology in shaping our collective imaginations of what is possible. Amid a surge in collective action in 2020, government officials took steps to restore the legitimacy of carceral policies and diffuse the momentum behind calls to decarcerate prisons and jails. Data-driven technology played a key role in this effort. For example, in March 2020 Attorney General Barr released a memo, outlining the steps penal officials should take to minimize the spread of the virus in federal detention facilities. In the memo, Barr emphasized the need to balance the well-being of inmates with the Bureau of Prison's mandate to keep communities safe, arguing that the release of large numbers of incarcerated people would result in "profound risks to the public," such as "violence or heinous sex acts" (Barr 2020b) He

directed penal officials to use an actuarial risk assessment called PATTERN to determine which federal inmates could be safely transferred to home confinement — only those with a "minimum" risk score would be given priority for release (Barr 2020a).

The memo immediately sounded alarm bells among community advocates, who were already familiar with PATTERN's limitations. The tool had been developed prior to the pandemic, as part of the First Step Act, a bipartisan bill that was framed as a "first step" towards criminal justice reform amid widespread demands to roll back the punitive policies that drive mass incarceration and racial disparities in the criminal legal system. At the time of its adoption, proponents of PATTERN claimed that the tool would serve as an efficient means of debiasing decision-making processes, while simultaneously realigning prison and courtroom operations with the rehabilitative mission of the justice system (Office of the Attorney General 2019). According to the Attorney General, PATTERN was designed to provide a merit-based pathway to early release — incarcerated people could reduce their risk scores if they participated in rehabilitative programming, which would then increase their likelihood of having their sentences reduced. The tool was supposed to not only support pathways out of prison, but also minimize racial bias in the way release decisions were made.

Yet, upon closer inspection, it was clear that the math behind PATTERN didn't add up. For one, negative risk factors significantly outweighed all the mitigating factors on the tool, which made it virtually impossible for many incarcerated people to meaningfully reduce their risk scores.[1] These limitations were further compounded by the fact that access to rehabilitative

[1] For example, if a male is between 18-25 years old at the time of assessment, then he is automatically assigned 30 risk points. If he received his first conviction during this same period of time (18-25 years of age), then he'll receive an additional 8 points, which totals up to 38 points from just two risk categories. By contrast, the average risk reduction that one can receive for participation in rehabilitative programming ranges from 1-3 points per course,

programming was extremely limited in the federal system, with waitlists often numbering in the thousands at any given time (Bronner Group 2016). As a result, the inclusion of mitigating "rehabilitation" factors was merely tokenistic. PATTERN was a rigged game, one which stacked the cards against incarcerated people so high that it was virtually impossible for them to succeed.

In the summer of 2019, I'd co-authored a written testimony about PATTERN, calling into question its utility as a tool for supporting more equitable outcomes. For example, White people were more than four times as likely to be classified as minimum risk than their Black counterparts (Leadership Conference on Civil and Human Rights 2019). In our statement, my co-authors and I pointed out that this was due in large part to the fact that many risk factors, such as criminal history, reflected historical disparities in the way that different racial groups are targeted and treated by law enforcement and the courts. Decades of research has shown that, for the same behavior, people of color are more likely to be arrested, prosecuted, convicted, and sentenced to harsher punishments than their White peers (The Sentencing Project 2018; Demuth and Steffensmeier 2004; Rehavi and Starr 2014; Mauer 2010). This results in fundamental distortions in the data on which PATTERN was built, which then fuels the systematic inflation of Black people's perceived criminal risk.

The adoption of PATTERN was a classic example of what sociologists call "institutional decoupling," whereby organizations cope with external pressures to change by adopting rituals and technologies that appear neutral or even progressive, without necessarily changing their day-to-day practices (Meyer and Rowan 1977) . As Meyer and Rowan (1977, 340) argue, institutional rules and tools "function as myths which organizations incorporate, gaining

with the maximum number of points one can earn being -12 for taking more than ten courses. (Office of the Attorney General 2019)

legitimacy, resources, stability, and enhanced survival prospects." Amid a groundswell of abolitionist organizing, the adoption of tools like PATTERN entrenched the legitimacy of the carceral state by making it harder to detect structural racism or understand the historical context in which inequitable outcomes emerged (Spade 2015; Ray 2019). I've come to understand such tools as a form of "police science," or tools which foreground statistical objectivity in order to reframe law enforcement activities as unbiased and fair (Wang 2018). Such interventions diffuse the momentum behind collective calls for change, by giving the appearance of taking-action while maintaining the status quo.

In the spring of 2020, these concerns about PATTERN were compounded by the dire life and death circumstances experienced by people behind bars. I joined up with a group of advocates to draft a response to AG Barr's policy, arguing that PATTERN's risk predictions would leave thousands of vulnerable people exposed to the deadly virus. These concerns were exacerbated by the fact that officials secretly adjusted the thresholds used to classify incarcerated people as minimum risk (MacDougall 2020). As a result, less than two percent of the federal inmate population was eligible for home confinement (MacDougall 2020). This trend was consistent across most penal institutions during COVID-19. While many jails decreased their populations in the early months of the pandemic, most had abandoned those early reforms by the summer of 2020, which resulted in an increase in their overall populations (Widra 2020). Racial disparities in incarceration also increased significantly during this time, as White people were released more quickly and at higher rates than their Black and Latinx peers (Rivera 2020; Sawyer and Jones 2021).

The events of 2020 laid bare the extent to which penal officials were willing to keep vulnerable people behind bars, as well as the role of technology in doing so. PATTERN was

used as an administrative alibi for neglect, one which upheld a racist cognitive schema that rationalized the uneven distribution of life-saving intervention during a key moment of crisis. Lofty ideals such as "safety" and "merit" were used to smooth over this harsh reality — officials claimed that PATTERN was essential for preserving public safety and that only those who had clearly demonstrated their desire to "rehabilitate" themselves warranted consideration for release. But PATTERN was not designed to forecast the risk of violence or "heinous sex acts," as the Attorney General put it. Even before PATTERN's risk thresholds were secretly adjusted, conditioning release on PATTERN's minimum risk category meant that even people with an 89% chance of "success" would not have been prioritized for transfer (Leadership Conference on Civil and Human Rights 2019).

Moreover, risk was defined in extremely broad terms — as being arrested or having any technical violation within 3 years of release. Advocates argued that this definition of "minimum risk" offered virtually no insight into public safety[2] and made no sense when compared to the devastating risks of COVID-19. As the Leadership Conference on Civil and Human Rights (2019) argued, "Risk assessments that include minor offenses or technical violations in their definition of 'risk' will inflate risk scores and incarceration rates, and exacerbate racial inequalities. In the context of COVID-19 and this memo, this means a much higher risk of illness and of fatality."

As I watched the ways that penal officials used PATTERN during the early days of the pandemic, it became clear to me the ways that such technologies contribute to the expansion and

[2] Studies have shown that the people with the highest risk of arrest for a violent crime are not the same people with the highest risk for being arrested for any crime at all (Mayson 2017). As Mayson (2017, 562) argues, "If the goal is to avert violence, focusing on the likelihood of any future arrest targets the wrong individuals and fails to target the right ones."

adaptation of the moral economy underpinning carceral societies. With the help of PATTERN, penal officials were able to label the vast majority of people incarcerated in federal prisons as "too dangerous" for transfer to home confinement. In doing so, they reinforced the status of criminalized people as dangerous and therefore disposable, while simultaneously reviving an aura of innocence around violent social institutions. Such complementary processes of criminalization and sanitization are the essence of what the sociologist Ruha Benjamin (2019a) terms "captivating technology," or tools which appeal to our noble sensibilities while simultaneously limiting our capacity to imagine alternatives to the status quo. Such interventions frequently foreground lofty ideals such as "merit" and "safety." Yet, they ultimately entrench and expand the state's capacity to capture and control marginalized populations through the production of criminalizing narratives.

Against a backdrop of interlocking crises, such technologies often emerge as a key terrain of struggle. That's because the design and implementation of technology is not simply about the creation of software and hardware objects, but the negotiation of practices and possibilities for living life together differently (Suchman 2007). In contemporary struggles for social change, technologies like PATTERN are increasingly the targets of analysis and intervention by social movement actors, because they embody and make tangible the often invisible aspects and contradictions of our shared social imaginaries (Jasanoff and Kim 2015). Sociotechnical systems operate as performative scripts that reveal and operationalize shared visions of the social good. They can also provide a sense of coherence and continuity to social order, and are often the site of struggle during moments when the status quo is disrupted or in flux (Jasanoff and Kim, 2015).

The pandemic pushed me to grapple with the role of data-driven technologies as interventions on our collective imagination. It made me think about the ways that carceral data

regimes are used to reproduce narratives that make dominant groups feel comfortable with, or even good about, the continuation of harmful social policies. Captivating technologies are seductive, and their widespread appeal is key to understanding the ways that discriminatory and racist systems are maintained over time. As Ruha Benjamin (2020) warned in the spring of 2020, "[T]he global spread of a microscopic virus is placing the ravages of racism and inequity under the microscope. But the fact is, we don't all see the same thing! Racism has a way of actually distorting our vision."

Just as the U.S. Surgeon General tried to recast the increased vulnerability of communities of color as the by-product of their own bad behavior, penal officials used PATTERN to justify the disparate exposure of Black and Latinx people to the deadly virus behind bars. Such distorted visions of reality are key to understanding not only the enduring presence of the carceral state, but also the foundational role that penal institutions play in perpetuating our unjust social order. Numerous scholars have theorized the primary purpose of the criminal punishment system as the fabrication and maintenance of an unequal and racialized social order for the sake of wealth accumulation (Davis 2003; Gilmore 2007; Melamed 2019; Seigel 2018).

Wealth acquisition and dispossession are inextricably linked in capitalist systems, and the maintenance of a racist social imaginary is how such complementary processes of wealth extraction and accumulation are normalized and legitimated. As Ruth Wilson Gilmore (2020) explains, "capitalism requires inequality and racism enshrines it." As such, all capitalism is *racial* capitalism, in that it emerges from a mutually reinforcing relationship between racialized exploitation and capital accumulation (Robinson 2000). The carceral state is one of the primary institutions through which racist systems of wealth extraction are put into motion and sustained.

Moral panics about the alleged criminality and dangerousness of racialized Others are designed to obfuscate the system of privileges that enable White populations access to collective and cumulative forms of wealth over time.[3]

In spite of the enduring presence of such practices, racial regimes are inherently unstable, due to the fact that the people who are targeted by them are always engaged in struggles to transform the status quo and build freer futures (Kaba 2021; Kelley 2002; Keeling 2019). As a result, racial regimes are perpetually in a state of crisis. I use the term crisis in the sense defined by Stuart Hall and Bill Schwarz (1988, 96), when "the existing social formation can no longer be reproduced on the basis of the pre-existing system of social relations." When racial capitalism undergoes a crisis, the carceral state serves as an important vehicle for reasserting legitimacy of the system through processes of criminalization that mark large swaths of the population as both disposable and ripe for exploitation, while simultaneously sanitizing the violence necessary to keep such people "in their place" (Cacho 2012; Friedman 2021; Patterson 1982; Wang 2018). Such social structures require constant renovation through institutional policies and processes of collective meaning making (Lipsitz 2011; Omi and Winant 2015; Ray 2019; Robinson 2007). As a result, racism emerges as a highly innovative and future-oriented phenomenon, one that adapts to changing circumstances in order re-capture our imaginations during times when the status quo is in flux (Benjamin 2016b; 2016a; Lipsitz 2011).

[3] As Ananya Roy (2019, 228) argues, "Carcerality is not a side-show to racial capitalism; it is a necessary logic." For example, scholars have described the ways that municipal ordinances are weaponized as a pretext for evicting poor women of color from land that investors want to "re-develop" (Kurwa 2018; Ananya Roy 2019). In such situations, prior calls to the police often serve as evidence of criminal activity and are then used as a justification for the forced expulsion of working-class communities of color from their homes. Such practices constitute what Roy (2019) calls "racial banishment," a legally imposed form of spatial exclusion which depends on the criminalization of poor and racialized communities for the sake of wealth capture. This phenomenon is not limited to the United States. Scholars have similarly illustrated the ways that militarized police forces in Latin America and the Middle East operate to maintain spatial segregation and racialized forms of dispossession abroad (Bornstein 2008; Ferreira da Silva 2009; Wacquant 2008).

The events of 2020 pushed me to reckon with the limits of my own imagination in terms of how I participated in debates regarding technology and social justice. I was becoming increasingly aware of the ways that I'd been seduced by captivating technology, even as I worked to develop a critical stance in my research. For example, in 2018, I published a paper called "Interventions Over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment," in which my co-authors and I argued that the core ethical concern surrounding criminal risk assessment was not about optimizing accuracy or reducing bias. Rather, it was one of purpose. We outlined the various ways that prediction-oriented risk assessments fueled punitive practices in the courts, through the construction of criminal risk as a static state rather than as something that could change over time. We connected criminal risk forecasting to a broader set of logics which had fueled mass incarceration since the 1970's. But we didn't stop there. We then proposed that, rather than use machine learning for risk prediction, we should use it to surface co-variates that could be fed into causal inference models. Then we could try to identify the causal drivers of crime and test interventions for reducing them. We argued that such an approach could support legal practitioners in thinking through a broader set of rehabilitation strategies to lower criminal risk, rather than just incarceration.

Back in 2018, that argument seem well-reasoned and even progressive to me. But I now see how it fails to challenge the core captivating premise surrounding the use of data as a vehicle for criminal justice reform — namely, that the best way to address the injustices of such a system is through the increased measurement and management of criminalized populations. In the years since we published "Interventions Over Predictions," I've had many enlightening conversations with community organizers who helped me to see the various pitfalls of this approach. For example, members of the Massachusetts Community Bail Fund helped me to understand how

such debates distracted from larger issues regarding the ways that courtroom officials strayed from the letter of the law in order to keep more and more people behind bars. They pushed me to fundamentally reframe the potential utility of court data, away from problematic measures of individual "risk" in marginalized populations, toward studying the behaviors of decision makers who possess the most agency to drive outcomes in the system.

Others challenged us on the premise that we could use data to inform strategies for rehabilitative programming, arguing that the criminal legal system was life-sucking at its core. They pointed us to numerous examples in which data regimes had been expanded under the guise of informing rehabilitative interventions, only to be used to further surveil and profile vulnerable populations.[4] They pointed us to the work of people like the Chicago Community Bail Fund, which had recently published a report entitled, "Punishment is not a Service." In the report, advocates documented the various ways that pretrial interventions, such as electronic monitoring and mandatory curfews, had resulted in new forms of punishment, rather than creating avenues for rehabilitation.

My co-authors and I were receptive to these arguments, but we were still seduced by the prospect of carrying out causal inference research in the courts. We thought that studies like the ones we'd outlined in our paper might still be useful as a means of convincing officials to roll back harmful policies. Perhaps we could wield the cultural capital of data-driven discourse in

[4] For example, Andrew Ferguson (2016) illustrates this in the case of using predictive algorithms to create a "strategic subjects" list that identified individuals who are at risk of being involved in gun violence. The project was initially framed as a community health intervention, whereby at-risk subjects would be targeted for support services to minimize the incidence of gun violence. However, the algorithm quickly evolved into a tool used for targeting and arresting individuals who were considered "suspects" in incidence of gun violence. In instances where these individuals were charged with a specific crime, they were punished more severely than people who were not on such a list.

order to advocate for less punitive practices, by showing how ineffective they were. But then community advocates like James Kilgore helped us to understand the fundamental limitations of using criminal legal data in such studies. He warned us that if we tried to repurpose administrative data from the legal system to measure the impact of certain practices, then it would likely result in the erasure of harm, rather than reveal the ineffectiveness of punitive policies. Moreover, Kilgore explained that advocates didn't need a causal inference study in order to understand the harms of such practices, because it was already apparent in the testimonies of the people who had been subjected to them. He warned us that the partial insights gleaned from computational research could undermine more robust bodies of historically grounded and community-centered knowledge that were already available.

By 2020, I could see how the ideas that we'd presented in "Interventions Over Predictions" represented a failure of the imagination. Although we'd challenged a core assumption of the discourse surrounding pretrial risk assessment, we'd failed to envision an alternative approach to bail reform that moved beyond the measurement and management of risk in criminalized populations. In the years since we'd published the article, we had gotten better at probing the assumptions and problem formulations which so powerfully shaped our work. We'd even turned down an offer to conduct a study like the one proposed in our paper, in light of the arguments outlined above. Through relationships with abolitionist organizers, I'd become better at reckoning with the ways I was complicit in upholding harmful social systems through my engagements with data in the criminal legal system.

But I'd also grown to respect how seductive narratives about technology and social justice could be, and how prone I was to falling for them. Even as my ability to perceive the world in more liberatory terms grew, I found myself un-learning the same old habits and re-

learning the same lessons, over and over again. As I've become more familiar with the project of abolition, I've come to understand that this cyclical process of un-learning and re-learning is integral to the work of imagining the world more rigorously (Nagar and Shirazi 2019; Gilmore 2018; bergman and Montgomery 2017). It involves continuously interrogating our own good intentions and identifying the ways that we've been seduced by captivating visions of our role in making the world a better place. As Bonilla-Silva (2019) argues, it is important to understand the racialized emotions that drive active participation in maintaining violent and racist systems. While much attention has been given to the fear-based motives for upholding carceral systems of domination (Ivanov, Novisky, and Vogel 2021), less work has been done to understand the positive emotions that make harmful policies seem appealing to dominant groups. Rather than strive to create a framework that completely prevents such entanglements with dominant social structures, an abolitionist approach to change requires us to constantly interrogate the ways that we are bound up with oppressive systems, "because to remain troubled is to sustain a space of movement… that is itself improvised and cyclical" (bergman and Montgomery 2017, 7).

One of the core goals of this dissertation is to identify the specific mechanisms through which contemporary data regimes enable social hierarchies to endure amid struggles to build a freer world, with a focus on how the carceral state deploys data to recapture our collective imaginations during times of crisis. As I watched the events of 2020 unfold, I became particularly interested in exploring the role of captivating technology in sustaining the affective logics that underpin social inequality. By examining the ways that carceral technologies are framed as vehicles for compassion and care, we can unpack the mechanisms through which dominant ideologies make people feel good about sustaining violent systems over time.

Uncovering the affective dimensions of technology design and implementation are key to "un-inventing" harmful data regimes. As Mackenzie (1996, 203) explains, technologies are profoundly shaped by the social preconditions which make them seem desirable, or even indispensable — "to make a technology real, you have to move it from an idea on the drawing-board into something that people want…that they feel they cannot live without." He argues that it is actually possible to un-invent harmful technologies such as nuclear weapons, by attending to the social factors which sustain the demand for them over time. Building from Mackenzie's concept of "un-inventing" nuclear weapons, Keyes and Austin (2022) argue that the design of contemporary algorithmic systems is often dependent on tacit, emotionally-charged understandings of the world. They argue that such knowledge must be continuously interrogated, as a means of "creating space for unlearning (sometimes subtle, but always pernicious) ignorances and biases" which shape the work of technology design (Keyes and Austin 2022, 10). This requires multiple levels of inquiry — ranging from understanding how the political economy of data shapes which questions we choose to pursue, to interrogating the ways our personal feelings lure us into replicating unjust social dynamics in our work.

Abolitionist community organizers have been at the forefront of efforts to identify the structural mechanisms through which data regimes, and discourse on "ethical AI," fuels the expansion of carceral systems. This work often involves questioning progressive-sounding proposals for improving data systems, in order to show the ways that such ideas actually perpetuate the very systems they purport to reform (Hassein 2017; Stop LAPD Spying Coalition 2020). For example, the Stop LAPD Spying Coalition (2020) has written about the limits of research which focuses on the incompleteness and inaccuracies of criminal legal data, arguing that "the centering of technology, data bias, and dirty practices does not provide a broad enough

lens to understand the motivations and violences of algorithmic policing." They warn that a narrow focus on the technical specifications of carceral tools can actually lead to the expansion of harmful systems, as law enforcement agencies often use such arguments to lobby for increased resources. Similarly, a growing number of scholars have outlined the limits of terms such as diversity and inclusion, transparency, and bias as anchor concepts for debates regarding ethical data practice (Hoffmann 2019; 2020; Birhane 2021; Barabas 2020; D'Ignazio and Klein 2020). By delineating the limits of such critical concepts, these scholars push for the expansion of the analytical lens used to evaluate the harms of data regimes. This kind of work is critical to disrupting what Benjamin (2019a) calls the "techno quo" or the default assumptions and narratives that perpetuate social inequality over time.

Other groups, such as the Community Justice Exchange, have developed concepts like "data criminalization" to describe the ways that data are used to brand certain people as threats to the public (Community Justice Exchange 2022b). They have written extensively about the various mechanisms through which "data criminalization" occurs, in order to shed light on the often-invisible apparatus that fuels racial profiling in the present day. Scholars have also contributed to this effort, by outlining the ways that digital technologies are part of a long lineage of data-driven discourse used to construct racialized suspect populations over time (M'charek 2020; Browne 2015; Muhammad 2011; Benjamin 2019b; Sutherland 2019).

Perhaps the most exciting area of work for me comes from activists and scholars who have developed novel strategies for revealing and disrupting such harmful data practices. Many of these practices aim to co-opt and repurpose surveillance tools for the sake of survival and escape (Browne 2015). For example, the Anti-Eviction Mapping Project engages in "counter-mapping" to illustrate the relationship between the displacement of working-class communities of color and

the arrival of high-earning tech employees from Silicon Valley (Maharawal and McElroy 2018). Similarly, the Stop LAPD Spying Coalition (2021) has used public records of the LAPD's predictive policing platform to surface the ways that law enforcement creates containment zones around poor and housing-insecure communities by criminalizing their proximity to wealth. Such interventions invert the narrative lens used to interpret data, away from evaluating the supposed risk of criminalized people and toward understanding the ways data are used to justify ongoing criminalization.

Others have sought to interrogate the feelings, motivations, and mistakes stemming from their own data-centered research. For example, Keyes and Austin (2022), offer up a reflection on their experiences carrying out an audit of a facial recognition dataset, by specifically attending to the messy emotional experiences that they had while doing the work. They argue that such autoethnographic reflections create a much-needed "space for unlearning" biases and toxic attachments to oppressive systems. Similarly, Pierre et al (2021), use case studies drawn from their work in the field to demonstrate the value of self-critique as a way of surfacing the various "epistemic burdens" that community groups face when working on data-centered participatory design projects.

This dissertation builds from this wide-ranging body of work, in order to identify specific mechanisms through which data-centered research and ethics discourse prop up carceral systems. But the aim of this dissertation is not only to develop a keener eye for identifying potential harms, but also to cultivate a more robust vocabulary for talking about the role of technology in supporting efforts for transformational social change. In the context of the carceral state, language is a powerful tool for naturalizing harm against historically marginalized groups, but it is also an important site of struggle (hooks 1989). As Sylvia Wynter (2021) explains "Language

is the way that we will carry ourselves out of these problems we have… it is one part of a bigger project of developing a transformation of knowledge and therefore the transformation of the whole of society, by using a different language to address these intellectual concerns." This often involves developing a shared language that simultaneously problematizes oppressive ways of thinking while offering up alternative concepts to replace them. For example, abolitionist and scholar Nabil Hassein has challenged efforts to increase the representation of dark-skinned individuals in the training data for facial recognition software, arguing that efforts to render such software more accurate through the inclusion of underrepresented faces would do more harm than good. As Hassein (2017) argues,

> "The reality for the foreseeable future is that the people who control and deploy facial recognition technology at any consequential scale will predominantly be our oppressors. Why should we desire our faces to be legible for efficient automated processing by systems of their design? . . . The struggle for liberation is not a struggle for diversity and inclusion — it is a struggle for decolonization, reparations, and self-determination."

Rather than lobby for more diverse representation in data sets, Hassein offers ups a different set of values. These values center the pursuit of agency and healing, rather than increased inclusion in a harmful system. Hassein's statement demonstrates the ways we might not only critique efforts to optimize a particular technology, but also approach technologies as a battleground for asserting alternative imaginaries. As Jasanoff and Kim (2015, 6) describe, such imaginaries "encode not only visions of what is attainable through science and technology, but also of how life ought, or ought not, to be lived; in this respect they express a society's shared understandings of good and evil."

Abolitionist thinkers like Hassein are leading efforts to cultivate a shared vocabulary that adds weight and specificity to visions of alternative futures. For example, the Community Justice Exchange has developed the concept of "data liberation" to accompany their critiques of "data

criminalization" (Community Justice Exchange 2022a). Organizers at the Media Justice project have described their work as being "data informed" rather than "data driven," arguing that the primary role of data in their work is to build collective power, rather than as some top-down strategy for implementing narrow technical interventions (Media Justice 2021). A number of critical scholars have also outlined frameworks such as "design justice" (Costanza-Chock 2020), "data feminism" (D'Ignazio and Klein 2020), "techno-affects" (Amrute 2019), and "viral justice" (Benjamin 2022) that can serve as practical guides for data practitioners looking to support justice-oriented work.

These approaches don't mandate the wholesale rejection of data and technology. Rather, they offer up first-principles frameworks for fundamentally reformulating the questions we ask, the way we characterize existing data, and how we identify and fill gaps in existing data regimes. In the context of the criminal legal system, this requires shifting away from forecasting criminal behavior and towards understanding processes of criminalization, from supporting law and order towards increasing community safety and self-determination, and from surveilling "risky" populations towards increasing accountability of state officials. Shifting these fundamental assumptions is very important, as they inform 1) how we diagnose the problems we aim to solve, 2) what data we consider important for understanding the problem and 3) how we make claims based on the available data. In this dissertation, I'll draw from my own field work to explore how movement leaders and critical scholars challenge and reformulate the language surrounding the adoption of algorithmic technologies in the carceral state in order to mobilize and strengthen collective efforts to imagine alternative futures.

By exploring the above themes, my goal is to document some of my ongoing journey in trying to more rigorously imagine the world through and beyond technology. This involves

cultivating a radical imagination in my work. At its core, a radical imagination refers to the ability to uncover the root causes of oppression and then devise strategies for overcoming them (Haiven and Khasnabish 2014; Davis 1989). The radical imagination is a catalyzing force — it fuels the nitty-gritty work of abolitionist world-building by providing the language needed to make sense of the present, as well as envision and enact freer futures. Joy is the term that many people use to describe this process (bergman and Montgomery 2017; Benjamin 2022; Hope et al. 2019). Joy is the growth of people's capacity to perceive the world more expansively and live in ways that were previously unimaginable (bergman and Montgomery 2017). As Benjamin (2020) explains, "[I]f racism distorts our vision, then perhaps joy is one way of clarifying it, helping us to discern what we want to keep and what we must let go of, as individuals, as societies…"

Cultivating joy is also a collective pursuit — it's about being in relationship with people who push us to perceive the world more expansively and hold us accountable to living in ways that manifest the latent potential for justice in our everyday lives (Gumbs 2014). But such collective endeavors come with their own challenges, especially for people who are drawn to computational work. One of the major appeals of computation for many people is that it provides the opportunity to traverse multiple high-stakes domains while engaging in large-scale problem-solving (Selbst et al. 2019). In theory, the same basic computational methods might be used to develop a cure for malaria or to weigh in on important policy debates regarding public safety. Yet such interventions often ignore the nuances of local context and operate without concern for prior and ongoing grassroots efforts (Graziani 2020). Few computational practitioners take the time to cultivate relationships with directly impacted communities. And when they do, those interactions often lack the structures and commitments necessary to develop a shared sense of accountability and reciprocity in the encounter (Benjamin 2016c). As a result, computational

practitioners miss valuable opportunities to learn about the ways they might break cycles of domination and forge new kinds of mutually enabling-relationships in their work.

However, there's a growing body of scholarship which aims to help technology researchers and practitioners "organize ourselves" (Pierre et al. 2021) before seeking out partnership with community groups. For example, Hope et al (2019) have reflected on the ways that they unintentionally excluded the most vulnerable voices from their efforts to design better breastfeeding technology, as well as how they iterated on their initial approach in order to "design for equity" and center the experiences of marginalized parents in their subsequent projects. Graziani and Shi (2020) have written about strategies for overcoming the unequal terrain between academia and activism in order to foster mutually beneficial collaborations between technology researchers and community groups. And Irani and Alexander (2021) have discussed practical ways that technologists can support community advocacy against carceral tech, often by engaging in the more mundane aspects of making sense of harmful technologies.

This dissertation is premised on the fact that transformative social work is always already present and in formation (Kaba 2021), and explores the ways that computational practitioners might contribute to this ongoing work without assimilating it into dominant power structures. As Pierre et al (2021, 9) explain, this requires us to reorient "academic knowledge production away from primarily providing design 'solutions' to community 'problems' and toward directly addressing the unexamined political issues involved in academic and design researchers' engagement with minoritized communities." By addressing such issues, the goal is to increase the ability for technology designers to identify and align themselves with the emergent potential for justice that exists in our daily lives (Gumbs 2014).

By reflecting on the challenges and insights gleaned from my own collaborations with grassroots organizations, I aim to contribute to this important body of research, by identifying practical ways for data practitioners to participate in and support social transformation, rather than control it. Such work requires computational practitioners to embrace a theory of change that decenters data and focuses instead on cultivating new ways of relating "at the margins" (Shah 2015; Dutta and Pal 2010).

In order to explore the themes outlined above, I have structured my dissertation around three guiding questions:

1. What are the mechanisms through which data-driven technologies and research are used as a means of entrenching the carceral state during times of crisis?
2. What is the critical vocabulary that social movement leaders and critical scholars have identified for naming and responding to the harms that arise from data-driven tech?
3. In what ways do computational practitioners forge solidarities with transformative social movements in order to engage with data regimes in ways that posit radical alternatives to the status quo?

The core chapters of this dissertation represent various experiments in trying to cultivate a radical imagination through an exploration of the above questions. These chapters are organized in a two-part structure. With the exception of Chapter 6, the bulk of each chapter is comprised of writing that I have published in various venues over the course of my doctoral program. Each chapter opens with an autoethnographic introduction, where I offer some insight into how the concepts introduced in each chapter were developed, and how grappling with those concepts has helped me to work through specific ethical challenges in my work. Many of these revelations emerged from moments of "affective shock" (Shotwell 2011, xvi–xx), when life as usual was disrupted in ways that revealed the hidden contours of our social world. As Keyes and Austin

(2022) explain, such moments of rupture are essential sites of inquiry for computational practitioners who are committed to un-inventing harmful technologies, because they make it possible to reflect on and address the broader social conditions which enable such technologies to persist.

In Chapter 3, I reflect on how a scandal at my home institution during the early days of my PhD pushed me to think about the ways that my research was entangled with strategic corporate lobbying efforts to circumscribe conversations regarding "ethical AI." I then connect this reflection to a broader discussion of the ways that the political economy of data makes it challenging for data researchers to scrutinize powerful people and institutions in their work. This discussion is based on an article that I co-authored with Colin Doyle, Karthik Dinakar, and JB Rubinovitz, in which we explore the usefulness of a critical concept from anthropology called "studying up," in contemporary debates regarding ethical data practice. In our article, we outline the various challenges that data scientists face when trying to shift the algorithmic gaze "upward" to study people and institutions in positions of power. We then reflect on how "studying up" can be used as a method for revealing unarticulated assumptions and the hidden contours of the political economy underlying data-intensive work.

In Chapter 4, I reflect on some of the emotional discomfort I've experienced in choosing *not* to pursue research opportunities that I understood to be harmful. I then explore how the analytic of refusal helped me to process the insights that I learned from choosing not to pursue particular lines of inquiry in my work. Enclosed in this chapter is an article that I published in 2022, outlining two complementary modalities of refusal in computation: "refusal as resistance" and "refusal as re-centering the margins." By exploring these two modes of refusal, I hope to provide a practical vocabulary for identifying and resisting the ways that sociotechnical systems reinforce

dependency on oppressive structural conditions, as well as offer a framework for flexible collective experimentation towards more just futures.

The goal of these chapters is not necessarily to coin a new academic term or develop a novel framework for ethical data research. Rather, my aim is to revisit ideas that have been around for a long time and put them into conversation with the present, in order to cast contemporary debates on ethical data practice in a new light. This involves cultivating a thoughtful appreciation of the ways that prior struggles for justice can illuminate and inform our current efforts to transform the status quo. Such an approach hinges on an understanding of the role of history as a powerful resource for the radical imagination. As Jonathan Arac (1986, xviii) explains, the process of remembering can be a practice which "transforms history from a judgment on the past in the name of a present truth to a 'counter-memory' that combats our current modes of truth and justice, helping us to understand and change the present by placing it in a new relation to the past." By situating contemporary debates regarding data-driven reform within a much longer arc of collective struggle, we can engage in a what Chidgey (2019) calls "foot stepping," or tracing down the paths that others have left behind in order to forge a new pathway forward.

Chapter 5 opens with a reflection on the value and limitations of open letters as an abolitionist advocacy strategy. I also discuss the challenges of maintaining an abolitionist ethic of care in collaborations that traverse multiple axes of power and difference. Following this essay, I include two open letters which I co-authored as part of larger advocacy strategies to roll back harmful carceral technologies. These letters represent deeply interdisciplinary writing, wherein my co-authors and I combine historical analysis with technical expertise in order to produce arguments about the fundamental limitations of technologies that claim to predict and measure criminogenic risk.

Chapter 6 is also written in an epistolary format, and is inspired by Laurence Ralph's (2020) concept of ethnographic lettering, which he describes as both a mode of ethnographic writing and a research methodology that enables him to participate in "an 'unfinished' and ever-evolving dialogue" with the people and groups involved in the social problems he studies (2020, 194). This letter is addressed to a former collaborator of mine, wherein I reflect on lessons learned from a project that we never got the chance to complete. I explore the practical challenges of trying to use data to support decarceral outcomes in the U.S. legal system. Specifically, I discuss the role of "moral debt" and "fuzzy logic" in maintaining a culture of fear in the courts, as well as the limits and opportunities of trying to transform carceral organizations from the inside out.

Throughout all of these chapters, I offer up an approach to ethical computation that de-centers the *data* in "data ethics" in favor of cultivating a relationship-based theory of change, building our collective capacity to experiment, learn, and build new worlds together. This involves rigorously reflecting on what's worked and what hasn't in my own small experiments to un-invent carceral technology and make the world a more livable place for everyone. Such efforts are not the prelude to abolition, but the very practice itself (Gilmore 2018).

Works referenced

Alagraa, Bedour. 2021. “What Will Be the Cure?: A Conversation with Sylvia Wynter.” *Offshoot*, January. https://offshootjournal.org/what-will-be-the-cure-a-conversation-with-sylvia-wynter/.

American Civil Liberties Union. 2020. “COVID-19 Model Finds Nearly 100,000 More Deaths Than Current Estiimates, Due to Failures to Reduce Jails.” American Civil Liberties Union. https://www.aclu.org/sites/default/files/field_document/aclu_covid19-jail-report_2020-8_1.pdf.

Amrute, Sareeta. 2019. “Of Techno-Ethics and Techno-Affects.” *Feminist Review* 123 (1): 56–73. https://doi.org/10.1177/0141778919879744.

Arac, Jonathan, ed. 1986. *Postmodernism and Politics*. Theory and History of Literature 28. Manchester: Manchester University Press.

Barabas, Chelsea. 2020. “Beyond Bias: Re-Imagining the Terms of ‘Ethical Al’ in Criminal Law.” *Georgetown Journal of Law and Critical Race Theory* 12 (2): 83–112.

Barr, William. 2020a. “Prioritization of Home Confinesment As Appropriate Response to COVID-19 Pandemic.” Office of Attorney General.

———. 2020b. “Increasing Use of Home Confinement at Institutions Most Affected by COVID-19.” Bureau of Prisons, Department of Justice. https://www.justice.gov/file/1266661/download.

Benjamin, Ruha. 2016a. “Racial Fictions, Biological Facts: Expanding the Sociological Imagination through Speculative Methods.” *Catalyst: Feminism, Theory, Technoscience* 2 (2): 1–28. https://doi.org/10.28968/cftt.v2i2.28798.

———. 2016b. “Innovating Inequity: If Race Is a Technology, Postracialism Is the Genius Bar.” *Ethnic and Racial Studies* 39 (13): 2227–34. https://doi.org/10.1080/01419870.2016.1202423.

———. 2016c. “Informed Refusal: Toward a Justice-Based Bioethics.” *Science, Technology, & Human Values* 41 (6): 967–90. https://doi.org/10.1177/0162243916656059.

———, ed. 2019a. *Captivating Technology: Race, Carceral Technoscience, and Liberatory Imagination in Everyday Life*. Durham: Duke University Press.

———. 2019b. *Race after Technology: Abolitionist Tools for the New Jim Code*. Medford, MA: Polity.

———. 2020. “Black Skin, White Masks: Racism, Vulnerability & Refuting Black Pathology.” April 15. https://aas.princeton.edu/news/black-skin-white-masks-racism-vulnerability-refuting-black-pathology.

———. 2022. *Viral Justice: How We Grow the World We Want*. 1st ed. Princeton: Princeton University Press.

bergman, carla, and Nick Montgomery. 2017. *Joyful Militancy: Building Thriving Resistance in Toxic Times*. AK Press.

Birhane, Abeba. 2021. “Algorithmic Injustice: A Relational Ethics Approach.” *Patterns* 2 (2): 100205. https://doi.org/10.1016/j.patter.2021.100205.

Bonilla-Silva, Eduardo. 2019. “Feeling Race: Theorizing the Racial Economy of Emotions.” *American Sociological Review* 84 (1): 1–25. https://doi.org/10.1177/0003122418816958.

Bornstein, Avram. 2008. “Military Occupation as Carceral Society: Prisons, Checkpoints, and Walls in the Israeli-Palestinian Struggle.” *Social Analysis* 52 (2). https://doi.org/10.3167/sa.2008.520207.

Bronner Group. 2016. “Federal Bureau of Prisons Education Program Assessment: Final Report.” Chicago, IL.
Browne, Simone. 2015. *Dark Matters: On the Surveillance of Blackness*. Duke University Press.
Cacho, Lisa Marie. 2012. *Social Death: Racialized Rightlessness and the Criminalization of the Unprotected*. Nation of Newcomers: Immigrant History as American History. New York: New York University Press.
Chappell, Dale. 2020. “More Than Half of Chicago’s COVID-19 Cases Linked to Cook County Jail | Prison Legal News.” *Prison Legal News*, October 2020. https://www.prisonlegalnews.org/news/2020/oct/1/more-half-chicagos-covid-19-cases-linked-cook-county-jail/.
Chavez, Nicole. 2020. “A Movement to Push Police out of Schools Is Growing Nationwide. Here Is Why.” CNN. June 28, 2020. https://www.cnn.com/2020/06/28/us/police-out-of-schools-movement/index.html.
Chen, Yea-Hung, Alicia R Riley, Kate A Duchowny, Hélène E Aschmann, Ruijia Chen, Mathew V Kiang, Alyssa C Mooney, Andrew C Stokes, M Maria Glymour, and Kirsten Bibbins-Domingo. 2022. “COVID-19 Mortality and Excess Mortality among Working-Age Residents in California, USA, by Occupational Sector: A Longitudinal Cohort Analysis of Mortality Surveillance Data.” *The Lancet. Public Health* 7 (9): e744–53. https://doi.org/10.1016/S2468-2667(22)00191-8.
Chidgey, Red. 2019. *Feminist Afterlives*. New York, NY: Springer Berlin Heidelberg.
Ciaramella, C.J. 2020. “8 of The Top 10 Biggest U.S. Coronavirus Hotspots Are Prisons and Jails.” *Reason.Com*, April 29, 2020. https://reason.com/2020/04/29/8-of-the-top-10-biggest-u-s-coronavirus-hotspots-are-prisons-and-jails/.
Community Justice Exchange. 2022a. “Activity 2: Where We’ve Been, Where We’re Going.” From Data Criminalization to Prison Abolition. 2022. https://abolishdatacrim.org/en/resources/activity-2-where-weve-been-where-were-going.
———. 2022b. “Overview: The Invisible Machinery of Data Criminalization.” From Data Criminalization to Prison Abolition. 2022. https://abolishdatacrim.org/en/report/overview-invisible-machinery-data-criminalization.
Costanza-Chock, Sasha. 2020. *Design Justice: Community-Led Practices to Build the Worlds We Need*. Information Policy. Cambridge, Massachusetts: The MIT Press.
Courtwatch MA. 2020. “Stories from Court.” CourtWatch MA. August 1, 2020. https://www.courtwatchma.org/stories-from-court.html.
Davis, Angela Y. 1989. *Women, Culture & Politics*. 1st ed. New York: Random House.
———. 2003. *Are Prisons Obsolete?* Open Media Book. New York: Seven Stories Press.
Demuth, Stephen, and Darrell Steffensmeier. 2004. “The Impact of Gender and Race-Ethnicity in the Pretrial Release Process.” *Social Problems* 51 (2): 222–42. https://doi.org/10.1525/sp.2004.51.2.222.
D’Ignazio, Catherine, and Lauren F. Klein. 2020. *Data Feminism*. The MIT Press. https://doi.org/10.7551/mitpress/11805.001.0001.
Dixon-Román, Ezekiel, T. Philip Nichols, and Ama Nyame-Mensah. 2020. “The Racializing Forces of/in AI Educational Technologies.” *Learning, Media and Technology* 45 (3): 236–50. https://doi.org/10.1080/17439884.2020.1667825.
Dutta, Mohan, and Mahuya Pal. 2010. “Dialog Theory in Marginalized Settings: A Subaltern Studies Approach.” *Communication Theory* 20 (4): 363–86. https://doi.org/10.1111/j.1468-2885.2010.01367.x.

Eubanks, Virginia. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.
Ferguson, Andrew Guthrie. 2016. "Policing Predictive Policing." *Wash. UL Rev.* 94: 1109.
Ferreira da Silva, Denise. 2009. "No-Bodies: Law, Raciality and Violence." *Griffith Law Review* 18 (2): 212–36.
Friedman, Brittany. 2021. "Toward a Critical Race Theory of Prison Order in the Wake of COVID-19 and Its Afterlives: When Disaster Collides with Institutional Death by Design." *Sociological Perspectives*, April, 073112142110054. https://doi.org/10.1177/07311214211005485.
Gabriel, Kay. 2022. "Abolition as Method." *Dissent Magazine*, 2022. https://www.dissentmagazine.org/article/abolition-as-method.
Gilmore, Ruth Wilson. 2007. *Golden Gulag: Prisons, Surplus, Crisis, and Opposition in Globalizing California*. American Crossroads. Berkeley: University of California Press.
———. 2018. "Prisons and Class Warfare: An Interview with Ruth Wilson Gilmore." Verso. August 2, 2018. https://www.versobooks.com/en-gb/blogs/news/3954-prisons-and-class-warfare-an-interview-with-ruth-wilson-gilmore.
———, dir. 2020. *Geographies of Racial Capitalism with Ruth Wilson Gilmore*. Antipode Foundation. https://www.youtube.com/watch?v=2CS627aKrJI.
Graziani, Terra. 2020. "Data for Justice: Tensions and Lessons from the Anti-Eviction Mapping Project's Work between Academia and Activism," 16.
Gumbs, Alexis Pauline. 2013. "Sometimes Freedom Means You Have To Burn It Down: Harriet Tubman and Abolitionist Vision That Don't Quite." *The Abolitionist*, 2013, Fall edition.
———. 2014. "Prophecy in the Present Tense." *Meridians* 12 (2): 142–52. https://doi.org/10.2979/meridians.12.2.142.
Haiven, Max, and Alex Khasnabish. 2014. *The Radical Imagination: Social Movement Research in the Age of Austerity*. Halifax ; Winnipeg : London: Fernwood Publishing ; Zed Books.
Hassein, Nabil. 2017. "Against Black Inclusion in Facial Recognition." *Digital Talking Drum* (blog). August 15, 2017. https://digitaltalkingdrum.com/2017/08/15/against-black-inclusion-in-facial-recognition/.
Hoffmann, Anna Lauren. 2019. "Where Fairness Fails: On Data, Algorithms, and the Limits of Antidiscrimination Discourse." *Under Review with Information, Communication, and Society*.
———. 2020. "Terms of Inclusion: Data, Discourse, Violence." *New Media & Society*, September, 146144482095872. https://doi.org/10.1177/1461444820958725.
hooks, bell. 1989. "CHOOSING THE MARGIN AS A SPACE OF RADICAL OPENNESS." *Framework: The Journal of Cinema and Media*, no. 36: 15–23.
Hooks, Gregory. 2020. "COVID-19 Spread Faster in Counties with Large Prisons - and to Nearby Counties: Marion County Is a Disturbing Example." December 2020. https://www.prisonpolicy.org/reports/covidspread_marion.html.
Hooks, Gregory, and Wendy Sawyer. 2020. "Mass Incarceration, COVID-19, and Community Spread." *Reports* (blog). December 2020. https://www.prisonpolicy.org/reports/covidspread.html.
Hope, Alexis, Catherine D'Ignazio, Josephine Hoy, Rebecca Michelson, Jennifer Roberts, Kate Krontiris, and Ethan Zuckerman. 2019. "Hackathons as Participatory Design: Iterating Feminist Utopias." In *Proceedings of the 2019 CHI Conference on Human Factors in*

Computing Systems, 1–14. Glasgow Scotland Uk: ACM. https://doi.org/10.1145/3290605.3300291.
Howard, Ashley M. 2022. "Violence and Disposability at the Intersection of Twin Pandemics." *Labor Studies Journal* 47 (4): 493–500. https://doi.org/10.1177/0160449X221121646.
Irani, Lilly, and Khalid Alexander. 2021. "The Oversight Bloc." *Logic Magazine*, December 25, 2021. https://logicmag.io/beacons/the-oversight-bloc/.
Ivanov, Stefan, Meghan A. Novisky, and Matt Vogel. 2021. "Racial Resentment, Empathy, and Support for Release during COVID-19: Results from a Survey Experiment." *Socius: Sociological Research for a Dynamic World* 7 (January): 237802312110052. https://doi.org/10.1177/23780231211005222.
Jasanoff, Sheila, and Sang-Hyun Kim. 2015. *Dreamscapes of Modernity: Sociotechnical Imaginaries and the Fabrication of Power*. University of Chicago Press.
Kaba, Mariame. 2021. *We Do This 'Til We Free Us: Abolitionist Organizing and Transforming Justice*. Haymarket Books.
Keeling, Kara. 2019. *Queer Times, Black Futures*. Sexual Cultures. New York: New York University Press.
Kelley, Robin D. G. 2002. *Freedom Dreams: The Black Radical Imagination*. Boston: Beacon Press.
Keyes, Os, and Jeanie Austin. 2022. "Feeling Fixes: Mess and Emotion in Algorithmic Audits." *Big Data & Society* 9 (2): 205395172211137. https://doi.org/10.1177/20539517221113772.
Khan-Cullors, Patrisse, and Asha Bandele. 2017. "When They Call You a Terrorist: A Black Lives Matter Memoir. New York, NY: St."
Khazanchi, Rohan, Charlesnika T. Evans, and Jasmine R. Marcelin. 2020. "Racism, Not Race, Drives Inequity Across the COVID-19 Continuum." *JAMA Network Open* 3 (9): e2019933. https://doi.org/10.1001/jamanetworkopen.2020.19933.
Kurwa, Rahim. 2018. In *"Freedom Is a Place: Land, Rent, and Housing."* UCLA Luskin.
Leadership Conference on Civil and Human Rights. 2019. "The Leadership Conference et al Comment Letter to Department of Justice on PATTERN." http://civilrightsdocs.info/pdf/policy/letters/2019/The%20Leadership%20Conference%20et%20al%20Comment%20Letter%20to%20Department%20of%20Justice%20on%20PATTERN%20%20First%20Step%20Act%209%203%202019.pdf.
Lipsitz, George. 2011. *How Racism Takes Place*. Philadelphia: Temple University Press.
MacDougall, Ian. 2020. "Bill Barr Promised to Release Prisoners Threatened by Coronavirus — Even as the Feds Secretly Made It Harder for Them to Get Out." *ProPublica*, May 26, 2020. https://www.propublica.org/article/bill-barr-promised-to-release-prisoners-threatened-by-coronavirus-even-as-the-feds-secretly-made-it-harder-for-them-to-get-out?token=8x14roDzBQ5p8Y0hCArCtdwGWCNrGooB.
Mackenzie, Donald. 1996. "Uninventing the Bomb?*." *Medicine and War* 12 (3): 202–11. https://doi.org/10.1080/13623699608409285.
MacKenzie, Donald A. 1993. *Inventing Accuracy: A Historical Sociology of Nuclear Missile Guidance*. 4. print. Inside Technology. Cambridge, Mass.: MIT Press.
Maharawal, Manissa M., and Erin McElroy. 2018. "The Anti-Eviction Mapping Project: Counter Mapping and Oral History toward Bay Area Housing Justice." *Annals of the American Association of Geographers* 108 (2): 380–89. https://doi.org/10.1080/24694452.2017.1365583.

Mauer, Marc. 2010. “Justice for All? Challenging Racial Disparities in the Criminal Justice System.”
Mayson, Sandra G. 2017. “Dangerous Defendants.” *Yale LJ* 127: 490.
M’charek, Amade. 2020. “Tentacular Faces: Race and the Return of the Phenotype in Forensic Identification.” *American Anthropologist* 122 (2): 369–80. https://doi.org/10.1111/aman.13385.
McKittrick, Katherine. 2014. “Mathematics Black Life.” *The Black Scholar* 44 (2): 16–28. https://doi.org/10.1080/00064246.2014.11413684.
McLeod, Allegra M. 2018. “Envisioning Abolition Democracy.” *Harvard Law Review* 132: 1613.
Media Justice, dir. 2021. *Points of Connection: Mapping Electronic Monitoring to Challenge E-Carceration*. https://www.youtube.com/watch?v=w_j-r8xqIZM.
Melamed, Jodi. 2019. “Operationalizing Racial Capitalism: Administrative Power and Ordinary Violence.” Yale University, November 19. https://www.youtube.com/watch?v=o3Z9sOGf6BA&t=2675s.
Meyer, John W., and Brian Rowan. 1977. “Institutionalized Organizations: Formal Structure as Myth and Ceremony.” *American Journal of Sociology* 83 (2): 340–63.
Muhammad, Khalil Gibran. 2011. *The Condemnation of Blackness*. Harvard University Press.
Nagar, Richa, and Roozbeh Shirazi. 2019. “Radical Vulnerability.” In *Keywords in Radical Geography: Antipode at 50*, edited by Antipode Editorial Collective, Tariq Jazeel, Andy Kent, Katherine McKittrick, Nik Theodore, Sharad Chari, Paul Chatterton, et al., 236–42. Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/9781119558071.ch44.
Office of the Attorney General. 2019. “The First Step Act of 2018 : Risk and Needs Assessment System.” Department of Justice.
Omi, Michael, and Howard Winant. 2015. *Racial Formation in the United States*. Third edition. New York: Routledge/Taylor & Francis Group.
Patterson, Orlando. 1982. *Slavery and Social Death: A Comparative Study*. Cambridge, Mass: Harvard University Press.
Patton, Stacey. 2020. “Perspective | The Pathology of American Racism Is Making the Pathology of the Coronavirus Worse.” *Washington Post*, April 11, 2020. https://www.washingtonpost.com/outlook/2020/04/11/coronavirus-black-america-racism/.
Perez, Matt. 2020. “Surgeon General Tells People Of Color To Avoid Alcohol, Drugs To Protect Against Coronavirus; Defends ‘Big Mama’ Comments.” Forbes. April 10, 2020. https://www.forbes.com/sites/mattperez/2020/04/10/surgeon-general-tells-people-of-color-to-avoid-alcohol-drugs-to-protect-against-coronavirus-defends-big-mama-comments/.
Perilous Chronicle. 2020. “First 90 Days of Prisoner Resistance to COVID-19: Report on Events, Data, and Trends.” Perilous Chronicle. https://perilouschronicle.com/2020/11/12/covid-prisoner-resistance-first-90-days-full-report/.
Pierre, Jennifer, Roderic Crooks, Morgan Currie, Britt Paris, and Irene Pasquetto. 2021. “Getting Ourselves Together: Data-Centered Participatory Design Research & Epistemic Burden.” In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–11. Yokohama Japan: ACM. https://doi.org/10.1145/3411764.3445103.

Pilkington, Ed. 2020. “Covid-19 Death Rate among African Americans and Latinos Rising Sharply.” *The Guardian*, September 8, 2020, sec. World news. https://www.theguardian.com/world/2020/sep/08/covid-19-death-rate-african-americans-and-latinos-rising-sharply.

Ralph, Laurence. 2020. *The Torture Letters: Reckoning with Police Violence*. Chicago ; London: The University of Chicago Press.

Ray, Victor. 2019. “A Theory of Racialized Organizations.” *American Sociological Review* 84 (1): 26–53. https://doi.org/10.1177/0003122418822335.

Rehavi, M. Marit, and Sonja B. Starr. 2014. “Racial Disparity in Federal Criminal Sentences.” *Journal of Political Economy* 122 (6): 1320–54. https://doi.org/10.1086/677255.

Rivera, Alissa. 2020. “Illinois Failing Key Pillar of COVID-19 Response.” *Restore Justice* (blog). June 15, 2020. https://restorejustice.org/early-releases-exacerbate-racial-inequity/.

Robinson, Cedric J. 2000. *Black Marxism: The Making of the Black Radical Tradition*. Chapel Hill, N.C: University of North Carolina Press.

———. 2007. *Forgeries of Memory and Meaning: Blacks and the Regimes of Race in American Theater and Film before World War II*. Chapel Hill: University of North Carolina Press.

Roy, Ananya, ed. 2019. “Racial Banishment.” In *Keywords in Radical Geography: Antipode at 50*. Chichester, West Sussex, United Kingdom ; Hoboken, NJ: Wiley-Blackwell for the Antipode Foundation Ltd.

Roy, Arundhati. 2020. “‘The Pandemic Is a Portal.’” *Financial Times*, April 3, 2020. https://www.ft.com/content/10d8f5e8-74eb-11ea-95fe-fcd274e920ca.

Sawyer, Wendy, and Alexi Jones. 2021. “New Data on Jail Populations: The Good, the Bad, and the Ugly.” *Briefings* (blog). March 17, 2021. https://www.prisonpolicy.org/blog/2021/03/17/jails/.

Schwarz, Bill, and Stuart Hall. 1988. “Questions of Theory.”

Seigel, Micol. 2018. *Violence Work: State Power and the Limits of Police*. Durham : London: Duke University Press.

Selbst, Andrew D., Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. “Fairness and Abstraction in Sociotechnical Systems.” In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68. Atlanta GA USA: ACM. https://doi.org/10.1145/3287560.3287598.

Shah, Nishant. 2015. “Networked Margins: Revisiting Inequality and Intersection.” Digitally Connected: Global Perspectives on Youth and Digital Media. http://www.ssrn.com/abstract=2585686.

Shotwell, Alexis. 2011. *Knowing Otherwise: Race, Gender, and Implicit Understanding*. University Park, Pa: Pennsylvania State University Press.

Sims, Kaitlyn M., Jeremy Foltz, and Marin Elisabeth Skidmore. 2021. “Prisons and COVID-19 Spread in the United States.” *American Journal of Public Health* 111 (8): 1534–41. https://doi.org/10.2105/AJPH.2021.306352.

Spade, Dean. 2015. *Normal Life: Administrative Violence, Critical Trans Politics, and the Limits of Law*. Revised and Expanded edition. Durham, [North Carolina]: Duke University Press.

Stop LAPD Spying Coalition. 2020. “The Algorithmic Ecology: An Abolitionist Tool for Organizing Against Algorithms.” *Stoplapdspying* (blog). March 3, 2020. https://stoplapdspying.medium.com/the-algorithmic-ecology-an-abolitionist-tool-for-organizing-against-algorithms-14fcbd0e64d0.

———. 2021. “Automating Banishment.” Los Angeles: Stop LAPD Spying Coalition. https://automatingbanishment.org/assets/AUTOMATING-BANISHMENT.pdf.
Suchman, Lucille Alice. 2007. *Human-Machine Reconfigurations: Plans and Situated Actions*. 2nd ed. Cambridge ; New York: Cambridge University Press.
Sutherland, Tonia. 2019. “The Carceral Archive: Documentary Records, Narrative Construction, and Predictive Risk Assessment.” *Journal of Cultural Analytics*. https://doi.org/10.22148/16.039.
The Sentencing Project. 2018. “Report to the United Nations on Racial Disparities in the U.S. Criminal Justice System.” The Sentencing Project. https://www.sentencingproject.org/reports/report-to-the-united-nations-on-racial-disparities-in-the-u-s-criminal-justice-system/.
United Nations, The. 2020. “UN Rights Chief Urges Quick Action by Governments to Prevent Devastating Impact of COVID-19 in Places of Detention.” UN News. March 25, 2020. https://news.un.org/en/story/2020/03/1060252.
Wacquant, Loïc. 2008. “The Militarization of Urban Marginality: Lessons from the Brazilian Metropolis.” *International Political Sociology* 2 (1): 56–74. https://doi.org/10.1111/j.1749-5687.2008.00037.x.
Wang, Jackie. 2018. *Carceral Capitalism*. Semiotext(e).
Widra, Emily. 2020. “As COVID-19 Continues to Spread Rapidly, State Prisons and Local Jails Have Failed to Mitigate the Risk of Infection behind Bars.” *Briefings* (blog). December 2, 2020. https://www.prisonpolicy.org/blog/2020/12/02/jail-and-prison-covid-populations/.
Williams, Brie, Cyrus Ahalt, David Cloud, Dallas Augustine, Leah Rorvig, and David Sears. 2020. “Correctional Facilities In The Shadow Of COVID-19: Unique Challenges And Proposed Solutions | Health Affairs Blog.” *Health Affairs* (blog). March 2020. https://www.healthaffairs.org/do/10.1377/hblog20200324.784502/full/.
Williams, Brie, and Leann Bertsch. 2020. “A Public Health Doctor And Head Of Corrections Agree: We Must Immediately Release People From Jails And Prisons.” *The Appeal*, March 27, 2020. https://theappeal.org/a-public-health-doctor-and-head-of-corrections-agree-we-must-immediately-release-people-from-jails-and-prisons/.

Chapter 2

Research Methods

This dissertation is organized around a collection of working concepts that I hope can be useful for computational practitioners interested in supporting transformative action against and beyond harmful data regimes. Rather than develop a set of cut-and-dry best practices, I aim to provide a framework for more flexible and iterative conversations regarding the fast-changing realm of data and social change, one that is grounded in a critical vocabulary that surfaces taken for granted assumptions and identifies root causes of ongoing inequality. This requires "reading data" in incredibly dynamic and unstable social and political contexts (McGlotten 2016; Cifor et al. 2019), in order to discern how various stakeholders enact social realities and give them meaning (Goldkuhl 2012). As such, I've collected and analyzed my data within an interpretivist tradition, focusing on specific, contextualized environments and acknowledging that the knowledge produced from observing these contexts is situated and partial. The end result is a set of what I've termed "working concepts," ideas that 1) do work (name invisible processes, interrogate assumptions, pose alternatives, etc. and 2) are in constant need of revision and refinement through praxis.

This dissertation is grounded in a set of collaborative projects that took place between 2017 and 2021. The methods I use to explore my research questions are varied and reflect the unique challenges and opportunities that arise when engaging with three distinct groups of research subjects and collaborators: 1) government and policy insiders working on criminal legal reform, such as administrators of the courts, judges, policymakers, and funders of criminal justice reform; 2) abolitionist organizers and other grassroots community leaders, such as members of the National Bail Fund Network; and 3) interdisciplinary groups of scholars who

organize collective action campaigns within the academy, such as the Coalition for Critical Technology.

My research design is premised on an understanding that the actions (or inactions) of academic researchers are never neutral. We play an active role in either maintaining the status quo or transforming it through our research practices. In my research design, I positioned myself as an active participant in ongoing conversations regarding the use of digital technologies in contested political debates. However, the nature of my participation in these debates has evolved over time, as I learned how to better identify and address my entanglements with structures of oppression and build stronger, more accountable relationships across power differentials. Similar to other critical scholars, I characterize my position as a researcher as "a fluid space of crossing borders and, as such, a contradictory one of collusion and oppositionality, complicity, and subversion" (Villenas 1996, 729). In the early stages of my research, I saw myself as a potential bridge-figure, someone who might facilitate productive dialogues across diverse groups of people with competing agendas. I sought out collaborations with a wide range of social actors, including government insiders and community organizers, in the hopes that I might help to establish common ground and foster mutual understanding across these groups.

However, over time I came to align myself increasingly with people and organizations anchored in an abolitionist worldview. My relationships with these groups profoundly shaped each phase of this dissertation, from the formulation of my research questions, to data collection and analysis, and the dissemination of the results. As I shifted from being an academic who occasionally dropped into movement spaces to becoming a member of a community of struggle, the goals of my research changed. My thought partners pushed me to go beyond armchair cultural criticism, to embrace a pragmatic and action-oriented stance, in order to explore not only

what ‘is’ but what ‘might be’ through my inquiry. As Dewey (1931, 8) argues, “an empiricism which is content with repeating facts already past has no place for possibility and for liberty.” This action-oriented approach to research design is aligned with Ruth Wilson Gilmore’s notion of scholar activism. As Gilmore (2007, 27) explains, “in scholarly research, answers are only as good as the further questions they provoke, while for activists, answers are as good as the tactics they make possible. Where scholarship and activism overlap is in the area of how to make decisions about what comes next.” In the subsequent chapters of this dissertation, I aim to do two kinds of bridging work: 1) bridging across disciplines, working closely with computer scientists, historians and activists to make grounded, historicized claims that have direct implications for how we think about the validity of specific computational practices; and 2) bridging scholarship into action, creating frameworks and arguments that can inform real-world practices outside of the academy.

As a critical data scholar from MIT, often working in collaboration with data scientists and legal experts, I have been invited to give input on the design, implementation, and evaluation of data-intensive systems. As such, I place myself into a class of researchers whom I refer to as “computational practitioners,” which includes a wide range of actors in academia, industry, and government who are engaged in data-centered discourse, research, and design. Computational practitioners often occupy privileged positions within organizations as either valued insiders or welcomed outsiders whose technical skills are in high demand. I had the opportunity to observe and participate in both public and private conversations with policymakers, criminal legal professionals, and funders of criminal justice reform regarding the limits and opportunities of data-driven reforms in the criminal legal system.

In the first year of my research, I spent several months conducting ethnographic fieldwork, during which I made connections with court administrators and technical support staff at professional convenings, such as the National Association of Pretrial Service Agencies Conference. I also attended policy gatherings and roundtable discussions hosted by private foundations who were investing millions of dollars in "data-driven reform" in the criminal legal system. Such meetings gave me the opportunity to observe the ways that the problem of mass pretrial incarceration was framed and how data-driven interventions such as actuarial risk assessment were posited as possible solutions. I often followed up with participants from these convenings to have more casual one-on-one conversations, where I had the opportunity to probe into the ways that the official narratives surrounding data-driven reforms were "decoupled" (Christin 2017) from the daily practices of the courts.

During this time, I also joined an interdisciplinary team of academic researchers who were studying pretrial risk assessments as part of a joint research initiative called the Ethics and Governance of AI Initiative. This research program was housed between the MIT Media Lab and the Harvard Law School, and its expressed goal was to explore novel ways of combining legal, humanistic, and scientific modes of thinking in order to contribute to conversations regarding the ethical stakes of data-intensive systems. Our team initially included me, Karthik Dinakar (data scientist), and Madars Virza (cryptographer). Our research was anchored in a collaboration with an administration of the courts in a mid-sized U.S. state. During this collaboration, our team worked with the state's administration of the courts to design a "judge dashboard" that could surface trends in the ways judges set bail and responded to specific pretrial reforms. As the resident social scientist on the team, my work was initially situated as what Dourish and Button (1998) term a "service science," whereby I used qualitative research methods, such as interviews

and surveys, in order to inform the design and implementation of a technological system. However, over time, I took on more of a leadership role, in which I challenged and reformulated the core assumptions, goals, and epistemic stakes surrounding our project, based on my ethnographic research. In this way, I situate my fieldwork within the lineage of other STS-oriented social scientists such as Lucy Suchman and Susan Leigh Star, who have taken on consultant-like roles within the organizations they are studying (Suchman 1987; Bowker, Timmermans, and Star 1997; Ribes 2019). Such a role gave me the opportunity to not only study the data systems housed within the administration of the courts, but to also shape the sociotechnical landscape in which these systems were being designed and implemented (Ribes 2019). As an active collaborator on an applied data science project in this context, I gained access to conversations and spaces that gave me deeper insight into the challenges and complexities of using data to reduce pretrial jail populations.

By immersing myself as a courtroom "insider," I was able to enter both public and private spaces surrounding the courts (i.e. courtrooms, judge chambers, administrative offices). This behind-the-scenes access gave me the opportunity to observe the beliefs, practices, and taken for granted assumptions that shaped the day-to-day functioning of the courts (Gusterson 1997). Moreover, by forging a collaboration that hinged on the use of administrative court data, I had the opportunity to map out the contingencies and ambiguities that arose in the construction of criminal legal data, as well as the political and ideological struggles that shaped how that data was interpreted and used. This approach was inspired by the work of Nicole Gonzalez Van Cleve (2016), whose multi-year ethnography of the Cook County court system demonstrates the ways that prolonged interaction with courtroom insiders enables better insight into the gaps that

emerge between the formal rule of law and how legal professionals interpret those rules in practice.

In addition to conducting participant observation, I interviewed court administrators and judges on their views regarding the use of data to support pretrial reform. These interviews provided me with an opportunity to more fully interrogate the beliefs and rationales of key decision makers in the courts. It also gave me the chance to have more frank conversations about sensitive topics, such as why judges systematically stray from the letter of the law when setting bail. As Van Cleve (2016), argues in her reflections of her own courtroom ethnography, being a young White woman who has never practiced law had some clear advantages in this setting. My demographic and professional background made me seem "non-threatening" (in terms of gender, race, age and professional experience) and teachable to many of my interlocutors, which often led to surprisingly candid conversations. After hundreds of hours of field visits, interviews, and data analysis, this project was ultimately shut down due to resistance from judges to having their decisions formally monitored by court administrators. However, I gleaned many rich insights along the way, as we attempted to use court data to support culture change and create structures of accountability for judges.

During this same period of time, I engaged in participatory action research in collaboration with grassroots organizers working on bail reform and other campaigns aimed at abolishing the carceral state. In this dissertation, I reflect on two participatory action research projects related directly to bail reform, which I carried out from 2018 to 2020. The first project was geared towards supporting a Boston-based organization as they collected, analyzed and disseminated insights gleaned from a local court watch initiative. Court watching is a community-based advocacy strategy that is designed to track the behavior of key decision

makers such as judges and prosecutors in the courts. I first learned about court watching in 2018, during a convening that I co-hosted at MIT, where we brought together community organizers from around the country to strategize around abolitionist approaches to using data to address mass pretrial incarceration.

At this convening, I learned that local leaders were about to launch a court watching initiative in Suffolk County (Boston). I offered to support the project in whatever way I could, which initially involved attending weekly meetings with Atara Rich-Shea, the leader of the project, to discuss implementation details, such as the design of the data collection forms and approaches to training community members in data collection. These weekly meetings spanned from March to October 2018, and eventually expanded to include two research assistants that I hired to support this project. I supervised these research assistants as they carried out different administrative tasks, such as coordinating volunteers and carrying out quality assurance processes with the court watch data.

Once the project was launched, I volunteered as a court watcher myself, in order to get first-hand experience with the challenges and insights generated by court watching activities. In the fall of 2018, I worked with Colin Doyle and Atara Rich-Shea to enlist students at MIT and Harvard in larger scale efforts to help with the tedious work of entering handwritten court observations into a database. During these convenings, I got to observe both the technical details of managing the administrative side of the project, as well as the ideological work that Atara and her colleagues did, educating students about the purpose and goals of the project.

Finally, I worked with other volunteers to clean up and analyze the data once it was digitized. This involved in-depth conversations about the quality of the data that had been collected and the types of arguments and campaigns that could be built off that data. This project

enabled me to build stronger relationships with the network of organizers who attended the February 2018 convening mentioned above. I became a known quantity, someone who could be called upon if the network needed support on something that fell within the realm of my expertise. And I became comfortable with calling on them whenever I sought out guidance or accountability on a project I was working on.

The second project aimed to the develop a "judge risk assessment" tool, which used criminal legal data to critically engage with mainstream narratives regarding how data might support bail reform in the courts. I collaborated with Rebekah Agwunobi (undergraduate computer science student), Colin Doyle (legal scholar), and JB Rubinovitz (data scientist) to develop this tool over the course of 2019 and 2020. We also built a website to share the project and provide educational materials for students and organizers interested in reframing the issue of bail reform as a challenge of judge accountability, not defendant risk. During the development of the tool and design of the website, our team engaged in ongoing conversations with leaders of the National Bail Fund Network, who gave us constructive feedback on the ways this work might support broader abolitionist interventions surrounding the use of pretrial risk assessment. Most of these conversations took place online, as the organizers we spoke with were located around the United States, in places like New York, California, Pennsylvania and Texas. I supplemented my fieldnotes from these conversations with one-on-one interviews with community organizers, reflecting on the strengths and weaknesses of the tactics we'd developed to complete both the court watching and judge risk assessment projects. In these conversations, we compared and contrasted our approaches with the tactics of other advocates around the country, such as the Criminal Court Data Dashboard in Harris County, Texas.

Through these conversations, I began to appreciate the potential value of investing time and energy in organizing within the academy. My interlocutors encouraged me to reframe academic labs and research initiatives as potential sites of political resistance, similar to the way critical race scholars had conceptualized law schools in the 1980's (Delgado and Stefancic 2013). As knowledge workers with significant cultural capital and technical expertise, it was essential that academics worked to ensure that "AI ethics" research did not ultimately undermine movements for liberation by narrowing the discourse surrounding racial justice and algorithmic systems down to a set of abstract and ideologically impoverished concepts. As such, I began to reframe my academic community as a third site of research, a place where I could engage in participant observation and collaborate on interventions that re-conceptualized academic spaces as places where important political work took place.

My first significant intervention in this space was to co-author with Karthik Dinakar and Colin Doyle an open letter that outlined the fundamental limitations of actuarial risk assessments as a tool for bail reform. This letter was signed by twenty-seven prominent scholars from a wide range of disciplines, including computer science, law, sociology and anthropology. We then worked with organizers and policy advocates to disseminate this letter to lawmakers across the country. This effort gave me the opportunity to think through an interdisciplinary approach to developing better language for talking about the limitations and potential harms of algorithmic systems in the criminal legal system.

During COVID-19, the nature of my research methods significantly changed. By then, my "insider" collaboration with the state-wide administration of the courts had been placed on an indefinite pause and I was working primarily to support an informal network of grassroots community organizations as they organized local and national campaigns to release incarcerated

people from prisons and jails. During this time, the effort to reduce jail and prison populations took on new urgency, as the bloated U.S. penal system became a major vector for the spread of COVID-19 (American Civil Liberties Union 2020; Chappell 2020; Hooks 2020). As a computational practitioner, part of my job was to comb through government documents in an effort to understand how state officials were using technology to respond to demands to reduce prison and jail populations, as well as quell collective action efforts behind bars. This work was similar to other efforts that technical experts have made to support community organizing efforts, by using their specialized knowledge to identify possible harms and counter official claims about what a given technology can and cannot do (Irani and Alexander 2021).

Most of my research during this time period is drawn from what Ortner calls "interface events," or sites where a population which is hard to access presents itself to the public (Ortner 1995; Souleles 2018). As various scholars have noted, one of the biggest challenges of studying elite system actors is gaining access. Studying authority figures often requires an "eclectic" mix of research methods, in which the researcher engages with a wide range of texts and informants across dispersed field sites in order to surface patterns and insights (Gusterson 1997). This was particularly true during COVID-19, when my interactions with collaborators and research subjects were largely limited to virtual online spaces.

The insights gleaned from this period of time were developed through a close reading of a diverse set of texts, including industry marketing materials, official government statements, newspaper articles, and social movement advocacy materials. From February to December 2020, I collected and analyzed official press releases and policies published by the Federal Bureau of Prisons, the U.S. Attorney General's office and the National Sheriff's Association regarding their response to COVID-19. The starting point of this research emerged after Attorney General Barr

released a memo in March 2020, mandating that a federal risk assessment tool called PATTERN be used to determine who in the federal prison system was eligible for transfer to home confinement (Barr 2020). I was part of a small group of academics and organizers who penned an open letter arguing against the use of PATTERN to make such decisions. After this letter, I continued to work with this group to track the official COVID-19 response policies of penal officials around the country. Occasionally, an organizer within this network would send me an email, requesting that I look into a specific technology that was reportedly being implemented as part of the COVID-19 response in a specific penal facility. Such requests served as the starting point for deeper investigation, wherein I gathered promotional materials, how-to guides, and white papers from private technology companies who marketed their products specifically as penal risk management tools during the pandemic.

In order to synthesize this wide range of texts, I was guided by Ruha Benjamin's (2019, 45) concept of *thin description*, as a way of engaging in "fields of thought and action too often disconnected… tracing links between individual and institutional, mundane and spectacular, desirable and deadly..." I combine this approach with critical discourse analysis (Fairclough 1995; Van Leeuwen 2008) in order to make sense of the ways social actors gave meaning to contested social policies during the pandemic. This hinged on an understanding of texts as key sites of struggle, which reveal how competing ideologies contend for dominance at a given moment in time. While much of this data collection was done in isolation, most of the knowledge claims that emerged during this time were developed through dialogues with other members of the network.

In addition to this work, I collaborated with a coalition of researchers to organize a campaign against the propagation of racist and carceral research within the academy from May

to July of 2020. The thrust of this campaign was an open letter to an academic publisher, demanding that they rescind an offer to publish a study which claimed to identify or predict "criminality" using biometric and/or criminal legal data. This letter garnered over two thousand signatures from academics and social movement leaders, and sparked a broader conversation around the ways technology and data science fueled the criminalization of racialized communities. This project gave me the opportunity to think collectively with an interdisciplinary group of academics, combining historical analysis with notions of technical validity in order to formulate arguments that bridged the divide between humanistic inquiry and strictly technical debates. It was also an opportunity to experiment with forming an "otherwise set of relations" (Abdurahman 2021) amongst ourselves as academic workers, by centering an ethic of care (Hobart and Kneese 2020) in the way we collaborated and interacted with one another throughout the project. We also explored unconventional ways of forming inter-coalitional ties with abolitionist organizers operating outside of the academy, opening up new lines of communication and structures of accountability for our work.

Throughout all of the above projects, I maintained three kinds of fieldnotes, which enabled me to record and reflect on the complex process of "sense-making" that emerged from my fieldwork (Carstensen-Egwuom 2014). By writing fieldnotes that operated in different registers, I was able to capture a more nuanced and transparent account of my research process. The first set of notes were written in my fieldwork diary, which I used to record the details of the events I participated in and observed, including snippets of direct quotes and large chunks of paraphrased conversations, as well as background observations regarding the context of specific interactions (i.e. what a judge was wearing during our interview, how specific field relationships were established, etc.). The second set of notes were what I call "logistical notes," which

included technical documentation for the various projects described above (i.e. database schemas, annotated data dictionaries), as well as notes from conversations with collaborators regarding the implementation details of a given project. The last set of notes consisted of "personal reflections," where I tried to capture my evolving emotional state throughout the research, including the doubts, anxieties, and pleasures that I encountered while doing my fieldwork.

In many ways, this last set of field notes has become the most important and generative source of empirical data for me, even though I did not set out to record these experiences in the beginning. This is not uncommon in qualitative research — other scholars have written about how these types of "out of category" data emerged through unsatisfactory attempts at writing up their research from traditional field notes (St. Pierre 1997). I started writing personal reflections about a year into my fieldwork, after I struggled to shake a lingering sense of anxiety that I felt after a conversation with a community organizer and mentor of mine. In our conversation, he pointed out some potential pitfalls of my "insider" research project at the administration of the courts, including specific ways that the research might be harmful to the people he worked with. I came away from that meeting with greater clarity on how I wanted to move forward with the work, but I also felt anxious about how I might renegotiate the terms of my collaboration with my counterparts in the state administration of the courts. After writing up my observational notes from the conversation, I opened up a separate document to work through what exactly I was feeling and why. This opened up a number of interesting lines of inquiry regarding the ways that power and access to court data shaped the types of questions that I was pursuing in my research and the methods I used to explore them.

Over time, I grew to appreciate the ways that my personal reflections enabled me to pay close attention to the 'micropolitics of research' at all stages of the work, in order to place myself in the same critical plane as my subject matter (Bhavnani 1993). This laid the groundwork for me to incorporate reflexivity as a central aspect of my research project, rather than as an afterthought. Reflexivity involves an ongoing reflection on the ways that power, privilege, and oppression shape the way we see and understand the world (Pillow 2003). As a research practice, the goal of reflexivity is to produce more transparent and honest accounts of the research process, by interrogating the ways that one's positionality as a researcher shapes the construction of knowledge (Pillow 2003; Hertz 1997). At its worst, critical reflexivity becomes a rote exercise of listing one's membership in various social groups. Such boilerplate disclosures often act as a kind of ablution for privilege, rather than as an entry point into more thoughtful conversations regarding the ways we might critically engage with a given piece of research. As Van Maanen (2011, 99) notes, "confessions, endlessly replayed, begin to lose their novelty and power to inform."

But at its best, critical reflexivity can serve as a kind of compass for researchers and their audience, one that can help to uncover unavoidable tensions and generative contradictions throughout the research process. As Keyes and Austin (2022, 10) argue, it is especially important to reflect on the emotional stakes of technological research processes, "not only the sense of injustice experienced by marginalized actors…but the sense of discomfort or defensiveness this is likely to prompt in developers themselves." By making such discomforts explicit, we are able to surface, reflect on, and address the tacit understandings of the world that so powerfully shape the design of sociotechnical systems. Moreover, such reflections can be a great resource for forming hypotheses, particularly with regard to how the inner life of researchers shapes their

external and professional goals, and the ways that they go about pursuing them (Carstensen-Egwuom 2014).

At the heart of such an approach is an appreciation for what Shotwell (2011, xvi–xx) terms moments of "affective shock," when "the implicit may be visible at sites of a certain rupture in habitual activity…moments of strong emotion and unpremeditated reaction." By embracing a "reflexivity of discomfort" (Pillow 2003) that pivots on such shocks, I aim to grapple with the unsettling messiness of doing engaged research, rather than seek out a tidy and transcendent end-point to the conversation. My personal reflections have been an essential practice and resource for me in the process of exploring the ways that computational researchers like myself might unlearn hegemonic modes of thinking, feeling, and being, in order to re-invent a different, reparative role for ourselves and our work.

Given this approach, much of the writing in this dissertation takes on a distinctly auto-ethnographic style, in which I write about events and insights which have shaped the way I approach my research. One key aspect of this work is to critically examine the types of collaborations I have forged over the years, acknowledging that my collaborators are people with their own desires and agendas, and exploring the trade-offs of working with different types of people as I navigated issues of access and trust across various power differentials. Over time, I have become increasingly interested in experimenting with more relational forms of ethnographic writing, ones that would enable me to center specific relationships in my analysis and form arguments as part of an ongoing dialogue, rather than as a definitive endpoint to the discussion. I became particularly intrigued by Laurence Ralph's (2020) notion of "ethnographic lettering," which he describes as both a mode of ethnographic writing and a research

methodology that enables him participate in "an 'unfinished' and ever-evolving dialogue" with the people and groups involved in the social problems he studies (2020, 194).

In his book, *The Torture Letters*, Ralph structures his chapters as letters addressed to various people whom he encounters during the research process. These letters give him the chance to reframe his research subjects as interlocutors and the primary audience for the knowledge he produces. It also enables him to build in accountability for the claims he makes in his work. As Ralph (2020, 197) describes, "Letter writing forced me out of the typical, discipline-imposed shell of objectivity, making me account for myself as a speaking person, a person who helped fashion and develop a set of ideas that sometimes demanded a reply." Moreover, letter writing can create avenues for us to engage in unfinished conversations, to "'write back' to relatives, institutional agents and future relations," whom we might not otherwise have easy access to in the present (Cisneros 2018, 192).

I was drawn to this idea of letter writing as a relational practice, one that invited ongoing conversation and accountability for the arguments that I made. So I decided to experiment with incorporating this mode of writing in Chapter 6, by writing a letter to my main collaborator at the administration of the courts, where we attempted to build a judge dashboard. I have given this letter to my collaborator and invited her to respond in whatever format she finds most appropriate. In many ways, this process looks similar to the "member checks" (Reason and Rowan 1981, 248) that I have conducted throughout the writing process, inviting collaborators and interview subjects to review my tentative findings and offer feedback on my thinking. While I have intentionally carved out time and space to conduct member checks at various stages of the writing process, such checks have also occurred organically. One of the benefits of forging long-term relationships with many of my collaborators and research subjects is that my work is

continuously being shaped by our ongoing conversations. Whenever I need to gut check an aspect of my analysis or verify a fact from my fieldnotes, I am often able to reach out to the relevant people in real time to ask for their input.

This is especially true for the network of organizers whom I've been working alongside for the past five years. Given the ongoing and dynamic nature of many of the issues that we have worked on together (i.e. pretrial reform, data criminalization, etc.), we have had many opportunities to check in with one another and make sense of new developments when they occur. For example, over the past month I've had multiple correspondences over email with two organizers (one based in Massachusetts and one based in California) about a case study that I write about in Chapter Four of this dissertation. There was recently a new development in this case study, and these two individuals reached out to get my read on the situation and brainstorm possible responses. This gave me the opportunity to revisit the case study and update my thinking, as well as check in on other aspects of our work together.

For example, at the time, I was drafting Chapter Six of this dissertation and had a few places in the text which I'd flagged as needing additional verification from one of the collaborators who'd reached out over email. We were able to discuss both the recent developments in the case study mentioned above, as well as revisit events and policies that we hadn't discussed for years. This correspondence led to an exchange of some useful supplementary documents that I had not previously had access to, which is currently shaping the arguments I am developing for Chapter Six. Given the significant impact these interactions have had on my analysis, I do my best to engage in feminist citational practice, recognizing that "new knowledge claims are rarely worked out in isolation from other individuals and are usually developed through dialogues with other members of a community" (Hill Collins 1990, 260).

In addition to one-on-one member checks, I have had ample opportunity to triangulate my data by drawing from multiple perspectives throughout the research process. By conducting interviews and collaborating with people who occupy various positions within the landscape of criminal legal reform, I have been able to surface complexity and tensions in the ways people perceive and experience the same set of events (Vogl, Schmidt, and Zartler 2019). This has enabled me to not only identify patterns, but points of divergence in my data. For example, when developing insights into the ways mass pretrial incarceration is fueled by courtroom culture, I was able to triangulate data from my interviews with judges, observations from working with courtroom administrators, and the informal conversations that I had with organizers working on similar issues around the country.

By drawing on this diverse set of data points, I was able to surface the 'sometimes-similar and sometimes-contradictory tellings' (Santoro 2014, 127) of a given set of events or the impacts of a specific courtroom intervention. Moreover, I was able to surface diverging views and multiple truths, which speak to the ways violent systems are maintained, even in the face of criticism. Such divergences can offer insight into the affective logic and moral economy underpinning carceral societies. By examining the ways that carceral technologies are framed as vehicles for compassion and care by some, and violent tools of exclusion by others, we can unpack the ways such technology serves as a battleground for broader forms of culture change. The end result of all this is the set of working concepts outlined in this dissertation. Working concepts are not so much a set of top-down best practices as they are a loose set of ideas that are in a state of constant revision. My hope is that these working concepts will resonate with others who are on their own journey to cultivate liberatory relationships through and beyond technology.

Works Referenced

Abdurahman, Khadijah. 2021. “A Body of Work That Cannot Be Ignored.” *Logic Magazine,* December 25, 2021. https://logicmag.io/beacons/a-body-of-work-that-cannot-be-ignored/.

American Civil Liberties Union. 2020. “COVID-19 Model Finds Nearly 100,000 More Deaths Than Current Estiimates, Due to Failures to Reduce Jails.” American Civil Liberties Union. https://www.aclu.org/sites/default/files/field_document/aclu_covid19-jail-report_2020-8_1.pdf.

Barr, William. 2020. “Prioritization of Home Confinesment As Appropriate Response to COVID-19 Pandemic.” Office of Attorney General.

Benjamin, Ruha. 2019. *Race after Technology: Abolitionist Tools for the New Jim Code*. Medford, MA: Polity.

Bhavnani, Kum-Kum. 1993. “Tracing the Contours.” *Women’s Studies International Forum* 16 (2): 95–104. https://doi.org/10.1016/0277-5395(93)90001-P.

Bowker, Geoffrey, Stefan Timmermans, and S.L. Star. 1997. “Infrastructure and Organizational Transformation: Classifying Nurses’ Work.” In *Information Technology and Changes in Organizational Work,* 344–70. London, UK: Chapman and Hall.

Carstensen-Egwuom, Inken. 2014. “Connecting Intersectionality and Reflexivity: Methodological Approaches to Social Positionalities.” *Erdkunde* 68 (4): 265–76.

Chappell, Dale. 2020. “More Than Half of Chicago’s COVID-19 Cases Linked to Cook County Jail | Prison Legal News.” *Prison Legal News,* October 2020. https://www.prisonlegalnews.org/news/2020/oct/1/more-half-chicagos-covid-19-cases-linked-cook-county-jail/.

Christin, Angèle. 2017. “Algorithms in Practice: Comparing Web Journalism and Criminal Justice.” *Big Data & Society* 4 (2): 205395171771885. https://doi.org/10.1177/2053951717718855.

Cifor, Marika, P. Garcia, T.L. Cowan, J. Rault, Tonia Sutherland, A. Chan, J. Rode, Anna Lauren Hoffmann, N. Salehi, and Lisa Nakamura. 2019. “Feminist Data Manifest-No.” 2019. https://www.manifestno.com/home.

Cisneros, Nora Alba. 2018. “‘To My Relations’: Writing and Refusal toward an Indigenous Epistolary Methodology.” *International Journal of Qualitative Studies in Education* 31 (3): 188–96. https://doi.org/10.1080/09518398.2017.1401147.

Delgado, Richard, and Jean Stefancic, eds. 2013. *Critical Race Theory: The Cutting Edge*. Third edition. Philadelphia, Pennsylvania: Temple University Press.

Dewey, John. 1931. “The Development of American Pragmatism.” *Philosophy and Civilization,* 3–13.

Dourish, Paul, and Graham Button. 1998. “On "Technomethodologyn: Foundational Relationships Between Ethnomethodology and System Design.” *Human Computer Interaction* 13: 395–432.

Fairclough, Norman. 1995. *Critical Discourse Analysis: The Critical Study of Language*. Language in Social Life Series. London ; New York: Longman.

Gilmore, Ruth Wilson. 2007. *Golden Gulag: Prisons, Surplus, Crisis, and Opposition in Globalizing California*. American Crossroads. Berkeley: University of California Press.

Goldkuhl, Göran. 2012. “Pragmatism *vs* Interpretivism in Qualitative Information Systems Research.” *European Journal of Information Systems* 21 (2): 135–46. https://doi.org/10.1057/ejis.2011.54.

Gusterson, Hugh. 1997. “Studying up Revisited.” *PoLAR*, 114.

Hertz, Rosanna, ed. 1997. *Reflexivity & Voice*. Thousand Oaks, Calif: Sage Publications.

Hill Collins, Patricia. 1990. *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment*. Perspectives on Gender, v. 2. Boston: Unwin Hyman.

Hobart, Hi‘ilei Julia Kawehipuaakahaopulani, and Tamara Kneese. 2020. “Radical Care.” *Social Text* 38 (1): 1–16. https://doi.org/10.1215/01642472-7971067.

Hooks, Gregory. 2020. “COVID-19 Spread Faster in Counties with Large Prisons - and to Nearby Counties: Marion County Is a Disturbing Example.” December 2020. https://www.prisonpolicy.org/reports/covidspread_marion.html.

Irani, Lilly, and Khalid Alexander. 2021. “The Oversight Bloc.” *Logic Magazine*, December 25, 2021. https://logicmag.io/beacons/the-oversight-bloc/?utm_source=Haymarket+Newsletter&utm_campaign=e36c62639b-EMAIL_Newsletter_2017_11_20_HOLIDAY1_COPY_01&utm_medium=email&utm_term=0_a36ffbc74a-e36c62639b-333136034&mc_cid=e36c62639b&mc_eid=46c155fae6.

Keyes, Os, and Jeanie Austin. 2022. “Feeling Fixes: Mess and Emotion in Algorithmic Audits.” *Big Data & Society* 9 (2): 205395172211137. https://doi.org/10.1177/20539517221113772.

McGlotten, Shaka. 2016. “Black Data: No Tea, No Shade: New Writings in Black Queer Studies,” 262–86.

Ortner, Sherry B. 1995. “Resistance and the Problem of Ethnographic Refusal.” *Comparative Studies in Society and History* 37 (1): 173–93. https://doi.org/10.1017/S0010417500019587.

Pillow, Wanda. 2003. “Confession, Catharsis, or Cure? Rethinking the Uses of Reflexivity as Methodological Power in Qualitative Research.” *International Journal of Qualitative Studies in Education* 16 (2): 175–96. https://doi.org/10.1080/0951839032000060635.

Ralph, Laurence. 2020. *The Torture Letters: Reckoning with Police Violence*. Chicago ; London: The University of Chicago Press.

Reason, P, and J Rowan. 1981. “Issues of Validity in New Paradigm Research.” In *Human Inquiry: A Sourcebook of New Paradigm Research*, 77:239–62. New York: John Wiley & Sons, Inc. https://www.cambridge.org/core/product/identifier/S000305540024853 3/type/journal_article.

Ribes, David. 2019. “STS, Meet Data Science, Once Again.” *Science, Technology, & Human Values* 44 (3): 514–39. https://doi.org/10.1177/0162243918798899.

Santoro, Ninetta. 2014. “Using a Multiple Perspectives Framework: A Methodological Approach to Analyse Complex and Contradictory Interview Data.” *Ethnography and Education* 9 (2): 127–39. https://doi.org/10.1080/17457823.2013.839387.

Shotwell, Alexis. 2011. *Knowing Otherwise: Race, Gender, and Implicit Understanding*. University Park, Pa: Pennsylvania State University Press.

Souleles, Daniel. 2018. “How to Study People Who Do Not Want to Be Studied: Practical Reflections on Studying Up.” *PoLAR: Political and Legal Anthropology Review* 41 (S1): 51–68. https://doi.org/10.1111/plar.12253.

St. Pierre, Elizabeth Adams. 1997. “Methodology in the Fold and the Irruption of Transgressive Data.” *International Journal of Qualitative Studies in Education* 10 (2): 175–89. https://doi.org/10.1080/095183997237278.

Suchman, Lucille Alice. 1987. *Plans and Situated Actions: The Problem of Human-Machine Communication*. Cambridge [Cambridgeshire] ; New York: Cambridge University Press.

Van Cleve, Nicole Gonzalez. 2016. *Crook County: Racism and Injustice in America’s Largest Criminal Court*. Stanford, California: Stanford Law Books, an imprint of Stanford University Press.

Van Leeuwen, Theo. 2008. *Discourse and Practice: New Tools for Critical Discourse Analysis*. Oxford Studies in Sociolinguistics. Oxford ; New York: Oxford University Press.

Van Maanen, John. 2011. *Tales of the Field: On Writing Ethnography*. 2nd ed. Chicago Guides to Writing, Editing, and Publishing. Chicago: University of Chicago Press.

Villenas, Sofia. 1996. “The Colonizer/Colonized Chicana Ethnographer: Identity, Marginalization, and Co-Optation in the Field.” *Harvard Educational Review* 66 (4): 711–31.

Vogl, Susanne, Eva-Maria Schmidt, and Ulrike Zartler. 2019. “Triangulating Perspectives: Ontology and Epistemology in the Analysis of Qualitative Multiple Perspective Interviews.” *International Journal of Social Research Methodology* 22 (6): 611–24. https://doi.org/10.1080/13645579.2019.1630901.

Chapter 3

Study up: Shifting the Algorithmic Gaze to Scrutinize Power

Introduction

Months before I enrolled in the PhD program in Media, Arts, and Sciences at the Media Lab, my future academic advisor and current boss, Joi Ito, reached out to me and some of my colleagues at MIT, inviting us to join a meeting with Mustafa Suleyman. By then, we were used to meeting with industry heavy hitters like Suleyman, the co-founder of DeepMind[5] and founding co-chair of the Partnership on AI (PAI), an industry-founded think tank whose stated goal is to provide "actionable guidance" on ethical AI (Partnership on AI n.d.). Joi liked to surround himself with what he called, "troublemakers," or people who weren't afraid to talk back to the powerful crowd of men and women with whom he spoke on a daily basis. My colleagues and I formed a motley crew of junior scholars from a wide range of disciplines, but there was one theme that united us all — a growing concern about corporate-led initiatives on "AI ethics."

In the weeks leading up to the meeting with Suleyman, our group raised serious concerns regarding the role that the Partnership in AI was trying to play in emerging legislation. The Partnership had recently drafted a policy document that was intended to guide the regulation and implementation of pretrial risk assessments, focusing primarily on "minimum requirements for responsible deployment," such as "validity and data sampling bias, bias in statistical predictions;…and post-deployment evaluation" (Partnership on AI 2019). By then, our group's research had led us to believe that the flaws with pretrial risk assessment were fundamental, and could not be remedied through minor technical fixes. We were concerned that the Partnership's

[5] Deep Mind is an AI startup acquired by Google for $500 million in 2014.

recommendations would eclipse more fundamental critiques from grassroots organizations such as the Movement for Black Lives, the National Bail Fund Network, and the Immigrant Youth Coalition, who had argued that risk assessments created feedback loops that fueled racial profiling and mass incarceration (Lo 2021; Community Justice Exchange 2019).

As one colleague explained in an email exchange with Joi, we were concerned that PAI's statement would "validate the use of RA [risk assessment] by emphasizing the issue as a technical one that can therefore be solved with better data sets, etc." I agreed, arguing that the "PAI statement is weak and risks doing exactly what we've been warning against, re: the risk of legitimation via these industry led regulatory efforts." Moreover, we argued that the Partnership's association with academic and non-profit institutions like MIT and the ACLU ran the risk of whitewashing Big Tech's role in developing systems that were used to harm vulnerable and marginalized groups.

Joi took our comments in stride, acknowledging the ways that industry-led efforts like the Partnership on AI might circumscribe the conversation around AI ethics in unhelpful ways. In response to our concerns, he wrote to my colleagues and me, saying "I do know, however, from speaking to Mustafa when he was setting up PAI that he was meaning for the group to be much more substantive and not just 'white washing.' I think it's just taking the trajectory that these things take."

In spite of our misgivings, we did eventually attend the meeting with Suleyman. I viewed such meetings as an opportunity to shape the direction of an important new field of research. We'd been talking with Joi about abolition and epistemic harm, ideas that didn't organically come up in polite conversation, or even rigorous debates, with Silicon Valley elites. We hoped

that through him, we might influence the broader conversation from the top down. And for a while, it looked like we were onto something. Joi had begun to talk about the stakes of AI in different terms, leaning on us heavily as he prepared to give talks in high-profile venues regarding the limits of the emerging field of AI ethics.[6]

But then, in the summer of 2019, everything shifted, as reports emerged of Joi accepting and then hiding financial ties with Jeffrey Epstein, a billionaire charged with child sex trafficking (Farrow 2019). Not long after these revelations came to light, Joi resigned from multiple roles at MIT and Harvard, as well as the boards of several prominent organizations, including the MacArthur Foundation, the Knight Foundation, and the New York Times Company. Soon after, the extent of Epstein's entanglements with the academy came to light, raising deeper questions around the ways that the financier had laundered his reputation as a serial child abuser by cultivating an alternative persona as a high-profile benefactor of cutting-edge scientific research (Oreskes 2020).

These revelations pushed my colleagues and I to fundamentally re-examine our relationship with Joi and, more generally, the top-down approach we'd embraced to influencing the conversation on AI ethics. We couldn't help but see the parallels between Joi's relationship with Jefferey Epstein, and our relationship with Joi. Like Epstein, Joi had benefited from his association with us, because we helped him to maintain a radical edge to his persona as a leader in AI ethics, all while he deepened his ties to Silicon Valley elites (Ochigame 2019). This left us

[6] For example, we helped him to prepare a talk for Harvard's *Science and Democracy Forum* in which he challenged "the promotion of voluntary "responsible practices" over enforceable regulations; the reduction of complex epistemological concerns to questions of "bias"; the attempt to settle political disputes with algorithmic formalisms of "fairness." (Harvard University n.d.) This language came directly from our conversations with him as we tried to reframe the terms of the debate.

with an unsettling question — had we whitewashed Joi's reputation the same way that he'd whitewashed Jeffrey Epstein's?

For one of my friends and colleagues, Rodrigo Ochigame, this revelation was enough for him to disengage from contemporary work in "AI ethics" altogether. He reflected on our relationship with Joi in an essay, quoting extensively from the conversations I mentioned above, in order to illustrate the ways that "Silicon Valley's vigorous promotion of 'ethical AI' has constituted a strategic lobbying effort, one that has enrolled academia to legitimize itself" (Ochigame 2019). Rodrigo's essay marked his final intervention in the conversation on AI ethics, as he decided to refocus his efforts completely on his brilliant dissertation project, exploring what he terms the "informatics of the oppressed" (Ochigame 2021).

I respected Rodrigo's decision to leave, but I wasn't sure that it was the right choice for me. For one thing, I had a number of projects still in the pipeline that I still thought were important and wanted to pursue. And I was interested in exploring what it might look like to resist corporate capture, beyond just walking away. Rodrigo was an invaluable thought partner for me during these troubling times. One of the most important take-aways I had from our conversations was the idea that this moment was not necessarily unique to the field of "AI ethics"— most academic disciplines had been forced to grapple with their entanglements with oppressive power structures at some point in time. Rodrigo encouraged me to seek out historical precedents that might help me make sense of the present moment. And so I decided to look for other Come to Jesus moments in the academy, times when scholars had been pushed to expand their conceptual and methodological toolkits in order to navigate the fraught socio-political terrain that shaped academic research.

It was during this process that I came across Project Camelot, a Cold War initiative whose stated goal was to enlist social scientists in research that would "make it possible to predict and influence politically significant aspects of social change in the developing nations of the world" (Horowitz 1967). The U.S. government earmarked $6 million dollars for the project, which would have made it the largest social science project in U.S. history, had the project not been shut down during the planning stages in 1965. Project Camelot marked a key moment of ascendance for the social sciences, which had historically occupied a junior status within the academy as compared to the natural sciences (Lemore, Bell, and Collins 1983). A key driver of this rise in status was a growing interest from government officials in using social science research to predict and control geopolitical events during the Cold War (Solovey 2001).

For many social scientists, Project Camelot represented an unprecedented opportunity to participate in a large-scale research project with abundant resources. For the "action intellectuals" of the time, the project offered a powerful chance to both produce first-rate science, while simultaneously solving some of the world's most pressing political challenges (Solovey 2001). Even the name, Camelot, was inflected with idealistic overtones. As the director of the Special Operations Research Office (SORO), a branch of the Army's Psychological Warfare office, explained, the name Camelot was in reference to the Arthurian legend, which he claimed was about ''the development of a stable society with domestic tranquility and peace and justice for all" (Solovey 2001, 182).

Yet, for some scholars, Project Camelot represented a grave threat to the integrity of their field. One such person was Johan Galtung, a Norwegian sociologist who had worked extensively in Latin America. Galtung was in Chile when he received a letter from the U.S. military, inviting him to attend a planning session for Project Camelot in August 1965 (Horowitz 1967). Galtung

declined the invitation and handed the letter over to a Chilean colleague, who then publicly confronted a representative of Project Camelot at a meeting in Santiago. The letter was published in the Chilean media soon after, causing an uproar from government officials across the globe, who considered the project to be essentially a "spying mission" for the U.S. military (Horowitz 1967).

Project Camelot was shut down not long after, but its ripple effects were felt for many years to come. Amid mounting anti-war sentiment and growing criticism regarding the role of social scientists in U.S. domestic programs, such as the War on Poverty, American scholars were pushed to examine the institutional arrangements which shaped their research agendas both at home and abroad (Solovey 2001). During this time, the social sciences underwent what Peter Novick (1988) terms an "epistemological revolution," in which academics challenged the idealized image of the social sciences as an objective and value-neutral endeavor.

Scholars began to have more open conversations around the ways that money shaped their research agendas, arguing that the relatively unlimited funding of Project Camelot had caused their peers to "not look a gift horse in the mouth," or dig too deep into the motives behind the U.S. military's sponsorship of their work (Horowitz 1967). Scholars like Galtung (1967) emphasized the asymmetrical nature of Project Camelot's research agenda — why had the study focused exclusively on 'counter-insurgency,' while giving no attention to the concept of 'counter-intervention,' or the way Latin American countries might prevent the United States from meddling in their affairs abroad? Such questions pushed social scientists to grapple with the ways that funding, prestige, and influence had shaped the questions they deemed worth asking and those that they'd chosen to ignore.

I heard echoes of my own experience in those early reflections on Project Camelot, written more than half a century prior. Like Project Camelot, the field of AI ethics had experienced a serious influx in funding and prestige. Cross institutional partnerships between the private sector, government agencies, and the academy meant that the scope of the problems researchers were invited to solve were large and sexy, many of them premised on the idea that advanced research methods could help those in power to better predict and control future outcomes. But, like Project Camelot, a recent slew of high-profile scandals was now raising serious concerns about the ways the academy's entanglements with powerful agendas might perpetuate harm under the guise of benevolent intentions.

I decided to track down the work of young scholars from this period of time, in the hopes that they might offer me a compass for how to navigate my current situation. That's when I came across a letter written by the anthropologist Laura Nader, a newly minted professor at UC Berkeley. Nader had penned a letter to the American Anthropological Association (AAA) in 1965, expressing grave concerns about the state of her field. In her letter, Nader argued that "masquerading as scholar-anthropologists to collect data that may be used against the people studied is doing research under false pretenses, is exploiting the image of the anthropologist for purposes of military considerations" (Nader 2020, 24). A few months later, an official from the AAA responded, assuring Nader that "There seems to me little doubt that anthropologists on a whole would object to their employment in file research under false pretenses, such as those you describe" (Nader 2020, 26). But Nader continued to ruminate on the Camelot controversy for several years to come. In 1969, she eventually formalized some of those thoughts in an article entitled "Up the Anthropologist: Perspectives Gained from Studying Up."

In this seminal article, Nader called for her fellow anthropologists to re-examine the default assumptions of their field, specifically the tendency to study exoticized cultures and downtrodden populations in distant lands (Nader 1972). She urged her peers to extend their field of study "beyond the ghetto," specifically to include powerful institutions and cultures. Nader's use of the term "ghetto" was intentional – a ghetto is defined as an isolated place, segregated from larger social structures. The call to study up was a call to contextualize traditional field sites in terms of their relationships to broader institutions and cultures of power, rather than as isolated cultural alcoves. For example, anthropologists might study banks and colonial administrations, or networks of white-collar crime, all of which created the preconditions necessary for specific marginalized and peripheral subcultures to emerge in the first place.

I'd actually read Nader's article before the Epstein scandal. The text had come up in conversations with some of my collaborators, as we tried to make sense of the unarticulated assumptions underlying the policy debate surrounding pretrial risk assessment. But Nader's words took on a new gravity as I read them against the backdrop of both the Camelot and Epstein controversies. If I replaced the words "anthropologist" with "data scientist" in the text, the arguments felt as if they could have been written yesterday, as relevant and useful as they'd been fifty years earlier. Nader helped me think about the ways that the political economy of academic research powerfully shapes what topics scholars consider important to study, and which topics we choose to turn a blind eye toward. Her writing helped me to make sense of how the partialities of cutting-edge research could whitewash the violence of powerful actors by creating ideological ghettos around marginalized groups, divorcing their plights from broader institutions and cultures of power.

As I drew parallels between Project Camelot and the present, I grew increasingly interested in interrogating the tendency for data scientists to cast the algorithmic gaze downward, similar to the ways Nader had done in the field of anthropology several decades prior. By putting my recent experiences in direct conversation with Nader's work, I'd have the opportunity to explore how the political economy of AI ethics shaped the default orientations of the field, and how an uncritical acceptance of such defaults normalized social hierarchies and legitimized violence against marginalized groups.

The following article was co-written with my collaborators —Colin Doyle, Karthik Dinakar, and JB Rubinovitz. In it, we build from Nader's work to argue that the field of algorithmic fairness would greatly benefit from reorienting its efforts "upward," by refocusing the algorithmic gaze on institutions and individuals which occupy powerful and privileged positions in society. Moreover, we frame studying up as a research method that helps to excavate the unnamed power dynamics that shape the discourse surrounding "AI ethics." When we "study up," it sheds light on the political economy of data science, by creating frictions which reveal the often-invisible forces which shape what questions get asked and how we go about answering them. Such a reorientation opens up new and interesting lines of inquiry, by surfacing ethical gray areas related to access, trust, and permission. These challenges are underexplored in the current literature, which is largely based on the premise that most data science work will focus on observing and influencing the behaviors and outcomes of people with relatively less power.

Ultimately, the goal of this article is to provide a useful analytic for orienting oneself in conversations regarding the potential uses of data to address tough social problems. By "studying up," computational practitioners can surface and then challenge unarticulated assumptions about

what (and who!) needs fixing, creating new avenues of inquiry that break free from the knee-jerk tendency to study marginalized groups in isolation from larger structures of power and privilege.

Works Referenced

Community Justice Exchange. 2019. “An Organizers Guide to Confronting Pretrial Risk Assessment Tools in Decarceration Campaigns.” https://static1.squarespace.com/static/60db97fe88031352b829d032/t/60dcd6e6ac9dae69c2243313/1625085675244/CJE_PretrialRATGuide_FinalDec2019Version.pdf.

Farrow, Ronan. 2019. “How an Élite University Research Center Concealed Its Relationship with Jeffrey Epstein | The New Yorker.” *The New Yorker*, September 6, 2019. https://www.newyorker.com/news/news-desk/how-an-elite-university-research-center-concealed-its-relationship-with-jeffrey-epstein.

Galtung, Johan. 1967. “Scientific Colonialism.” *Transition*, no. 30 (April): 10. https://doi.org/10.2307/2934342.

Harvard University. n.d. “Science & Democracy: ‘The Limits of Ethical A.I.’ with Joichi Ito.” Accessed March 22, 2023. https://environment.harvard.edu/science-democracy-limits-ethical-ai-joichi-ito.

Horowitz, Irving Louis. 1967. “The Life and Death of Project Camelot.” *Ethnographic Fieldwork: An Anthropological Reader*, 277–87.

Lemore, S. Dale Mc, Daniel Bell, and Randall Collins. 1983. “The Social Sciences since the Second World War.” *Social Forces* 61 (3): 914. https://doi.org/10.2307/2578145.

Lo, Puck. 2021. “From Data Criminalization to Prison Abolition.” Community Justice Exchange. https://abolishdatacrim.org/en/about.

Nader, Laura. 1972. “Up the Anthropologist: Perspectives Gained from Studying Up.”

———. 2020. *Laura Nader: Letters to and from an Anthropologist*. Ithaca [New York]: Cornell University Press.

Novick, Peter. 1988. *That Noble Dream: The “Objectivity Question” and the American Historical Profession*. 1st ed. Cambridge University Press. https://doi.org/10.1017/CBO9780511816345.

Ochigame, Rodrigo. 2019. “How Big Tech Manipulates Academia to Avoid Regulation.” *The Intercept*, December 20, 2019. https://theintercept.com/2019/12/20/mit-ethical-ai-artificial-intelligence/.

———. 2021. “Remodeling Rationality: An Inquiry Into Unorthodox Modes of Logic and Computation.” Massachusetts Institute of Technology. https://dspace.mit.edu/handle/1721.1/140189.

Oreskes, Naomi. 2020. “Jeffrey Epstein’s Harvard Connections Show How Money Can Distort Research.” Scientific American. September 1, 2020. https://doi.org/10.1038/scientificamerican0920-84.

Partnership on AI. 2019. “Report on Algorithmic Risk Assessment Tools in the U.S. Criminal Justice System.” Partnership on AI. file:///Users/chelsea.barabas/Downloads/Report-on-Algorithmic-Risk-Assessment-Tools.pdf.

———. n.d. “About Us: Advancing Positive Outcomes for People and Society.” Partnership on AI. Accessed March 22, 2023. https://partnershiponai.org/about/.

Solovey, Mark. 2001. “Project Camelot and the 1960s Epistemological Revolution: Rethinking the Politics-Patronage-Social Science Nexus.” *Social Studies of Science* 31 (2): 171–206. https://doi.org/10.1177/0306312701031002003.

Studying Up: Reorienting the study of algorithmic fairness around issues of power[7]

ABSTRACT

Research within the social sciences and humanities has long characterized the work of data science as a sociotechnical process, comprised of a set of logics and techniques that are inseparable from specific social norms, expectations and contexts of development and use. Yet all too often the assumptions and premises underlying data analysis remain unexamined, even in contemporary debates about the fairness of algorithmic systems. This blind spot exists in part because the methodological toolkit used to evaluate the fairness of algorithmic systems remains limited to a narrow set of computational and legal modes of analysis. In this paper, we expand on Elish and Boyd's (2018) call for data scientists to develop more robust frameworks for understanding their work as situated practice by examining a specific methodological debate within the field of anthropology, frequently referred to as the practice of "studying up". We reflect on the contributions that the call to "study up" has made in the field of anthropology before making the case that the field of algorithmic fairness would similarly benefit from a reorientation "upward". A case study from our own work illustrates what it looks like to reorient one's research questions "up" in a high-profile debate regarding the fairness of an algorithmic system – namely, pretrial risk assessment in American criminal law. We discuss the limitations of contemporary fairness discourse with regard to pretrial risk assessment before highlighting the insights gained when we reframe our research questions to focus on those who inhabit positions of power and authority within the U.S. court system. Finally, we reflect on the challenges we have encountered in implementing data science projects that "study up". In the process, we surface new insights and questions about what it means to ethically engage in data science work that directly confronts issues of power and authority.

INTRODUCTION

A key aim of the algorithmic fairness community is to develop frameworks and standards to evaluate algorithmic systems against social and legal principles such as equity, fairness, and justice. Initial efforts to grapple with the ethical and social implications of algorithmic systems have come in the form of technical "fairness criteria" – computational formalisms used to define and evaluate complex legal concepts such as "disparate impact," "equal opportunity," and "affirmative action" (Dwork et al. 2012; Feldman et al. 2015; Hardt, Price, and Srebro 2016).

---

[7] *This article is co-authored with Colin Doyle, Karthik Dinakar and JB Rubinovitz. It was originally published in the 2020 proceedings of the ACM FAccT* Conference in Barcelona, Spain.*

Preferred citation*:* Barabas, Chelsea, Colin Doyle, J. B. Rubinovitz, and Karthik Dinakar. "Studying Up: Reorienting the Study of Algorithmic Fairness around Issues of Power." Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 167–176. https://doi.org/10.1145/3351095.3372859.

But recent scholarship in the field of algorithmic fairness has begun to point out the limitations of this approach. The fairness of a data science project extends far beyond the technical properties of a given model and includes normative and epistemological issues that arise during processes of problem formulation (Passi and Barocas 2019), data collection and preparation (Selwood 2002), claims-making (Barabas 2020; Ochigame 2019) and everyday use in applied contexts (Selbst and Barocas 2018). Scholars have argued that narrow technical conceptualizations of algorithmic fairness elide more fundamental issues and, in the process, run the risk of legitimizing harmful practices based on fundamentally unsound truth claims about the world (Barabas 2020; Benjamin 2019; Elish and Boyd 2018; Stark and Hoffmann 2019).

In light of these concerns, some scholars have suggested that designers of algorithmic systems embrace a process-driven approach to algorithmic justice, one that explicitly recognizes the active and crucial role that the data scientist plays in constructing meaning from data (Costanza-Chock 2020; Elish and Boyd 2018; Greene, Hoffmann, and Stark 2019; Selbst and Barocas 2018). Researchers have problematized common- place characterizations of data science as an objective and neutral process (Kerssens 2019; Stark and Hoffmann 2019), arguing that data and their subsequent analyses are always the by-product of socially contingent processes of meaning making and knowledge production (Benjamin 2019; Jasanoff 1994; Passi and Barocas 2019; Stark and Hoffmann 2019). Rather than strive to develop narrow technical solutions to the issue of "fairness," scholars have recently encouraged the algorithmic fairness community to draw broader boundaries around what they conceive of as an "algorithmic system," to include social actors and key contextual considerations (Selbst and Barocas 2018).

Research within the fields of sociology, anthropology, and science, technology and society (STS) has long characterized the work of data science as a sociotechnical process, comprised of a

set of logics and techniques that are inseparable from specific social norms, expectations and contexts of development and use (Benjamin 2019; Elish and Boyd 2018; Eubanks 2018). But all too often the assumptions and premises underlying data analysis remain unexamined, even in contemporary debates regarding the fairness of algorithmic systems. This blindspot has appeared in part because the conceptual and methodological toolkit used to evaluate the fairness of algorithmic systems remains limited to a narrow set of computational and legal modes of analysis.

To grapple with the full social and ethical implications of data-intensive algorithmic systems, we must develop more robust methodological frameworks that enable data scientists to reflect on how their research practices and design choices influence and distort the insights generated from their work. To this end, Elish and Boyd (2018) have called for researchers to reconceptualize data science as a form of "computational ethnography," whereby data scientists actively participate in the production of partial and situated knowledge claims. Like ethnographers, data scientists "surround themselves with data (‘a field site’), choose what to see and what to ignore, and develop a coherent mental model that can encapsulate the observed insights" (Elish and Boyd 2018, 20). Elish and Boyd argue that reconceptualizing data science as a novel form of qualitative inquiry opens up a new set of methodological frameworks that can guide data-driven practices of modeling the world and illuminate the ways that power shapes and operates through the work of data science.

The algorithmic fairness community can benefit greatly from ongoing methodological debates and insights gleaned from fields such as anthropology and sociology, where scholars have been working for decades to develop more robust frameworks for under- standing their work as situated practice. In this paper, we expand on Elish and Boyd’s (2018) call for the

development of more reflexive data science practices, by examining a specific methodological debate within the field of anthropology, frequently referred to as the practice of "studying up." In what is now considered a classic anthropological text, Laura Nader (1972) called for her fellow anthropologists to expand their field of inquiry to include the study of elite individuals and institutions, who remained significantly under- examined in the anthropological cannon. Rather than study exotic cultures in far-flung lands, Nader appealed for a critical repatriated anthropology that would shed light on processes of exploitation and domination by refocusing the anthropological lens on the cultures of the powerful. As Nader (1972, 7) argued, "If one's pivot point is around those who have responsibility by virtue of being delegated power, then the questions change."

This paper is organized into two parts. In part one, we reflect on the contributions that the call to "study up" has made to the field of anthropology. Nader's provocation came at a time when anthropologists were grappling with the epistemological and methodological limits of their tradition. The call to "study up" was a call for scholars to move beyond their default orientations – their tendency to study the "underdog" in isolation from larger structural forces – in order to deal directly with issues of power and domination in their work. We draw parallels between this debate in anthropology and similar issues that data scientists are grappling with today in their pursuit of "fair, accountable, and transparent" algorithmic systems. We then make the case for why the field of algorithmic fairness would benefit from such a reorientation "upward." The political and social impacts of algorithmic systems cannot be fully understood unless they are conceptualized within larger institutional contexts and systems of oppression and control. Data science projects that "study up" could lay the foundation for more robust forms of accountability and a deeper understanding of the structural factors that produce undesirable

social outcomes via algorithmic systems.

In part two, we present a case study from our own work as a group of interdisciplinary researchers from the fields of computer science, sociology and law. This case study illustrates how research questions can be reoriented "up" in contemporary discourse regarding the fairness of algorithmic systems. For the past two years, we have been deeply engaged in the public and academic debate regarding the use of pretrial risk assessment as a means of bail reform in the United States. Pretrial risk assessment has become one of the prototypical examples of the ethical stakes of contemporary algorithmic decision-making regimes. But the ethical debate regarding these tools has by-and-large uncritically accepted the premise that the best way to address mass pretrial incarceration is by modeling and forecasting the risk of some of the most marginalized and disempowered factions of American society – individuals arrested and charged with a crime.

We discuss the limitations of contemporary fairness discourse regarding pretrial risk assessment before illustrating the insights gained when we reframe our research questions to focus on those who inhabit positions of power and authority within the U.S. court system. Finally, we reflect on the challenges we have encountered in implementing data science projects that aim to shift the gaze "upward." In the process, we develop new insights and questions about what it means to ethically engage in data science work that directly confronts issues of power and authority in the field. To do this, we draw on a feminist tradition of rigorously examining the "micropolitics of research" to unpack the ways that our positionality as researchers who are interacting with powerful institutions impacts the production of specific knowledge claims in our work (Bhavnani 1991). Our hope is that in doing so, we will expand the conversation about ethical engagement in data science and open up new lines of inquiry that, until now, have been

left unexamined by the algorithmic fairness community.

THE CALL TO "STUDY UP"

1.1 Anthropology

In her 1972 article, "Up the Anthropologist: Perspectives Gained from Studying Up," Laura Nader called on her fellow anthropologists to move beyond the study of people and cultures at the peripheries of Western society (who comprised the bulk of the anthropological canon), in favor of "studying up," to excavate the ways that power operates through elite institutions and positions of authority. As Peters and Wendland (2016) argue, "studying up" indicates a relative direction between the researcher and subject, to study someone who has what Edward Said (1978) referred to as "the relative upper hand" in terms of the amount of agency and authority they have in a given context (Peters and Wendland 2016) .

Nader's provocation came at a time when the social sciences were undergoing what historians have called "the epistemological revolution of the 1960's", whereby the predominant post-war model of social science as an objective and value-neutral enterprise (modeled after the natural sciences) was called into question (Novick 1988). These developments occurred amidst high-profile controversies, such as Project Camelot, which brought to light the various ways that scholarly practice in the social sciences was deeply intertwined with larger social and political agendas, both at home and abroad (Solovey 2001; Novick 1988) . These movements pushed anthropologists to critically examine the political nature of their work and to reflect on the blind spots they had developed while working in a dominant tradition that generally preferred studying "the underdog."

While the call to study up was a particularly influential provocation, Nader was part of a larger movement within anthropology in the late 1960's and early 1970's, which called for the

discipline to develop methods and theories that directly addressed issues of power and domination in the modern era. As Peters and Wendland (2016) have discussed, scholars like Berreman (1968, 392) "challenged anthropologists to recognize that they were 'involved whether they wish it or not' in the international power plays of the Cold War...[and] to recognize that we [anthropologists] are involved in dynamics that systematically privilege some people and render others suspect." Others argued that anthropology risked irrelevance if it continued to neglect powerful actors and institutional settings in their analysis (Gjessing 1968). Furthermore, scholars questioned the ways anthropologists framed their field sites as isolated islands of culture, arguing that modern society should increasingly be studied as a single, interdependent social system (Gough 1968).

It was in this context that Nader (1972, 1) urged her colleagues to study "the most powerful strata" of society, arguing that "the quality of life and our lives themselves may depend upon the extent to which citizens understand those who shape attitudes and actually control institutional structures." She called into question anthropologists' widely held predilection for studying "the downtrodden," subjects who were frequently conceptualized as members of isolated cultures in distant lands. Rather, she argued that these phenomena must be understood in terms of relationships that extend "beyond the ghetto," to include powerful institutions and cultures. "Studying up" might include, for example, the study of banks and colonial administrations, or networks of white-collar crime, all of which create the preconditions necessary for specific marginalized and peripheral subcultures to emerge in the first place. Nader's use of the term "ghetto" was intentional – a ghetto is defined as an isolated place, segregated from larger social structures. The call to study up was a call to contextualize traditional field sites in terms of their relationships to broader institutions and cultures of power, rather than as isolated cultural

alcoves.

Nader foresaw various cultural, ethical and methodological challenges to studying up. Researchers face structural barriers in accessing field sites where cultures of the elite and powerful can be observed. Studying up often requires breaking through networks of gatekeepers and navigating delicate negotiations regarding the purpose and authority of the research (Hertz and Imber 1995; Marcus 1983; Priyadharshini 2003). As Priyadharshini (2003, 423) notes, "the issue was not so much the difficulty in gaining entry, but of not gaining it on my own terms. Access to such sites often depends on how well researchers can negotiate their research aims, their methods, and their very identities with gatekeepers."

Barriers to access give rise to new ethical challenges, ones which upset the typical assumptions which underlie ethical frameworks intended to reduce abuses of power in anthropological research. At what point does partial disclosure of one's research aims or personal worldview become deception? Can researchers forge relationships under one premise (i.e. getting a job at the field site) while simultaneously carrying out ethnographic research? How does one navigate issues of self-censorship, particularly when the disclosure of previously hidden information might preclude access for the researcher, or other research colleagues, in the future?

While reflecting on the challenges researchers face in studying up, Nader (1972, 20) argues that "we should not necessarily apply the same ethics developed for studying the private, and even ethics developed for studying in foreign cultures (where we are guests), to the study of institutions, organizations, bureaucracies that have a broad public impact." This requires the researcher to venture into unknown territory, where the rules of the game are often ambiguous and no consensus exists about the best approach. In light of these challenges, it is absolutely critical for ethnographers of the powerful to develop novel methods for gaining access to subjects who

have the power to resist outside scrutiny. It also requires researchers to cultivate strong practices of reflexivity and to track the effects of such complexities on the knowledge claims that they ultimately produce. In the decades since Nader's initial provocation, anthropologists have developed a body of literature that reflects on the methodological and ethical complexities of studying up. These reflections on fieldwork offer the research community the opportunity to dig into the micro-politics of their work and grapple with the ways that their conclusions are shaped by the methods available to them and their positionality as researchers.

We share this brief history of "studying up" for two purposes. First, we believe that the field of algorithmic fairness would greatly benefit from reorienting efforts "upward," thereby challenging de- fault framings and assumptions that tend to cast the algorithmic gaze on the relatively poor and marginalized, in an attempt to model their behavior and characteristics. In the following section, we draw parallels between contemporary discourse on algorithmic fairness and the debates Nader and her colleagues were engaged in more than half a century ago. We then make a call for "studying up" in the field of data science, arguing that this reorientation is critical if we aspire to deploy "data science for social good."

Second, we build from Elish and Boyd's (2018) call for data scientists to expand their methodological toolkit to include strong practices of reflexivity. We posit that engaging in ethical data science requires moving beyond narrow technical solutions or definitions of fair- ness, to cultivate a rigorous process-driven approach to the aim of algorithmic fairness. We present a case study based on our work as an interdisciplinary group of researchers who have reoriented our research "upward" in a domain that has been widely used as a case study in the algorithmic fairness community – pretrial risk assessments. In the second half of this article, we draw from the anthropological tradition of writing reflections from our time in the field, surfacing

challenges we faced while negotiating access to data that would enable us to shift the algorithmic gaze upward. In doing so, our aim is to demonstrate the value of engaging in this type of reflective practice, as a means of producing more ethical outcomes in data science work more broadly.

1.2 Data Science

What can the field of data science learn from Nader's call to "study up" in anthropology? Like the anthropologists of the 1960's and 70's, data scientists today have been confronted by a series of high profile controversies that illustrate the various ways that their work is intertwined with larger political and social struggles (Guynn 2015; Hill 2012). These controversies have given rise to an influential community of researchers from both academia and industry who have formed a new regulatory science (Jasanoff 1994) under the rubric of "fair, accountable, and transparent algorithms" (Diakopoulos et al. 2017; Lepri et al. 2018). In addition, a growing body of critical scholarship seeks to problematize the notion of data science as an apolitical and benevolent enterprise, whereby data scientists develop technocratic solutions to complex social problems by uncovering "objective truths" found in the data (Benjamin 2019; Hoffmann 2019; Stark and Hoffmann 2019). But the reality is that data scientists still lack the methodological tools necessary to critically engage with the epistemological and normative aspects of their work.

As a result, data scientists tend to uncritically inherit dominant modes of seeing and understanding the world when conceiving of their data science projects. As various scholars have argued, such an uncritical acceptance of default assumptions inevitably leads to discriminatory design in algorithmic systems by reproducing ideas which normalize social hierarchies and legitimize violence against marginalized groups (Benjamin 2019; Hoffmann 2019). Discriminatory design does not require intentional malice or prejudice on the part of the data

scientist. As Benjamin (2016, 148) explains, "One need not harbor any racial animus to exercise racism... rather, when the default settings have been stipulated, simply doing one's job...is enough to ensure the consistency of white domination over time."

Similarly, D'Ignazio and Klein (2020) have characterized data science as an inherently conservative force, one which tends to reproduce logics and worldviews which maintain and legitimize the status quo. As they argue, data science "is characterized by extremely asymmetrical power relations, where those with power and privilege are the only ones who can actually collect the data but they have overwhelming incentives to ignore the problem, precisely because addressing it poses a threat to their dominance" (D'Ignazio and Klein 2020, 29).

Like the anthropologists of the mid-twentieth century, the default tendency is for data scientists to cast their gaze "downward," to focus on the relatively poor and powerless factions of society. This tendency is particularly widespread amongst projects which self-identify under the rubric of A.I. or data science "for social good." As Hoffmann (2019) has pointed out, data scientists tend to study disadvantage in a one-dimensional way, divorced from the normative conditions which produce complementary systems of advantage and privilege. She argues that the myopic focus on disadvantaged subjects pushes data scientists into a stance of patronizing benevolence, whereby they seek to understand the plight of those "relegated to the 'basement' of the social hierarchy" solely in terms of their own behaviors, relationships and pathologies (Hoffmann 2019, 11).

This downward orientation holds widespread appeal, because it creates discursive ghettos around marginalized populations via statistical discourse in ways that disconnect their plights from structural forms of oppression. In this way, today's data scientists have much in common with the anthropologists of the 60's and 70's, who struggled to develop more sophisticated ways of

contextualizing their field sites in terms of larger structural forces through which power and domination were exercised to maintain the status quo.

Nader's mandate to "study up" was a call for her colleagues to deal directly with issues of power and domination in their work. It's time for a similar provocation to be made within the field of data science. Data science projects which reorient themselves around "studying up" could lay the foundation for more robust forms of accountability and deeper understandings of the structural factors that produce undesirable social outcomes via algorithmic systems. Re-orienting the field of data science upward requires us to ask different questions. It requires us to develop a critical reflex when presented with opportunities to build models based on data collected by powerful institutions. It requires a new set of reflective practices which push the data scientist to examine the political economy of their research and their own positionality as researchers working in broken social systems. Data scientists who "study up" could provide tremendous benefit to society.

Like Nader, we also recognize that such a reorientation comes with its own set of challenges – challenges for which there are no easy solutions or quick technical fixes. These challenges can't be modeled with mathematical equations or resolved with clear-cut right or wrong answers. They require rigorous reflection on the ways that our design choices are shaped by external social forces. In the following sections we present a case study of our attempts to study up as data scientists engaged in a highly contentious debate regarding the use of predictive algorithms as a vehicle for bail reform. We outline the insights we gained by reframing the issue of bail reform in terms of the behaviors and cultural practices of those who occupy positions of power and authority in the US court system. We then reflect on the challenges we encountered in pursuing data science projects which cast the lens up, as well as the strategies

we developed for overcoming barriers to the research.

CASE STUDY: BAIL REFORM

U.S. bail reform provides an ideal domain for examining the benefits of "studying up" in algorithmic discourse. Ongoing reforms address an urgent social issue, rely heavily upon data science and algorithmic solutions, and have spawned an important, albeit limited, academic discourse on the fairness, equity, and transparency of algorithmic interventions. Consistent with our analysis above, existing data science interventions for bail reform have accepted the default orientations of the criminal justice system, whereby algorithmic solutions cast the gaze "down" to evaluate the riskiness of marginalized populations facing prosecution and the threat of pretrial detention. This framing prevails even though judges are ultimately responsible for making the bail decisions that have led to the current crisis.

By "studying up" on judges and judicial culture, we aim to upset and expand current framings of both the challenges of bail reform and the ethical stakes of algorithmic systems in criminal justice. This section begins with a brief introduction to bail reform and the role that algorithmic interventions have played in addressing the problem of mass pretrial incarceration. We then review the limits of current ethical discourse regarding these tools before providing a description of our efforts to reframe this debate in our own work. Finally, we close with reflections on the challenges and opportunities we encountered when pursuing a "study up" approach to data science in this domain.

2.1 Pretrial Algorithms and Limits of the Current Ethics Debate

The United States faces a crisis of mass incarceration. Current levels of incarceration defy all international and historical norms. Pretrial incarceration – the jailing of people awaiting trial who have been accused but not convicted of crimes – is a key driver of these extreme

incarceration rates. On any given day, American jails imprison nearly half a million people before their trial (Minton and Zeng 2015). Increases in pretrial incarceration rates are "responsible for all of the net jail growth in the last twenty years" (Wagner and Sawyer 2018). More than 30 years ago, the Supreme Court affirmed that "liberty is the norm, and detention prior to trial or without trial is the carefully limited exception" ("United States v Salerno" 1987). But these days the exception has become the rule. Today in America, there are more legally innocent people in jail than there were convicted people in jails and prisons in 1980 (Beck and Gilliard 1995). The harms of this system fall disproportionately on communities of color, as Black and Latinx people are more likely to be detained pretrial than similarly situated white people (Demuth and Steffensmeier 2004).

America's unique reliance on money bail has allowed for the spike in pretrial incarceration. Most states have procedures through which a judge can intentionally detain someone pretrial. Judges often skirt these procedures because they are purposefully burdensome for the government and only permit the detention of a small subset of defendants. It's much easier, but not constitutionally permissible, for judges to impose an unaffordable money bond on whomever they'd like to detain.

In recent years, bail reform has gained political salience with the public and both political parties. At the heart of the money bail system is an obvious injustice: before trial, the rich can go free while the poor stay in jail. As lawsuits combating the money bail system have proliferated across the country, states and counties are looking to stay ahead of legal liability and adopt bail reform on their own.

Nearly every jurisdiction that has attempted bail reform in re- cent years has moved to adopt pretrial risk assessment. The rapid proliferation of pretrial risk assessment has been propelled by

a specific, though often unarticulated, logic. Judges are concerned about releasing individuals who might flee the jurisdiction or commit a violent crime if released. But judges struggle to accurately identify those individuals who pose a true risk. As a result, judges incarcerate far too many people because they overestimate pretrial risk. This presents an opportunity for data science to help judges to distinguish "signal from noise" when making time sensitive decisions. Algorithms may more accurately predict a person's risk of flight or violence, thereby reducing pretrial incarceration rates. If a jurisdiction uniformly adopts risk assessments, then pretrial decisions could be more consistent and less susceptible to the prejudices of individual judges.

In recent years, the fairness and efficacy of pretrial risk assessments have been called into question. A high-profile exposé by Propublica argued that, not only are algorithmic risk assessments not very accurate, but the burden of that inaccuracy is disproportionately borne by historically marginalized groups, who are often subject to higher false positive rates (Angwin et al. 2016). This discrepancy in ac- curacy is usually talked about in terms of bias – critics argue that algorithmic tools run the risk of reproducing or amplifying pre- existing biases in the system. Propublica's reporting inspired a wave of scholarship attempting to define and measure fairness, equity, and bias within risk assessment instruments. This scholarship has evolved along two complementary tracks: 1) the development of formal fairness criteria that illustrate the trade-offs of different algorithmic interventions and 2) the development of "best practices" and managerialist standards for maintaining a baseline of accuracy, transparency and validity in algorithmic systems. These efforts include detailed technical formulas, lively debate over the correct or best measures of fairness, and explorations of how to achieve both procedural and substantive frameworks for maintaining fairness.

This technical debate might give the impression of a robust discourse regarding the fairness

of pretrial risk assessments. But the conversation is actually quite limited. The narrow focus on technical and managerial aspects of fairness fails to contend with key epistemological and normative assumptions that underpin the project of pretrial risk assessment and bail reform more broadly (Barabas 2020; Ochigame 2019). As scholars have pointed out, there are numerous challenges to modeling and effectively intervening on risks such as pretrial violence and flight (Barabas 2020; Gouldin 2016). Some of these challenges arise because pretrial violence and flight are exceedingly rare and hard to predict. To overcome this problem, most risk assessments make questionable empirical leaps when defining their outcome variables and inputs, leaps which could easily lead judges to overestimate pretrial risk and detain more people than is justified, while also exacerbating racial disparities (Barabas 2020; Ochigame 2019).

But more importantly, the current ethical discourse tacitly accepts the premise of pretrial risk assessment – namely, that the best or only way to reduce pretrial incarceration rates is by casting the algorithmic gaze "downward" to model the behavior of people who have been accused of a crime. A narrow technical focus on the fairest way to predict a pretrial defendant's behavior elides more fundamental issues about who to study, what to predict, and whose behavior to influence. At the stage of a bail hearing, pretrial defendants arguably have the least agency and power over whether they are incarcerated before trial. Rather than frame the issue of bail reform in terms of a judicial culture that permits excessive incarceration, current reform focuses algorithmic techniques exclusively on predicting the behavior of defendants. This discourse runs the risk of legitimizing harmful practices based on fundamentally unsound truth claims about the risk of pretrial defendants and the drivers of pretrial incarceration.

2.2 Reflections from the Field

As a group of multi-disciplinary researchers, we recognized the limitations of the current

discourse on algorithmic fairness as it pertains to pretrial risk assessment. Rather than engage in the fairness debate on its current terms, we sought to redefine the problem space by reorienting our work "up." To help us think through the best approach, we reached out to a group of community organizers working at the forefront of bail reform, who represent the largest organized resistance to risk assessments in general. Through a series of conversations and a one-day roundtable with academics and organizers, we began to develop a new vision for how data science could support a better approach to bail reform.

It became clear to us that missing from the ethical discussions of pretrial risk assessment is a sense of judges' agency and accountability. Studying up requires us to reframe the problem of mass pretrial incarceration in terms of the organizational context and courtroom cultures that have enabled pretrial detention rates to skyrocket. Mass pretrial incarceration represents a case of "institutional de- coupling" whereby the everyday practices of the courtroom diverge from state and federal laws regarding pretrial release (Christin 2017). From this angle, it seems perplexing that interventions seeking to change judicial behavior would focus exclusively on modeling the behavior of people awaiting trial, rather than on understanding how this decoupling emerged and persists. This perspective helped us develop an agenda for studying up: rather than focus on predicting the likelihood of a defendant failing, we would try to understand why American judges send so many people to jail, in spite of state and federal laws protecting against excessive pretrial detention. "Studying up" in the area of bail reform would require us to surface insights regarding past, present and future trends in the way judges make bail decisions.

It was not immediately apparent what the most effective approach might be for accessing data to support our work. We were unsure about which organizations or individuals to approach for collaboration, and it was unclear how we might go about acquiring data that would shed light

on judges' behavior.

The availability of datasets proved to be a tremendous challenge for this work. Many states do not publish or share court data. Some states share information about defendants, but rarely collect and share information about judges and other court actors. Most juris- dictions do not collect important information about bail hearings, including the amount of bail the prosecution and defense requested, the arguments that the attorneys made, the reasons why a judge imposed bail, or whether a defendant's ability to pay was assessed at any point in the process. For many places across the country, there is no available data on who has been detained pretrial. No jurisdiction tracks whether a person was detained because a judge set an unaffordable bail amount with the intent to detain that per- son or whether that person's pretrial detention was incidental or unintended. Nonetheless, some court data can be used to evaluate judge behavior, although it's often an incomplete patchwork. Courts do not collect data as a means of evaluating judge performance but by-and-large only collect data about judges incidentally, as a byproduct of court administration software.

In the end, we embraced what other scholars in the literature on studying up have termed an "eclectic" research strategy (Gusterson 1997; Souleles 2018). We developed a two-pronged approach to collecting and accessing data that could be used to study the behavior of judges. The first strategy was an "insider" approach to accessing official court records, whereby we negotiated access to court data through a collaboration with a state administrative agency. The second was an "outsider" approach to accessing and generating data that either did not exist or that the government would not release of its own accord. To this end, we collaborated with grassroots organizers who sought to generate their own data sets based on public court observation.

The Inside Approach: Negotiating access to government data.

Two years ago, we reached out to an administration of the courts (AOC) of a mid-sized state to negotiate access to data that could be used to evaluate judge behavior over a span of five years. The state had recently passed legislative reforms aimed at reducing their pretrial jail population, which included the adoption of a pretrial risk assessment tool and statutory guidelines for release. The AOC was interested in working with us to understand the various ways that judges had responded to these reform efforts.

In order to access this data, we held a series of meetings in which we pitched ideas for collaboration with the AOC. During these conversations it became clear that the AOC was interested in understanding the impacts of supervised conditions of release, such as electronic monitoring and mandatory drug testing, on pretrial outcomes such as missed court dates and re-arrest. Although this was beyond the scope of our original research question, we felt it was necessary for us to explore the topic in order to provide value to the AOC and therefore acquire the other data we needed regarding judge behavior. The AOC ultimately provided us with data concerning the pretrial decisions at the level of the individual judge. Our hope was that insights gleaned from this data could also inform the selection of judges for more in-depth qualitative interviews regarding their processes for making bail decisions. By selecting a diverse cross-section of judges to interview, our goal was to surface insights about judicial attitudes toward risk assessment and pretrial release more broadly.

As we began to delve into the data provided by the AOC, we were forced to grapple with the partiality of the court's data. The outcomes that were captured were only the outcomes that matter for court administration, such as a defendant missing a court date, being arrested, or being arrested for a violent crime. This limited our ability both to evaluate judges' bail decisions

and to measure the impact of supervised conditions of release. Judges ostensibly choose to impose pretrial interventions in lieu of incarceration to keep someone's life on track and avoid the negative consequences of pretrial detention. Research has shown that both pretrial incarceration and involvement in the criminal justice system more generally can destabilize people's lives, causing them to lose jobs, housing, custody of their children, and more. None of these outcomes appeared in this dataset.

These limitations on defendant outcome data ultimately caused us to decline the offer to study the impact of pretrial supervision on ethical grounds. Within the dataset, the impact of pretrial interventions or pretrial incarceration was only measured in terms of whether people showed up to court or got rearrested if released. This narrow data collection skews policy toward incarceration. Not only are the measures of negative effects of incarceration missing entirely, but incarceration is only measured in terms of its benefit. People in jail can't miss court dates and can't be rearrested. According to the data we were provided, a surefire way to improve pretrial results would be to incarcerate more people. The limited selection of outcome variables predetermined the conclusion of incarceration as an effective intervention, while simultaneously erasing the known harms of incarceration. Because we had no ability to measure other outcomes, we had to end our research on ethical grounds.

We had some anxiety about refusing to pursue the research question proposed to us by the courts. We did not want to miss an opportunity to inform an important conversation, in a way that would give us highly coveted access to a state's administrative data. But we later came to view this refusal as a generative act – it gave us the opportunity to have conversations with key decision makers from the courts about the limits and opportunities of the data that they had collected and to imagine alternative research questions that might be more suitable for this data.

These conversations allowed us to disrupt the default logics that tend to guide quantitative research in criminal justice. And it gave us the opportunity to discuss expanding the state's research agenda to account for the full set of harms and benefits of specific policies. All in all, the refusal of the initial research proposition opened up new research potentials that could be viewed as interventions in and of themselves. This course of action stands in stark contrast to the default approach that many researchers take when doing data science work, particularly with government entities and especially in criminal law.

Data scientists often uncritically accept the premises and framings of the research question given to them by actors in a specific domain. They abdicate responsibility for the social impacts of their work by characterizing their role as limited and technocratic (Green, Data science as political action). Rather than decline to carry out a study for which the data is an inappropriate or incomplete proxy for the outcome of interest, they provide a perfunctory acknowledgment of the imperfect nature of the data they use as proxies in the methods section of their papers and reports. In the case of our project, such an uncritical acceptance ran the risk of legitimizing harmful practices through the erasure of important outcomes related to public safety and crime reduction, such as housing, employment and family stability.

Researchers in the field of data science have begun to call for more careful consideration of the projects that researchers agree to take on, after thoughtful evaluation of the social context and the politics of their intervention (Selbst et al. 2019). These calls acknowledge that such careful consideration can lead to a decision to not move forward with the project. As Selbst et al. (2021, 13) have argued, "not all problems can or should be solved with technology. In such cases, taking ourselves out of the equation as a relevant social group requires and rewards researcher humility."

Yet few opportunities exist for researchers to share the insights gained from exploring and ultimately abandoning research ideas that they deemed harmful or inappropriate. This is an area where data science can benefit from conversations taking place in the field of anthropology, where the concept of "refusal" has been developed as a key methodological and theoretical framework for approaching fieldwork (Simpson 2007; McGranahan 2016; Tuck and Yang 2014; Ortner 1995). Simpson (2014) considers how refusal can structure possibilities, as well as introduce new discourses and subjects of research by shedding light on what was missed in the initial research proposal. As she argues, "There was something that seemed to reveal itself at the point of refusal-a stance, a principle, a historical narrative, and an enjoyment in the reveal" (Simpson 2014, 107). The field of data science should develop a robust practice of refusal whereby data scientists grow accustomed to critically engaging and reformulating the assumptions and epistemological leaps undergirding the research opportunities that they are offered.

Finally, similar to the anthropologists looking to "study up," we also found that access to our research subjects was limited and contingent. Judges reign supreme in their courtrooms and have not traditionally been subject to outside scrutiny or evaluations. During the course of our qualitative interviews with judges, one judge was suspicious of our motives and suspected that our research would impugn the character and professionalism of the judge and his colleagues. With one phone call to the administrative office of the courts, the judge succeeded in effectively canceling our study. Technically speaking, the study is still live, but it has been relegated to a low priority for the courts and our access to judges has been all but officially revoked.

Even the limited access that we were able to get can be largely attributed to our flexibility with location, the prestige of our academic institutions, and the open-ended nature of our funding. We were willing to study any jurisdiction in the United States that would offer us

access to data and the ability to interview judges. A court system can use its willingness to be studied by researchers from top universities to make claims to the public about the institution's accountability and commitment to excellence. Because lawyers and judges like to be associated with prestigious places, the names of our affiliated institutions may have opened doors that would be closed to other researchers. Our funding for this project was open-ended and not tied to a specific grant or work product, which gave us maneuverability and allowed us to try out multiple avenues for data collection. Few researchers have the benefit of such privileged circumstances in their work.

The Outside Approach: Creating "missing data sets".

Around the same time, we connected with activists and organizers who were working to access similar information about the behaviors of key decision makers in the courts. Across the country, community organizations have developed "court watch" programs to fill in the information that is missing from court datasets or not shared with the public. To shine a light on judicial behavior at bail settings, court watch volunteers observe bail hearings to collect both quantitative and qualitative data about how the proceedings were conducted and what decisions were made. We volunteered to assist one local court watching effort by providing technical assistance in the collection, cleaning, and analysis of court watch data.

Court watch programs can be very effective at reminding public officials that they are accountable, revealing to the public snap- shots of how our criminal courts operate, and uncovering common court practices that deviate from law and policy. But court watch programs have a limited capacity to generate datasets that can be useful for research. Collecting data about a vast bureaucracy takes a lot of time and is very work intensive. Courts are open five days a week and, in most urban places, judges make bail decisions every day. Court watch

programs rely entirely on volunteers to visit court, sit, and record their observations. It's nearly impossible to schedule shifts and locations to ensure that the sample data obtained is representative of the court system as a whole.

The quality of court watch data is not great. Not only does performance vary considerably from volunteer to volunteer, but the court process is fast, opaque, and complicated, which virtually guarantees that information is lost with real time transcription of events. An entire bail hearing, from attorneys' arguments to a judge's final decision, can happen in less than a minute, sometimes in as little as fifteen seconds. Much of the information about the case is not conveyed to the public in attendance but exists on paper for court personnel. The attorneys and judges use legal jargon and acronyms that may be unintelligible or unfamiliar to the volunteers observing court. In many states, cellphones and other electronic devices are prohibited from courtrooms. Volunteers cannot record the proceedings and can only record their findings on pen and paper. These barriers to access made it extremely challenging to collect data that met the minimum standards of quality necessary to be used for academic research purposes.

Perhaps surprisingly, we also encountered access challenges when working with community groups and bail funds. The prestige of our affiliations and credentials may open the doors of powerful institutions, but many grassroots organizations have a deep, warranted distrust of academics. For decades, the academy has played an integral role in propagating harmful data narratives, especially in criminal law (Muhammad 2011). Far too often scholars have used their power to influence local and national policy without consideration for the community groups that have done the hard labor of advocating for change and who are often denied a seat at the table. It took time to build trust and expand our role beyond a purely technical job of processing and cleaning data.

Speculative Practice: Judicial risk assessment.

Given the challenges we faced in studying judges directly and developing alternative datasets, we also engaged in a different type of practice to help denaturalize the assumptions that undergird pretrial risk assessment, by making judges the focus of predictive modeling. Following Ruha Benjamin's (2019) advice to construct a "sociotechnical imaginary that examines not only how the technical and social components of design are intertwined, but also imagines how they might be configured differently," we are developing a new risk assessment model.

Rather than predict the behavior of pretrial defendants, the judicial risk assessment tool predicts the behavior of judges. Our model generates a "failure to adhere" score based on a prediction of whether - for any given case - a judge will fail to adhere to the U.S. Constitution by imposing unaffordable bail without due process of law. The model's training data is a combination of the court data given to us by the AOC and demographic information about judges in the dataset that we collected through internet searches. Like existing risk assessments, demographic information drives the model. Just as age is the most predictive variable for a defendant being arrested while on pretrial release, age is the most predictive variable for a judge illegally incarcerating someone pretrial. The algorithm – a stochastic gradient boosting machine – exceeds the accuracy and ROC scores of pretrial risk assessments currently in use.

Our risk score draws direct parallels with existing risk assessment models on the market today. One of the key outcomes pretrial risk assessments attempt to model is whether a defendant will "fail to appear" for a future court date. This is frequently referred to as a defendant's "FTA" score. Mainstream FTA predictions achieve accuracy and ROC scores around 65% (DeMichele et al. 2018). Our risk assessment achieves 80% accuracy and a 79% ROC score when generating our alternative "FTA" prediction – of a judge's likelihood of "failing to adhere."

By traditional measures of accuracy and validity, our risk model far outstrips the performance of mainstream pretrial risk assessments. But this algorithm is not intended for practical use. Indeed, many of the technical and ethical problems with defendant risk assessments may also apply with a judicial risk assessment. But the project is more than just a parody or a gag. As a thought exercise and a counter-narrative, a risk assessment that "looks up" at judges can, as Benjamin (2019, 14) puts it, "[help us] better understand and expose the many forms of discrimination embedded in and enabled by technology." This imaginative work echoes the advice of the critical race scholar Derrick Bell (1994, 893), "To see things as they really are, you must imagine them for what they might be." Our goal in developing a judicial risk assessment was to render as intuitive the various ways that the methods and discourse surrounding risk assessment are limited and stigmatizing, by subjecting those in power to the very socio-technical processes which are typically reserved for only the poor and marginalized.

We conceived of this work as a collaboration with community groups, such as the Community Justice Exchange, that are at the forefront of local struggles over bail reform across the country. Our hope was that the judge risk assessment project could bolster their ongoing efforts to challenge the use of pretrial risk assessments as a vehicle for bail reform. We envisioned using the judge risk assessment as a starting point for larger conversations about the tenuous empirical assumptions underlying pretrial risk assessment, as well as the need for greater accountability for judge decisions regarding bail. But collaboration between academics and community requires patience and deference. Our vision of success has not always aligned with the near and mid-term goals of our partners on the ground. It has taken time to meaningfully incorporate our partner's feedback and adjust our timelines according to their shifting circumstances and strategies.

Constructing the judicial risk assessment also raises ethical concerns. Nothing in our memorandum of understanding with the administrative office of the courts prevents us from developing this project using court data. Nonetheless, it's highly unlikely that the state would have shared this data with us to build this kind of algorithm. And if the judicial risk assessment attracts enough notoriety, state courts may be even less likely to share data with us and other researchers in the future. This only raises more questions: of what value is this access if it is contingent upon refusing to question the unchecked assumptions and premises of the data regime itself.

CONCLUSION

Scholars in the field of algorithmic fairness are developing frameworks and standards to evaluate algorithmic systems against important social and legal principles. As social scientists have long argued, the fairness of a data science project extends far beyond the technical properties of a given model. Data science is a sociotechnical process, and the designers of algorithmic systems must embrace a process-driven approach to algorithmic justice that recognizes the active role that the data scientist plays in constructing meaning from data.

In this paper, we have expanded upon the call for data scientists to understand their work as situated practice by introducing the anthropological practice of "studying up." The case study from our own work with pretrial risk assessments illustrates what it looks like to reorient one's research questions about algorithmic fairness "up" at people and institutions in positions of power. This project is similar in spirit to other recent attempts to turn criminal law's algorithmic gaze upon the powerful, such as the Stanford Open Policing Project and the White Collar Crime Risk Zones. At the same time, this case study reveals many of the ethical and practical challenges of "studying up" in practice. Like anthropologists attempting to study up,

we encountered serious obstacles to gaining access to our research subjects. And the availability of meaningful datasets will continue to be a challenge for this kind of work. Across many domains, datasets reflect the mindset and values of the institutions that have the money and authority to collect and assemble this data. It will almost always be easier to study and scrutinize the marginalized and turn a blind eye to the powerful. Projects that question the ethics of existing data approaches will often require unearthing or assembling alternative datasets. Ethical data science will sometimes hinge on a politics of generative refusal, whereby the assumptions and logics undergirding a particular research problem is challenged and fundamentally reimagined in new terms. Data scientists and researchers will need patience, receptiveness, and humility to build trust with communities who, for too long, have been the objects, rather than the co-producers, of the algorithmic gaze. By turning the algorithmic lens on the powerful, researchers will face new ethical grey zones related to access, trust, and permission.

The field of algorithmic fairness needs to study up. Research in the field will be limited and distorted if it exclusively and uncritically accepts the data and values of powerful institutions. Deeply inquisitive study of algorithmic ethics will require creativity and resourcefulness. To generate insights into algorithmic equity and fairness, we must look beyond the existing data and their accompanying interpretations, to imagine the data for what it might be.

Works Referenced

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. "Machine Bias." *ProPublica*, May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Barabas, Chelsea. 2020. "Beyond Bias: Re-Imagining the Terms of 'Ethical Al' in Criminal Law." *Georgetown Journal of Law and Critical Race Theory* 12 (2): 83–112.

Beck, Allen J., and Darrell K. Gilliard. 1995. "Prisoners in 1994." 151654. Prisoners. Bureau of Justice Statistics. https://bjs.ojp.gov/library/publications/prisoners-1994.

Bell, Derrick. 1994. "Who's Afraid of Critical Race Theory?" *UNIVERSITY OF ILLINOIS LAW REVIEW* 1994 (4): 893–910.

Benjamin, Ruha. 2016. "Catching Our Breath: Critical Race STS and the Carceral Imagination." *Engaging Science, Technology, and Society* 2: 145–56.

———, ed. 2019. *Captivating Technology: Race, Carceral Technoscience, and Liberatory Imagination in Everyday Life*. Durham: Duke University Press.

Berreman, Gerald D. 1968. "Is Anthropology Alive? Social Responsibility in Social Anthropology." *Current Anthropology* 9 (5, Part 1): 391–96. https://doi.org/10.1086/200923.

Bhavnani, Kum-Kum. 1991. *Talking Politics: A Psychological Framing for Views from Youth in Britain*. European Monographs in Social Psychology. Cambridge [England] ; New York : Paris: Cambridge University Press ; Editions de la Maison des sciences de l'homme.

Christin, Angèle. 2017. "Algorithms in Practice: Comparing Web Journalism and Criminal Justice." *Big Data & Society* 4 (2): 205395171771885. https://doi.org/10.1177/2053951717718855.

Costanza-Chock, Sasha. 2020. *Design Justice: Community-Led Practices to Build the Worlds We Need*. Information Policy. Cambridge, Massachusetts: The MIT Press.

DeMichele, Matthew, Peter Baumgartner, Michael Wenger, Kelle Barrick, Megan Comfort, and Shilpi Misra. 2018. "The Public Safety Assessment: A Re-Validation and Assessment of Predictive Utility and Differential Prediction by Race and Gender in Kentucky." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3168452.

Demuth, Stephen, and Darrell Steffensmeier. 2004. "The Impact of Gender and Race-Ethnicity in the Pretrial Release Process." *Social Problems* 51 (2): 222–42. https://doi.org/10.1525/sp.2004.51.2.222.

Diakopoulos, Nicholas, Sorelle A. Friedler, Marcelo Arenas, Solon Barocas, Michael Hay, Bill Howe, H.V. Jagadish, and Kris Unsworth. 2017. "Principles for Accountable Algorithms and a Social Impact Statement for Algorithms :: FAT ML." In . https://www.fatml.org/resources/principles-for-accountable-algorithms.

D'Ignazio, Catherine, and Lauren F. Klein. 2020. *Data Feminism*. The MIT Press. https://doi.org/10.7551/mitpress/11805.001.0001.

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. "Fairness through Awareness." In , 214–26. ACM.

Elish, Madeleine Clare, and Danah Boyd. 2018. "Situating Methods in the Magic of Big Data and AI." *Communication Monographs* 85 (1): 57–80.

Eubanks, Virginia. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Publishing Group.

Feldman, Michael, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. “Certifying and Removing Disparate Impact.” In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259–68.

Gjessing, Gutorm. 1968. “The Social Responsibility of the Social Scientist.” *Current Anthropology* 9 (5): 397–402.

Gough, Kathleen. 1968. “New Proposals for Anthropologists.” *Current Anthropology* 9 (5): 403–35.

Gouldin, Lauryn P. 2016. “Distangling Flight Risk from Dangerousness.” *BYU L. Rev.*, 837.

Greene, Daniel, Anna Lauren Hoffmann, and Luke Stark. 2019. “Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning.” In .

Gusterson, Hugh. 1997. “Studying up Revisited.” *PoLAR*, 114.

Guynn, Jessica. 2015. “Google Photos Labeled Black People ‘Gorillas.’” *USA TODAY*, July 2015. https://www.usatoday.com/story/tech/2015/07/01/google-apologizes-after-photos-identify-black-people-as-gorillas/29567465/.

Hardt, Moritz, Eric Price, and Nati Srebro. 2016. “Equality of Opportunity in Supervised Learning.” *Advances in Neural Information Processing Systems*, 3315–23.

Hertz, Rosanna, and Jonathan B. Imber, eds. 1995. *Studying Elites Using Qualitative Methods*. Sage Focus Editions 175. Thousand Oaks, Calif: Sage Publications.

Hill, Kashmir. 2012. “How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did.” 2012. https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/.

Hoffmann, Anna Lauren. 2019. “Where Fairness Fails: On Data, Algorithms, and the Limits of Antidiscrimination Discourse.” *Under Review with Information, Communication, and Society*.

James, Andrea. 2021. “Andrea James | Facebook.” June 11, 2021. https://www.facebook.com/andrea.james.336.

Jasanoff, Sheila, ed. 1994. *Learning from Disaster: Risk Management after Bhopal*. Law in Social Context Series. Philadelphia: Univ. of Pennsylvania Press.

Kerssens, Niels. 2019. “De-Agentializing Data Practices: The Shifting Power of Metaphor in 1990s Discourses on Data Mining.” *Journal of Cultural Analytics*. https://doi.org/10.22148/16.037.

Lepri, Bruno, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. 2018. “Fair, Transparent, and Accountable Algorithmic Decision-Making Processes: The Premise, the Proposed Solutions, and the Open Challenges.” *Philosophy & Technology* 31 (4): 611–27. https://doi.org/10.1007/s13347-017-0279-x.

Marcus, George. 1983. *Elites, Ethnographic Issues.* Univ of New Mexico Press.

McGranahan, Carole. 2016. “Theorizing Refusal: An Introduction.” *Cultural Anthropology* 31 (3): 319–25. https://doi.org/10.14506/ca31.3.01.

Minton, Todd, and Zhen Zeng. 2015. “Jail Inmates at Midyear 2014.” NCJ-248629. Bureau of Justice Statistics.

Muhammad, Khalil Gibran. 2011. *The Condemnation of Blackness*. Harvard University Press.

Nader, Laura. 1972. “Up the Anthropologist: Perspectives Gained from Studying Up.”

Novick, Peter. 1988. *That Noble Dream: The "Objectivity Question" and the American Historical Profession*. 1st ed. Cambridge University Press. https://doi.org/10.1017/CBO9780511816345.
Ochigame, Rodrigo. 2019. "The Illusion of Algorithmic Fairness." Cambridge, Mass.
Ortner, Sherry B. 1995. "Resistance and the Problem of Ethnographic Refusal." *Comparative Studies in Society and History* 37 (1): 173–93. https://doi.org/10.1017/S0010417500019587.
Passi, Samir, and Solon Barocas. 2019. "Problem Formulation and Fairness." In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 39–48. Atlanta GA USA: ACM. https://doi.org/10.1145/3287560.3287567.
Peters, Rebecca Warne, and Claire Wendland. 2016. "Up the Africanist: The Possibilities and Problems of 'Studying up' in Africa." *Critical African Studies* 8 (3): 239–54. https://doi.org/10.1080/21681392.2016.1244945.
Priyadharshini, Esther. 2003. "Coming Unstuck: Thinking Otherwise about 'Studying up'"." *Anthropology & Education Quarterly* 34 (4): 420–37.
Said, Edward W. 1978. *Orientalism*. London: Routledge and Kegan Paul.
Selbst, Andrew D., and Solon Barocas. 2018. "The Intuitive Appeal of Explainable Machines." *Fordham Law Review* 87 (3): 1085–1139. https://doi.org/10.2139/ssrn.3126971.
Selbst, Andrew D., Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. "Fairness and Abstraction in Sociotechnical Systems." In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68. Atlanta GA USA: ACM. https://doi.org/10.1145/3287560.3287598.
Selwood, Sara. 2002. "The Politics of Data Collection: Gathering, Analysing and Using Data about the Subsidised Cultural Sector in England." *Cultural Trends* 12 (47): 13–84. https://doi.org/10.1080/09548960209390330.
Simpson, Audra. 2007. "On Ethnographic Refusal: Indigeneity, 'Voice' and Colonial Citizenship," 14.
———. 2014. *Mohawk Interruptus: Political Life across the Borders of Settler States*. Durham: Duke University Press.
Solovey, Mark. 2001. "Project Camelot and the 1960s Epistemological Revolution: Rethinking the Politics-Patronage-Social Science Nexus." *Social Studies of Science* 31 (2): 171–206. https://doi.org/10.1177/0306312701031002003.
Souleles, Daniel. 2018. "How to Study People Who Do Not Want to Be Studied: Practical Reflections on Studying Up." *PoLAR: Political and Legal Anthropology Review* 41 (S1): 51–68. https://doi.org/10.1111/plar.12253.
Stark, Luke, and Anna Lauren Hoffmann. 2019. "Data Is the New What? Popular Metaphors & Professional Ethics in Emerging Data Culture." *Journal of Cultural Analytics*. https://doi.org/10.22148/16.036.
Tuck, Eve, and K. Wayne Yang. 2014. "R-Words: Refusing Research." In *R-Words: Refusing Research*. 1 Oliver's Yard, 55 City Road London EC1Y 1SP: SAGE Publications, Inc. https://doi.org/10.4135/9781544329611.
"United States v Salerno." 1987. https://www.law.cornell.edu/supremecourt/text/481/739.
Wagner, Peter, and Wendy Sawyer. 2018. "Mass Incarceration: The Whole Pie 2018." 2018. https://www.prisonpolicy.org/reports/pie2023.html.

Chapter 4

Refusal: Getting at the "Code Beneath the Code"

Introduction

At the 2019 Conference on Neural Information Processing Systems (NeurIPS), one of the most well-respected computer science conferences in the world, the opening panel discussion on A.I. for social good didn't go quite as people might have expected.

"I'm not usually in spaces like this, and I'm not entirely convinced that I haven't surreptitiously walked into a terrorist den," began Sarah T. Hamid (2019), one of the core members of the Carceral Tech Resistance Network. As Sarah explained, "like terrorists, technologists in spaces like this have a concept of what social good is. … Everybody thinks that they're doing good things in the world … that's the scary thing. I don't know how it is that we have a disconnect between kids in cages and the work that's happening in spaces like this."

There was a smattering of applause from the audience as others shifted uncomfortably in their seats. Most of the morning had been devoted to presentations on how computer science might help tackle some of the world's toughest problems, ranging from online content curation to supporting the UN's Sustainable Development Goals. Throughout those talks, the line between problem and solution had been clear and uncomplicated. Now Sarah was pushing the audience to think more deeply about the ways their efforts might contribute to the very problems they claim to be solving with technology.

I watched on from the middle of the crowd as the moderator asked Sarah how she thought computer scientists should handle competing definitions of social good, citing a laundry list of mathematical formalisms that had been developed to grapple with ideas like "fairness" in A.I. Again, Sarah responded with an answer that sent ripples of anxious murmurs through the room.

"I think that's a very interesting question," she said, "I just don't think it's an important question." Sarah went on to describe a multiyear campaign led by the Stop LAPD Spying Coalition to dismantle Operation LASER, a predictive policing program that the Los Angeles Police Department used to deploy officers to "hot spots" for violent crime and gang-related activity. She told the audience of the painstaking efforts that the coalition undertook to learn about and fight against Operation LASER. "Meanwhile, seven people have died in LASER zones," she explained, "… while we're trying to understand how to defend our community against these data-driven programs … people are dying in these LASER zones." Sarah used this example to illustrate how abstract debates about social good were out of touch with the very real harms that communities face in their daily lives as a result of data-driven technologies. If computer scientists really hoped to make a positive impact on the world, Sarah argued, then they would need to start asking better questions.

Interactions like this can be deeply unsettling, especially for people who strive to use their time and talents to build technology to make the world a better place. But over the last few years, conversations like this have become foundational to my work. For example, during the early stages of my research, I reached out to the Massachusetts Bail Fund to learn more about their work. During our first meeting, my colleagues and I offered up a number of ideas on ways we might use data collected by the bail fund to understand different pretrial outcomes. At the time, our primary goal was to develop less biased techniques for forecasting the risk of people awaiting trial. The organizers from the bail fund were rather stoic about our proposals. There was a long pause after we finished our presentation and I wondered if we'd done something to offend them.

Eventually, Atara Rich-Shea, the executive director of the bail fund, leaned in and shared her frank perspective. She said the bail fund was frequently contacted by academics who were only interested in asking their own questions and that, for the most part, those questions were harmful to the people who she served. Atara went on to explain the ways that academics undermine the work of movements for liberation by asking questions that either siphoned people into categories of "deserving and undeserving," erased the violence of incarceration, or distracted from more pressing issues. Her colleagues also challenged some of the language we'd been using throughout the conversation, pointing out the ways that such language reinforced harmful assumptions about the people that they worked with. Ultimately, the bail fund was not interested in collaborating with us if we were not invested in building a shared vocabulary and approach to understanding the harms and potential uses of data in contentious debates such as bail reform.

Atara's comments were the beginning of many more fruitful conversations about how our group might ask better questions and set up structures to ensure that our work was accountable to the communities that are directly impacted by data-driven technologies. Over time, not only did the quality of our research questions improve, but we became increasingly capable of identifying and un-doing harmful tendencies in our research. A key aspect of this process involved turning down offers to collaborate on projects that were premised on flawed or harmful assumptions. Such work was challenging and it often left me wondering if the academy was really the place I wanted to invest in as a long-term home. What was the point of getting a PhD, I wondered, if the most useful thing I could do was to decline projects on offer? It was easy to slip into the mindset of "If I don't do this bad study, someone else will do it even worse than me."

Several months before the NeurIPS conference, I spoke with Sarah Hamid about the growing sense of frustration I felt with my work. I had recently turned down an opportunity to

carry out a study that, a year prior, I would have been eager to do. Though there were good reasons to not pursue that particular research question, I couldn't shake a lingering sense of disappointment. It felt like I was missing out on an opportunity to engage in an important policy conversation. And I wasn't exactly sure that declining to participate in such studies would make a difference. My efforts to reframe similar research conversations with potential collaborators had been met with mixed results. Emails went unanswered. Polite nods and pregnant silences filled the room. Doors were closed. And, perhaps most importantly, I didn't have a framework for processing the insights that I was learning from these interactions.

On our call, I told Sarah that I felt like there was no space within the academy to explore the knowledge that was produced through the process of deciding *not* to study something. That's when Sarah pointed me to Audra Simpson's work on refusal. Once again, I had the experience of reading work from outside the mainstream milieu of "AI ethics" that deeply resonated with my experiences working in the field. In her book *Mohawk Interruptus,* Simpson (2014) describes multiple unsettling moments that she experienced during her fieldwork. As a member of the Mohawk nation, Simpson was interested in exploring how members of her community navigated fluid notions of membership and belonging, outside of official, state-based definitions of citizenship. She'd experienced firsthand the complex ways her fellow tribe members forged a sense of community and inclusion through what she called "feeling citizenships" (Simpson 2014, 173). But when she tried to delve into these experiences with the people she interviewed, they shrugged off her questions, often feigning ignorance by saying things like "no one seems to know." At first, Simpson found these responses confusing.

"It was very interesting to me that [my interviewee] would tell me that 'he did not know' and 'no one seems to know'," she explained (2007, 77), "To me these utterances meant, 'I know

you know, and you know that I know I know...so let's just not get into this.' Or, 'let's just not say.'"

Over time, Simpson came to embrace these limits on what she could ask — and therefore what she could say — about Mohawk nationhood. She understood the reluctance of her interviewees to engage with her line of inquiry as a strategy for survival. By refusing to answer certain questions, they were placing an epistemological barrier between their community and the academic gaze, which had deep roots in the settler colonial project. As Simpson argues, who belongs to Indigenous nations and on what terms is a fundamental question of the settler state, long used to justify the dispossession of land and resources from Indigenous peoples (Simpson 2007). Instead of pursuing an ethnographically "thick" description of how Mohawk people navigated tribal membership and why, Simpson embraced her interlocutors' refusals to discuss such topics, and used such moments as a pivot point for developing novel theoretical insights regarding the construction and maintenance of statist forms of sovereignty and nationhood (Simpson 2014).

I found a lot of parallels between Simpson's accounts of her research process and my own. For Simpson, refusal represented an opportunity to draw boundaries around the representational territory she was willing to engage with in her work. Rather than deliberate on questions regarding who belonged to Indigenous nations and who didn't, Simpson flipped the questions around, to ask why such lines of inquiry were deemed important and whose agendas they served. Similarly, I'd learned through my interactions with community organizers to interrogate research agendas that pivoted on an impulse to divide people into categories of deserving and undeserving of their liberty and full membership in society. As Chi (2022) argues, this notion of refusal "flips the gaze and calls the legitimacy of the questioner into account."

Like Simpson, I initially struggled to find a place within the existing literature to articulate the insights that emerged from the "discursive wrestling" that stemmed from such acts of refusal (Simpson 2007, 74). But Simpson's work, along with the writing of other Indigenous and feminist scholars, became an entry point for me to understand refusal as a research method and a framework for talking about the insights that emerge through various modes of "unlearning," or reworking our default modes of operation to a point where we disrupt cycles of harm (Keyes and Austin 2022).

The following article is my attempt to interrogate what Tuck and Yang (2014, 811) term the "code beneath the code" of computational research through the lens of refusal. Getting at the code beneath the code of computational work requires us to ask whose voices are being silenced and what sacrifices are being demanded for the sake of the research. Building from the work of Simpson and others, I explore the liberatory potential of refusal as a practice of generative boundary setting. To refuse is to say no—to reject the default categories, assumptions and problem formulations which so often underpin data-intensive work. But refusal is more than just saying no; it can be a generative and strategic act, one which creates space to renegotiate the assumptions underlying socio-technical endeavors.

The following article explores two complementary modalities of refusal in computation: "refusal as resistance" and "refusal as re-centering the margins." By exploring these two modes of refusal, I hope to provide a vocabulary for identifying and resisting the ways that sociotechnical systems reinforce dependency on oppressive structural conditions, as well as offer a framework for flexible collective experimentation towards more just futures. Most of the examples presented in this article are drawn directly from my fieldwork, and are my attempt at synthesizing some of the less intuitive insights that I gleaned through my collaborations with

community organizers. Ultimately, I hope that this work can serve as a useful framework for other computational practitioners interested in engaging in tough but fruitful conversations like the one that Sarah broached at NeurIPS, so that we can all begin the long journey of "unlearning, relearning, and negotiating which stories can cross which borders…and with what intentionality" (Nagar and Shirazi 2019, 240).

Bibliography

Chi, Elisha. 2022. “Critical Theory for Political Theology 2.0: Refusal.” *Political Theology Network* (blog). November 15, 2022. https://politicaltheology.com/refusal/.

Hamid, Sarah, dir. 2019. *Panel Discussion Track 3 - Towards a Social Good? Theories of Change in AI*. NeurIPS Conference 2019. Toronto, Canada. https://slideslive.com/38923829/panel-discussion-track-3-towards-a-social-good-theories-of-change-in-ai?ref=og-meta-tags.

Keyes, Os, and Jeanie Austin. 2022. “Feeling Fixes: Mess and Emotion in Algorithmic Audits.” *Big Data & Society* 9 (2): 205395172211137. https://doi.org/10.1177/20539517221113772.

Nagar, Richa, and Roozbeh Shirazi. 2019. “Radical Vulnerability.” In *Keywords in Radical Geography: Antipode at 50*, edited by Antipode Editorial Collective, Tariq Jazeel, Andy Kent, Katherine McKittrick, Nik Theodore, Sharad Chari, Paul Chatterton, et al., 236–42. Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/9781119558071.ch44.

Simpson, Audra. 2007. “On Ethnographic Refusal: Indigeneity, ‘voice’ and Colonial Citizenship.” *Junctures: The Journal for Thematic Dialogue*, no. 9.

———. 2014. *Mohawk Interruptus: Political Life across the Borders of Settler States*. Durham: Duke University Press.

Tuck, Eve, and K. Wayne Yang. 2014. “Unbecoming Claims: Pedagogies of Refusal in Qualitative Research.” *Qualitative Inquiry* 20 (6): 811–18. https://doi.org/10.1177/1077800414530265.

Refusal in Data Ethics:
Re-Imagining the Code Beneath the Code of Computation in the Carceral State[8]

"[C]oding, in the guise of objective science, expands the project of settler colonial knowledge production— inquiry as invasion is built into the normalized operations of the researcher. Coding, once it begins, has already surrendered to a theory of knowledge. We ask, what is the code that lies beneath the code?"

— (Tuck and Yang 2014b, 811)

Abstract

In spite of a growing interest in ethical approaches to computation, engineers and quantitative researchers are often not equipped with the conceptual tools necessary to interrogate, resist, and reimagine the relationships of power which shape their work. A liberatory vision of computation requires de-centering the data in "data ethics" in favor of cultivating an ethics of encounter that foregrounds the ways computation reproduces structures of domination. This article draws from a rich body of feminist scholarship that explores the liberatory potential of refusal as a practice of generative boundary setting. To refuse is to say no—to reject the default categories, assumptions and problem formulations which so often underpin data- intensive work. But refusal is more than just saying no; it can be a generative and strategic act, one which opens up space to renegotiate the assumptions underlying sociotechnical endeavors. This article explores two complementary modalities of refusal in computation: "refusal as resistance" and "refusal as re- centering the margins." By exploring these two modes of refusal, the goal of this paper is to provide a vocabulary for identifying and rejecting the ways that sociotechnical systems reinforce dependency on oppressive structural conditions, as well as offer a framework for flexible collective experimentation towards more free futures.

Introduction

[8] This article was originally published in Engaging Science, Technology, and Society in 2022.

Preferred citation: Barabas, Chelsea. 2022. "Refusal in Data Ethics: Re-Imagining the Code Beneath the Code of Computation in the Carceral State." Engaging Science, Technology, and Society 8(2): 35–57. https://doi.org/10.17351/ests2022.1233.

In spite of a growing interest in ethical approaches to computation, engineers and quantitative researchers are often not equipped with the conceptual tools necessary to interrogate, resist, and reimagine the relationships of power which shape their work. Like most euphemisms, terms like "diversity" and "fairness" have begun to blunt our collective imagination, making it difficult to articulate and intervene on the ways that data production and analysis perpetuate complementary structures of privilege and harm (Hoffmann 2021).

A liberatory vision of ethical data practice requires de-centering the data in "data ethics" in favor of cultivating an ethics of encounter. As Bergman and Montgomery (2017) explain, an "ethics of encounter" is grounded in a relational theory of change, building our capacity for collective struggle and lived transformation through forging new ways of relating across difference (Bergman and Montgomery 2017, 238). In contrast to techno- centric efforts to render algorithms more inclusive, accurate, or fair, an ethics of encounter requires us to grapple with the ways technocratic modes of innovation maintain violent social relations, so that we might begin "an ongoing process of becoming otherwise" (Bergman and Montgomery 2017, 112).

The concept of "refusal" offers a foundational framework for computational practitioners interested in pursuing a relational ethics of encounter in their work. I use the term "computational practitioners" to indicate a wide range of actors in academia, industry, and government who are engaged in data-centered discourse, research, and design. Refusal is an especially useful concept for computational practitioners involved in the emerging subfield of data ethics. This article draws from my own experiences working as a member of this community. For the last six years, I have collaborated with data scientists, community organizers, and government officials on a variety of applied data projects related to the US

criminal legal system, as well as participated in a number of critical interventions designed to deepen the conversation regarding what constitutes ethical data practice in this domain. Through these efforts, I have been both the recipient and instigator of acts of refusal which aimed to shed light on unnamed power asymmetries and reorient conversations towards more transformative modes of collaborative problem- solving.

In addition to my own experiences, this article builds from a rich body of feminist scholarship that explores the liberatory potential of refusal as a practice of generative boundary setting (Tuck and Yang 2014b; Honig 2021; Simpson 2014; McGranahan 2016; Wright 2018). As Tuck and Yang (2014b, 239) argue, "instead of a settler-colonial configuration of knowledge that is petulantly exasperated and resentful of limits, a methodology of refusal regards limits of knowledge as productive, as indeed a good thing." In recent years, feminist thinkers have worked to delineate ways that researchers might refuse harmful data regimes while affirming commitments to more radical and transformative social and scholarly practices (Hoffmann 2021; Cifor et al. 2019; Barabas 2020). Some scholars have explored the ways that marginalized groups refuse co-optation into harmful sociotechnical systems (Benjamin 2016b; Gangadharan 2020; Zong and Matias 2020; Zong 2020), while others have introduced opportunities for technology designers to embrace refusal as a generative and collaborative practice in their work (Graeff 2020; Benjamin 2016b).

Refusal requires computational practitioners to push beyond easy liberal fixes and abstract formulations of fairness, to grapple with the practical and material ways that data science is used to undermine collective struggles for liberation. Refusal is a particularly important practice for computational practitioners to engage in today, when there is heightened

demand for data-driven reform amid widespread abolitionist efforts to roll back the tendrils of the carceral state. In the United States, philanthropic organizations and government agencies have funded a number of high-profile initiatives aimed at reforming the carceral state through the adoption of "smart," "evidence based," and "data driven" practices (Stoller 2019; Ball 2015). Elite research initiatives such as the Chicago Crime Lab, Harvard's Access to Justice Lab, and the Stanford Computational Policy Lab endow computational practitioners with the resources, access, and influence necessary to intervene in high-stakes debates regarding criminal legal reform and the administration of social services. In this context, "data science" is often framed as the key skill set needed to "drive social impact through technical innovation." (Stanford Computational Policy Lab n.d.).

At the same time, there are a growing number of examples of data practitioners pushing back against uncritical entanglements with the carceral state. For example, thousands of employees from large tech companies like Google and Microsoft have staged protests to pressure their companies to withdraw from lucrative military contracts (Wakabayashi and Shane 2018; Schneider and Sydell 2019). Scholars have come together to demand that academic journals rescind offers to publish studies which fuel the "tech to prison pipeline" (Barabas, Beard, et al. 2020). And student groups have organized petitions to cancel computational courses that treat marginalized communities as "laboratories for experimentation" (Kahn and Levien 2021).

All of these efforts have emerged at a time when the carceral state is undergoing a crisis of legitimacy. By carceral state, I mean the expansive system of state-sanctioned capture, confinement, and control that underpins our current unjust social order, both in the United States and globally. The carceral state is a fundamentally relational phenomenon that extends far

beyond brick and mortar prison walls—as Ruth Wilson Gilmore explains, "prison is not a building 'over there' but a set of relationships that undermine rather than stabilize everyday lives everywhere" (Gilmore 2007, 242). These relationships are grounded in punitive systems of control that undermine the capacity for people to protect one another and resolve conflicts themselves, while simultaneously buttressing existing relations of power and domination. The carceral state is also a global phenomenon. Domestic policing programs are deeply entangled with militarized security practices abroad—methods for surveilling and punitively controlling people in one context are often repurposed for other locales around the globe (Schrader 2019). In this article, I draw primarily from examples based in the US context because that is where my own fieldwork has been grounded, while also nodding toward other important strains of research and praxis that are taking place outside of the United States.

Historically, law enforcement agencies and policymakers have embraced science and technology as a stabilizing force during moments when the carceral state is undergoing a crisis of legitimacy (Wang 2018). Data collection and analysis have long played an important role in upholding the racialized and gendered systems of meaning and control that legitimize punitive social practices. For example, US crime statistics were used to conflate Blackness with danger and criminality during the Progressive Era in order to justify racial violence and systematically exclude African Americans from the broader public sphere (Muhammad 2019). Modernized crime reporting initiatives were used during the 1960s to conflate social unrest with criminality during the Civil Rights Era (Murakawa 2014). The US government exported their police surveillance practices abroad in order to construct and keep track of "enemies of the state" during the Cold War (Weld 2014; Schrader 2019).

Today, data collection serves as a powerful vehicle for carceral expansion into other important realms of life, such as educational, financial, and healthcare institutions, often under the guise of progressive care-based rhetoric (Brayne 2014; Friedman 2021; Katz 2020). In this context, it is important for data practitioners to interrogate what Tuck and Yang (2014b, 812) term the "code beneath the code" regarding who benefits from and who is harmed by data-intensive interventions within the ever-expanding dominion of the carceral state.

Refusal could serve as a powerful analytic for computational practitioners who are interested in unpacking the code beneath the code of data work. To refuse is to say no—to reject the default categories, assumptions and problem formulations which so often frame the work of data science. But it's more than that; refusal can be a generative and strategic act, one which opens up space for us to forge new kinds of relationships and renegotiate the assumptions and premises underlying socio-technical work (Benjamin 2016b). Refusal allows the data practitioner to reposition themselves as actively producing the conditions of inquiry in ways that are accountable to the groups who are often silenced by data-driven discourse—it breaks down harmful modes of knowledge production and imagines a new, reparative role for itself. This article explores two complementary modalities of refusal in data-intensive work: "refusal as resistance" and "refusal as re-centering the margins." In order to concretize these concepts, I draw from examples from my own fieldwork, as well as other projects that have catalyzed acts of refusal within my broader community of collaborators.

As Tuck and Yang argue, refusal is not about identifying an unproblematic object of study, but rather it's about cultivating "an ethic of studying to object" (2014b, 814). Refusal as resistance involves identifying and rejecting the ways that sociotechnical systems undermine

collective resistance and reinforce dependency on oppressive structural conditions. Building from Ruha Benjamin's concept of "secondhand refusal" (2016b, 982), I explore the transformative potential of refusal as resistance when practiced by powerful institutional actors like computer scientists and technology designers. Computational practitioners often occupy privileged positions within organizations as either valued insiders or welcomed outsiders whose technical skills are in high demand. In such contexts, computational practitioners are in a powerful position to negotiate and challenge the underlying theories of change associated with a given data project. When met with resistance, such individuals have the choice to decline to participate in the work.[9] By contrast, the people targeted by carceral data systems are often not given a chance to refuse participation. In this context, secondhand refusals are crucial, because they give computational practitioners the opportunity to voice dissent in solidarity with vulnerable and marginalized populations.

In my discussion of "refusal as resistance," I outline three common pitfalls that computational practitioners fall into when engaging in work regarding the carceral state: 1) "proving" harm 2) adopting deficiency narratives and 3) optimizing harmful systems. I explain how these pitfalls undermine collective efforts toward transformative change, and then explore

[9] Given the broadness of the term "computational practitioners" it's important to note that there is significant variation in the level of privilege and agency that individuals have access to when considering refusal as a course of action. For example, a software engineer at Google is likely to have much more leverage in a conversation regarding the design of a new digital platform than an hourly wage laborer doing manual data entry on that same platform. The price of walking away from such a project would also be much more burdensome for the hourly wage laborer. In this article, I am primarily concerned with exploring the potential of refusal for relatively privileged computational practitioners. This sub-group is quite large and can include highly paid salary workers, as well as students and early career researchers.

ways computational practitioners might resist these pitfalls in their work. Finally, I discuss ways we might view "refusal as resistance" as a generative project, one that creates space for new possibilities to emerge.

Refusal as re-centering the margins explores the potential of refusal as a transformative relational practice. Building on the work of bell hooks, I conceptualize the margins as a space of radical openness and possibility, rather than as a site of deprivation and need (hooks 1989; Shah 2015). Refusal as re-centering the margins hinges on an acknowledgement that transformative social work is always already present and in formation (Kaba 2021), and explores the ways that computational practitioners might contribute to this work without assimilating it into dominant power structures. Refusal as re-centering the margins requires an intentional redistribution of resources and power (Benjamin 2016b), as well as the cultivation of "common notions" (Bergman and Montgomery 2017, 42) that reshape our habits for working together across power differentials. Such an approach lays the foundation for flexible collective experimentation towards the possible, rather than the probable (McGranahan 2016; Benjamin 2016a).

Refusal as Resistance

"Proving" Harm

One of the most prominent theories of change underlying computational work is the idea that quantitative analysis itself can make an impact by proving harm. For example, perhaps one could develop an observational study to quantify the role that race plays in police-involved shootings (Fryer 2016), or construct a randomized control trial to measure the impacts of a new violence prevention program (Kubiak et al. 2015). One common, seemingly progressive hope underlying such studies is that, by rigorously documenting the effects of such policies, researchers might

bolster the legitimacy of personal testimonies or qualitative research that already exists and thus persuade policymakers and law enforcement agencies to abandon harmful practices.

Activists and feminist scholars have pushed back against this widespread intuition, arguing that damage-centered narratives that are designed to convince an outside adjudicator that violence has been perpetrated rarely lead to accountability or reparations for the aggrieved (Hartman 1997; Gilmore 2002; Tuck and Yang 2014a; Onuoha 2020). As Mimi Onuoha (2020) argues, "The idea that structural racism can be proven and overcome by gathering just enough or the right kind of evidence is nothing more than a myth. Historically, it has rarely been the case... the grand ritual of collecting and reporting this data has not improved the situation." This is particularly true in instances when harms are carried out against marginalized racial communities, such as Black and Indigenous peoples. For example, scientific approaches to documenting the harms of punitive legal practices which disproportionately impact people of color have proven largely ineffective in courts of law (Gilmore 2002).

Moreover, such approaches reinforce knowledge hierarchies that effectively silence the voices of directly impacted people and maintain the computational practitioner's privileged status as the primary arbiter of truth (hooks 1990; Tuck and Yang 2014a). All too often, the partial insights gleaned from computational research are used to undermine robust bodies of historically grounded and community- centered knowledge that are already available. The result of such work is the production of research that makes far-reaching claims based on faulty assumptions.

For example, a recent observational study made headlines by claiming that "contrary to conventional wisdom, parental incarceration has beneficial effects on children," especially for

Black children (Norris, Pecenco, and Weaver 2021, 1). The authors of the study do not substantially engage with the large body of research that already exists documenting the various physical, psychological, and social harms associated with parental incarceration. Rather, they measure the primary "beneficial impact" of parental incarceration in extremely narrow terms—the likelihood that a child will be charged, convicted, or incarcerated for a crime before the age of twenty-five. Such limited notions of impact are commonplace in quantitative research regarding the carceral state, because the maintainers of the criminal legal system collect only a narrow set of operational data about the system's impact, usually in terms of "recidivism."

Recidivism is a classic example of a quantitative metric which reveals virtually nothing about the conditions that people live under after they leave prison. Yet it continues to serve as the default standard in measuring post-carceral success, largely because it is a convenient data point to capture via administrative systems. Such myopic data collection results in serious consequences, often erasing the well-established harms of incarceration and recasting violent policies as benevolent interventions.

The state of existing data regimes presents a serious ethical challenge for quantitative researchers in this area. As Lily Hu argues, "Either [the researcher] buys herself the ability to work with troves of data, at the cost of implausibility in her models and assumptions, or she starts with assumptions that are empirically plausible but is left with little data to do inference on" (2021). As such, an ethical computational practice will often require practitioners to look beyond the limitations of current data regimes, to imagine the data for what it might have been.

For example, I was once part of a research team that was negotiating access to data housed in a state government's administration of the courts. During the course of our

interactions, one government official suggested that we use their data to evaluate the impact of electronic monitoring on pretrial outcomes. Although this was beyond the scope of our original research topic, our team considered taking on the project as a stepping stone to accessing the data we wanted.

Before agreeing to do the study, we consulted with James Kilgore, a researcher and organizer against the use of "e-carceration," or punitive technological interventions that deprive people of their liberty. Kilgore warned us against pursuing an impact study that was limited to data collected by the courts. He pointed us to a widely cited study on the effectiveness of electronic monitoring, wherein the authors used propensity score matching in order to compare the outcomes of people assigned to electronic monitoring to those who were not. The authors included one hundred and twenty-two covariates in their analysis, boasting that "the richness of [their] covariate set is quite extraordinary" (Bales et al. 2010, 47).

Yet as Kilgore (2017) points out elsewhere, these propensity scores did not include any of the variables that most analysts and formerly incarcerated people identify as the most important factors contributing to recidivism, such as employment, access to housing, and the strictness of the supervising law enforcement agent. As Kilgore (2017, 3) argues, without taking such factors into account, the study "becomes merely an exercise in statistical gamesmanship rather than serious research." After careful consideration, we decided not to proceed with the proposed study, largely due to concerns that the limited selection of data points would erase the harmful impacts of electronic monitoring and silence the testimonies of directly impacted people who were organizing against the use of this form of e-carceration.

Damage-centered research can also perpetuate harm in carceral contexts by serving as a distraction that diverts precious time, attention, and resources away from more transformative efforts for change. This is especially true during times of social upheaval. Calls for expanded data collection are often framed as pragmatic and politically neutral responses to widespread calls for change, but they have significant social and political consequences.

For example, during the height of the Covid-19 crisis in 2020, a network of grassroots bail funds reached out to me and other allied researchers with concerns about a study being launched by Harvard's Access to Justice Lab. The stated goal of the study was to evaluate whether the presence of counsel at first appearance in court increased the likelihood that a person would be released from jail (North 2020). The study was launched amid widespread calls to decrease jail populations around the country, given the fact that jails and prisons had become major vectors for the spread of Covid-19 (Schnepel 2020). Jim Greiner, the faculty director of the Access to Justice Lab, framed the study as an opportunity to glean "concrete evidence" on the potential benefits of legal counsel, which he argued could pave the way for useful policy changes (North 2020).

Yet, as others at Harvard Law School pointed out, the study was harmful and unnecessary because 1) there already exists evidence demonstrating that access to counsel improves outcomes and 2) available research reveals how pretrial detention harms detained people and their loved ones (Naples-Mitchell 2021). As Katy Naples-Mitchell (2021) of Harvard's Houston Institute argues, "Instead of randomizing access to interventions for which there is already an evidence base in order to prove the causal mechanism with greater precision, we should be spending our efforts, and our funds, on implementing those policies as widely and quickly as possible."

Naples-Mitchell points out that there is a stark disparity in the levels of evidence required to roll out expanded policing, surveillance, and incarceration compared to the proof required to undo such harmful systems. Rather than simply act on existing evidence, the Access to Justice Lab decided to reframe a beneficial intervention as an unknown. Meanwhile, thousands of people were needlessly exposed to a life-threatening virus behind bars. In the context of the carceral state, data- intensive studies like this impose serious burdens on directly impacted communities, used to investigate questions for which answers are already available, but often ignored.

Data-intensive projects like observational studies and randomized control trials have become increasingly popular in recent years as abolitionist social movements such as the Movement for Black Lives work to dismantle the carceral state. For example, in response to uprisings against police brutality in 2015, FBI Director James Comey (2015) argued that "the first step to understanding what is really going on in our communities is to gather more and better data related to those we arrest. . . . Data seems a dry and boring word but, without it, we cannot understand our world and make it better."

Khalil Gibran Muhammad (2015) argues that this statement frames data collection as a politically neutral response to the problem of police brutality, while simultaneously implying that further study is needed before we can truly understand what the core issues are. Rather than listen to the communities who are directly impacted by police violence, Comey asserts that expanded data collection is the only way to understand the problem and identify steps for improving it. Such interventions dilute and diffuse the momentum of transformative social movements by giving the appearance of taking action while maintaining the status quo (Hoffmann 2020).

Getting at the code beneath the code of computational work requires us to ask whose voices are being silenced and what sacrifices are being demanded for the sake of the research. As Ruha Benjamin (2016a, 2) argues, such questions ultimately push us away from narratives that are "meant to convince others of what is . . . to expand our visions of what is possible." Refusal is a practice that we can use to redirect academic analysis away from harmful, damage-centered narratives toward the institutions and policies that produce those narratives in the first place (Zahara 2016).

Adopting Deficit Narratives

One important aspect of refusal involves an intentional shift in how we interpret data, away from measuring the "antisocial" pathologies of "risky" individuals and towards a carceral system that surveils and punishes people in disparate ways. As Catherine D'Ignazio and Lauren Klein (2020, 167) argue, the language surrounding data analysis of poor or marginalized communities is often grounded in "deficit narratives," or descriptions that reduce a given population to their perceived deficiencies, rather than highlighting the richness that they possess. In the context of the carceral state, language is a powerful tool for naturalizing harm against historically marginalized groups, but it is also an important site of struggle (hooks 1989). As Kimberlé Crenshaw (2012, 1466) argues, when discussions of incarceration are framed around "at-risk" populations, the result is a "subtle erasure of the structural and institutional dimensions of social justice politics."

Such framings often show up in the way data are characterized in computational models. For example, data about arrests, convictions, and incarceration are often used to measure "public safety risk" even though numerous researchers have pointed out the fundamental measurement

errors that occur when computer scientists use such data as a shorthand for danger or risk (Barabas, Beard, et al. 2020; Dolovich 2011; Harris 2003; Prins and Reich 2018). Similarly, when computer scientists develop machine learning algorithms to identify or predict "criminality" using biometric and/or criminal legal data, they materially conflate the shared, social circumstances of being unjustly overpoliced with pathological criminality (Barabas, Beard, et al. 2020). Such claims are based on an ahistorical interpretation of data, devoid of notions of structural harm and state-sanctioned racialized violence.

Yet these flawed measures persist because they are politically useful as rhetorical tools for justifying violent and punitive decisions (Stop LAPD Spying Coalition 2018). Such criminalizing tactics are not new. As Naomi Murakawa (2014, 151) argues, attempts to modernize law enforcement and the courts have historically served to legitimize and expand the carceral state, maintained through a "politics of pity" that depends on portrayals of Black people as damaged and potentially violent. The persistence of this interpretive lens is not surprising, given the political economy of data-intensive work. Computation often relies on data and analytical tools that are housed in powerful institutions, which subtly shape research agendas and problem framings by acting as gatekeepers. In addition, the data collection regimes of powerful institutions tend to skew toward the surveillance of marginalized populations who often do not have the political capital to successfully resist such tracking (Eubanks 2018).

As a result, available data, and the accompanying interpretations of that data, tend to erase systemic violence and shift blame and risk onto marginalized groups. These data then serve as the basis for dividing people into categories of deserving and undeserving of life-changing intervention (Kaba 2020; Gilmore 2017). One of the most important forms of refusal we can

engage in as computational practitioners is to reject these default problem framings and reorient the algorithmic gaze "upward" to examine the people and institutions who occupy positions of power (Barabas, Doyle, et al. 2020; Benjamin 2016b). Recent efforts to resist pretrial risk assessment offer a great example of how we might support such shifts in computational work.

Pretrial risk assessments are a classic example of the downward orientation underlying most data- driven reforms. These tools were introduced as a solution to mass pretrial incarceration, a serious and growing problem in the US (Barabas, Doyle, et al. 2020). This issue is primarily driven by a severe case of "institutional decoupling" in the US courts (Christin 2017), whereby judges increasingly diverge from state and federal laws which mandate that pretrial detention be the limited exception, not the rule. Proponents of pretrial risk assessment claim that such tools could help to course correct judges' behavior by offering a more accurate and objective view into who poses a serious risk to the community. Yet, to date, these tools have not made a significant impact on the way judges make decisions regarding pretrial release (Stevenson 2017).

In 2016, I was a part of a team of researchers who were interested in developing a computational platform that would enable court administrators to audit the accuracy and biases of algorithmic risk assessments. Our hope was that we could increase the rate at which judges released people pretrial by improving the transparency and reducing the biases of such tools. This approach was aligned with the broader discourse surrounding pretrial risk assessments, where there was growing interest in the development of strategies to audit and reduce the biases of high-stakes decision-making tools (Kleinberg et al. 2018; Chouldechova 2017).

Early on in the project, we were challenged to fundamentally re-think our approach by a group of community organizers working on pretrial justice. We'd originally reached out to the group to see if they would be interested in co-designing our audit platform. Rather than collaborate with us on our terms, the organizers engaged in an act of refusal, pushing us to re-think the project entirely. According to them, the main problem with pretrial risk assessments wasn't that they were inaccurate or that judges didn't adhere to them. Rather, it's that they focus exclusively on modeling the supposed risk of people awaiting trial, instead of shedding light on a punitive courtroom culture run amok. They argued that the same court data could be used to surface insights regarding the way judges make pretrial release decisions. Such a reorientation would shift the solution framing away from problematic measures of individual "risk" in marginalized populations, toward studying the behaviors of decision makers who possess the most agency to drive pretrial outcomes.

In light of these comments, our team reformulated our understanding of the problem we aimed to address and ultimately pursued a different intervention. Rather than seek to improve the accuracy of existing risk assessments, we developed an alternative tool that measured the risk that a given judge would unlawfully deprive someone of their liberty before their trial. The goal of this work was to shift the conversation away from measuring the risk of individual defendants to increasing the accountability of powerful system actors. This project exemplifies a deliberate shift in the unit of analysis, away from the people who bear the brunt of our unjust social order, and toward the institutions which perpetuate such harmful policies. Computational work which shifts its focus onto the study of powerful institutions could lay the foundation for more robust forms of accountability and shed light on the structural factors that produce unjust outcomes.

Refusal as resistance, then, requires researchers to cultivate an approach to analyzing data "within a matrix of commitments, histories, allegiances and resonances" (Tuck and Yang 2014b, 811) that informs what can and cannot be known through computational analysis. Such refusals hinge on an acknowledgement that quantitative work is always interpretive and requires having a fluent command of the power dynamics underlying the way questions are framed and the data used to answer those questions (Gangadharan 2020; Irani et al. 2010).

Optimizing of Harmful Systems

One of the most effective means of co-opting the demands of transformative social movements is through the adoption of technocratic reforms that embrace progressive rhetoric while normalizing and expanding systems of harm (Benjamin 2019; Katz 2020; Murakawa 2014; Spade 2015). Such interventions are what Ruth Wilson Gilmore (2007, 242) calls "reformist reforms" because they widen, rather than dismantle, the carceral state's net of social control. An important and generative mode of refusal, then, involves disengaging from computational work that aims to optimize the efficiency and expand the reach of carceral systems through ever more data collection, maintenance, and analysis.

One of the primary vehicles for carceral expansion is through bolstering the state's capacity to produce facts on the ground that support punitive policies (Melamed 2019; Sutherland 2019). Computational systems are particularly useful vehicles for this kind of expansion, due to the depoliticized nature of mainstream engineering culture and education, which encourages engineers to tinker on the edges of large bureaucratic systems without ever grappling with the downstream consequences of their efforts. By recasting the dirty work of the carceral state in terms of bureaucratic optimization challenges and technical data audits, officials

are able to abstract away the pain and misery enabled by carceral systems while simultaneously giving computational practitioners a chance to demonstrate their technical prowess (Katz 2020).

Such "win-win" scenarios stabilize structural violence by strengthening dominant systems of meaning and control rather than dismantling them to make room for alternative approaches to safety and justice. As Ruth Wilson Gilmore (2014, 13) explains, "big answers are the painstaking accumulation of smaller achievements. But dividing the problem into pieces in order to solve the whole thing is altogether different from defining a problem solely in terms of the bits that seem easiest to fix . . . [it] is the difference between reformist reform—tweak Armageddon—and non-reformist reform—deliberate change that does not create more obstacles in the larger struggle."

The depoliticized nature of engineering culture means that technologists tend to dismiss epistemic questions regarding how data are interpreted and operationalized as "non-technical" concerns, irrelevant to the "real" work of engineering (Cech 2013). This ideology assumes that political and social contexts can and should be kept separate from the purely technical concerns of engineering (Cech and Sherick 2015). As a result, criminalizing logics and discourses are easily embedded into these systems, because the interpretation and operationalization of data analysis is considered outside the purview of system engineers.

State agencies often then interpret the results of algorithmic systems according to a default philosophy of punishment in marginalized communities (Roberts 2019). For example, a company called LEO Technologies partnered with the National Sheriff's Association during Covid-19 to "support the health and safety of our deputies and inmates" (Thompson 2020). LEO offers a product called Verus, a transcription service that uses natural language processing to

produce real-time transcriptions of incarcerated people's phone calls, as well as keyword-based searches and alerts. LEO's marketing materials are couched in a language of compassionate care, claiming that incarcerated people's phone calls "contained valuable intelligence" that could be used to listen for "cries for help from vulnerable inmates . . . information that could potentially save lives" (LEO Technologies 2020b).

However, a few months into the pandemic, LEO came under fire for their "absurd" attempts to mitigate the impacts of Covid-19 in prisons (Lacy et al. 2020). Rather than removing obvious obstacles to care, prisons opted to eavesdrop on prisoners' phone calls to supposedly identify people who might be sick, as well as "other threats" related to the disease. As it turns out, those other "threats" included outlandish and unsubstantiated events that shifted the focus away from the dire, life and death situation that incarcerated people faced behind bars during the pandemic, recasting the prisoners themselves as the imminent danger to be managed.

For example, LEO claimed that by searching for keywords like "kill him" they had identified inmates with plans "to kill anyone who was discovered to have the virus in order to contain the spread" and by searching for the word "coughing" they had identified a person who "had declared that inmates would begin coughing on guards if coronavirus was discovered in the jail" (LEO Technologies 2020a). Rather than mobilize resources to increase access to testing and treatment, prison officials invested in software that helped them manufacture a different narrative, one that recast captive populations as the real threat to be mitigated during a deadly pandemic. Meanwhile, transmission and death rates in prison skyrocketed at rates three to five times higher than the general population (Saloner et al. 2020).

This example powerfully illustrates the ways data are used to prop up narratives that justify the systematic abandonment and sacrifice of people's lives for the sake of preserving the life of carceral institutions (Friedman 2021). In response to public outcry after the revelation of these supposed use cases, LEO Technologies CEO Scott Kernan asserted that, "Our company points [prisons] to a point in the call where a word or phrase was said that they have concerns about. What they do with that information is purely out of our control" (Lacy et al. 2020). Kernan's position is emblematic of a broader stance within engineering culture, which offloads responsibility for harm propagated by sociotechnical systems onto circumstances "outside of their control." In their marketing materials, the company prefers to emphasize the technical performance of their system, highlighting that it is automated (enabling continuous surveillance), fast (offering "real-time" intelligence), and objective (without "implicit bias") (LEO Technologies n.d.).

Refusal in this context requires computational researchers to actively engage with the material and epistemic stakes of their work, by recognizing the ways that techno-reforms ultimately justify and expand the violent practices of the carceral state (Katz 2020). By disengaging with efforts to expand, improve and maintain the carceral state's administrative power through data collection and analysis, we create space for experimenting with new ways of improving people's life chances beyond the constraints of the criminal punishment system. In this sense, refusal can be understood as a beginning that starts with an end (Barabas 2020). This stance pivots on a rejection of the notion that carceral systems are simply broken and in need of fixing. As Mariame Kaba (2021, 36) urges, let us not focus on asking "'What do we have now, and how can we make it better?' Instead, let's ask, 'What can we imagine for ourselves and the world.'"

Resistance as a Generative Act

The above sections outline three important opportunities for "refusal as resistance" in computational work. In the context of the carceral state, these modes of resistance hinge on second-hand refusals from computational practitioners who decline to engage in data projects that are often lucrative and professionally beneficial, but that ultimately reproduce harmful social structures.

Refusal as resistance is more than just a decision to decline what's on offer—it opens up opportunities for alternative social formations to emerge, ones which re-center the communities who are often targeted and silenced by data analysis. Refusal as resistance embraces limits on knowledge production and technological capacity, framing those limits as productive boundaries that ultimately create space for us to renegotiate the assumptions and key vocabularies underlying data work. As Sara Ahmed (2017) argues, such refusals should be understood as generative acts, a kind of "counter-institutional project . . . creating new paths for others to follow." Resistance in this context lays the groundwork for us to learn from dispossessed people without serving up stories of harm or peddling in criminalizing narratives.

But refusal as resistance, especially in the form of secondhand refusals enacted by relatively privileged actors, is only a first step. At the end of the day, computational researchers and toolmakers must not only create space but also cede space in order to enact new modes of relating to one another outside of techno-centric theories of change. In the following section, I outline a framework for refusal as re-centering the margins which explores some of the practical aspects of doing this work.

Refusal as Re-centering the Margins

The crux of refusal as re-centering the margins lies in identifying practical ways for data practitioners to participate in and support social transformation rather than control it. All too often, data analysis is used as a tool for circumscribing political agendas and maintaining the "savior status" of technical experts (Raji, Scheuerman, and Amironesei 2021). While traditional strategies of mass organizing are chronically underfunded, philanthropic and governmental organizations channel resources toward projects with quantifiable outcomes that are housed in elite institutions. Social justice work becomes a career track pursued by specialized professionals who maintain a monopoly over decision-making and policy discourse (Spade 2015). Refusal as re-centering the margins moves beyond technocratic notions of social change to develop a shared imagination of what transformative change looks like and what we need to do to get there. Re-centering the margins is easier said than done. Such work requires computational practitioners to embrace a theory of change that decenters data and focuses instead on cultivating new ways of relating "at the margins" (Shah 2015; Dutta and Pal 2010). As bell hooks (1989, 20) teaches us, the margins are a site of radical possibility, a "central location for the production of a counter hegemonic discourse that is not just found in words but in habits of being and the way one lives." As hooks also reminds us, the margins are not always a comfortable place to operate from—and efforts from computational practitioners to engage at the margins might be met with skepticism and distrust. Benjamin (2016b) explains that such distrust often stems from a long history of asymmetrical social relationships, which are reproduced through an uneven distribution of material resources and symbolic power.

Rebuilding trust at the margins requires the careful cultivation of spaces where silences can be broken so that new shared vocabularies and habits can be formed (Lorde 1984). A first step toward creating such spaces involves an intentional redistribution of power and resources, beyond diversity and inclusion models of participation in hegemonic structures. Only then can we cultivate the collective capacity to rebuild trust, which may entail taking responsibility for prior and ongoing complicity with dominant structures.

Open dialogue about past and present harms creates opportunities for us to radically reshape our habits and ways of being together across difference. To that end, refusal as re-centering the margins also involves cultivating what Bergman and Montgomery (2017) call "common notions," rather than a rigid set of best practices. As I explore below, common notions are a flexible and pragmatic set of shared sensibilities that guide us in fostering mutually enabling relationships through the enactment of values such as accountability, reciprocity and embeddedness.

Redistributing Resources and Power

A key aspect of refusal involves cultivating relationships that are anchored in an intentional redistribution of resources and decision-making power (Benjamin 2016b). Technical solutions and tools often divert material resources and symbolic power away from directly impacted communities and into the hands of technocratic elites. When computational practices have a monopoly over what's considered innovative or cutting-edge, then we end up erasing a much wider array of community-led activities that are at the heart of transformational work (Benjamin 2019). This problem cannot be solved by simply making data-driven processes more diverse and inclusive. As Anna Lauren Hoffmann (2020, 11) explains, inclusion in technical

design processes "normalizes the dependency on exclusive forms of expertise in ways that do not address but feed and maintain the potential for violence." Refusal as re-centering the margins requires us to resist the terms of inclusion on offer and think through new modes of relating to one another outside of techno- centric theories of change.

This is not to say that data interventions do not have any place in social change efforts, but they should be conceived of as supplementary resources, rather than as the focal point of the work. Expert interventions might take on a supportive role in service of larger strategies for mass mobilization in social change processes (Spade 2015). For example, James Kilgore points out that social movements can recognize the value in data, without necessarily being "data-driven" (Media Justice 2021). Kilgore is a part of a collaborative data project called "Mapping E-carceration" which tracks ongoing developments related to e- carceration around the country in order to support local and national organizing efforts. As his collaborator Eteng Ettah describes, the primary role of data in this project is to inspire more people to join the collective struggle against ever-expanding regimes of e-carceration. "Stories and narratives move folks to action," Ettah (2021) argues, "This is data . . . directly pushing back against the notion that the lived experiences of black and brown folks do not matter, because they do."

Kilgore and Ettah's use of the term data includes qualitative information and personal testimonials, which are valued for their affective and narrative potential. The use of such data inverts the default knowledge hierarchies implicit in most data regimes, which place quantitative data as the most highly valued form of information. This project illustrates the ways that community organizers have used an expansive notion of data as a resource for building collective

power, rather than as a strategy for circumscribing the agenda and monopolizing resources for narrow technical interventions.

Technical experts can play a valuable role in supporting such grassroots strategies for data activism. For example, in 2018 CourtWatch MA launched a community data collection project that aimed to shift "the power dynamics in our courtrooms by exposing the decisions judges and prosecutors make about neighbors every day" (Courtwatch MA 2018). The project faced a number of challenges due to rules regarding outside data collection. For example, all observations had to be recorded by hand with pen and paper because it was prohibited for members of the public to enter the courtroom with digital devices such as cell phones or computers. This presented a massive administrative challenge for the organizers of CourtWatch MA, who then had to develop a strategy for digitizing and analyzing that data.

I was part of a small group of computational practitioners who supported this effort, which included developing a data entry and clean-up strategy and recruiting students from local universities to help with the tedious work of entering handwritten observations into the database. It's important to note that this kind of technical support is not flashy and will not likely be considered "cutting-edge" in the field of computer science (Hope et al. 2019). As Terra Graziani and Mary Shi (2020, 399) point out, "there is often a gap between the work that is rewarded by the academy and the work that most directly empowers communities."

However, it is this kind of contribution that can actually move the needle on community-led engagements with data. The goals and strategies for CourtWatch MA were developed by community organizers, who then called on the support of technical experts to assist them in implementing their vision. Technical experts played a similar supporting role in campaigns

against expanded surveillance infrastructure in places like San Diego. As Irani and Alexander (2021) explain, technology workers lent their expertise in order to counter official claims about what new surveillance systems could or couldn't do. They engaged in tedious work, such as combing through government contracts, in order to identify specific threats and concerns regarding how surveillance data were used and by whom. In this way, data expertise served as a practical resource for grassroots efforts, rather than as a tactic for seizing resources and control of the work.

Cultivating Common Notions

Refusal offers a relationship-centered approach to ethics that builds our collective capacity to create emergent alternatives to existing social conditions. A key aspect of this work involves undoing hegemonic desires to control processes of social change and addressing our complicity with oppressive social structures, so that space is created for more radical and collective forms of transformation. Yet within the field of computation, the overwhelming tendency is for practitioners to envision ethics as a set of rigid rules and decontextualized best practices that leave oppressive modes of engagement intact. As Sareeta Amrute (2019, 57) argues, such frameworks are often the result of top-down efforts to "future proof" technologies against potential harm, yet such rules-based frameworks are wholly inadequate for addressing the types of relational and structural harms explored in this article. By contrast, the concept of "common notions" could offer an alternative approach to ethical data practice. Common notions are based on an inherently experimental and improvisational approach to social change that stems from a "constant working on each other" (Bergman and Montgomery 2017, 115). The

essence of this process pivots on an ethics of encounter, or "holding a conversation rather than following a recipe" (Irani et al. 2010, 1317).

Bergman and Montgomery (2017) define common notions as an evolving set of shared values that arise out of frank (and often uncomfortable) conversations across difference. As Audre Lorde (1984, 86) argues, "it is not difference which immobilizes us, but silence. And there are so many silences to be broken." A first step toward developing common notions, then, involves breaking long held silences so that we can engage in conversations which "enable us to see and feel the toxicity of some of our attachments" (Bergman and Montgomery 2017, 116) or the ways that we reproduce structures of domination in our work. Nagar and Shirazi (2019, 239) talk about this in terms of cultivating "radical vulnerability," or the process of "reminding ourselves and one another of the violent histories and geographies that we inherit and embody despite our desires to disown them."

I have experienced the value of such conversations first-hand. As I mentioned in the introduction to this article, I was once part of a team of researchers who initiated a meeting with leaders from the Massachusetts Bail Fund in order to explore potential collaborations. During the meeting my team offered up a number of ideas on ways we might use data collected by the bail fund to run various studies. By and large, our ideas were met with silence. Toward the end of the meeting, Atara Rich-Shea, the executive director of the bail fund, leaned in and shared her frank perspective. She said the bail fund was frequently contacted by researchers who were only interested in asking their own questions and that, for the most part, those questions were harmful to the people she served. She went on to explain the ways that academics undermine the work of organizations like hers by pursuing research that either divided people into categories of

"deserving and undeserving," distracted from more pressing issues, or erased the violence of carceral policies.

Atara broke her initial silence during our brainstorm to tell us that the bail fund was not interested in sharing their data with us if there was no accountability for how that data would be used. She also challenged some of the language we'd been using throughout the conversation, pointing out the ways that that language reinforced harmful assumptions about the people that she worked with. Atara's comments were the beginning of many more fruitful conversations about how our group might begin to ask better questions and set up structures to ensure that our work was accountable to community organizations like hers. Over time, not only did the quality of our questions improve, but we became increasingly capable of identifying and un-doing harmful tendencies in our work.

In order to create space for such conversations to occur, computational practitioners must rethink the default extractive modes of engagement that are inculcated in their field. One of the major appeals of computational work is that it provides the opportunity to traverse multiple high stakes domains while engaging in large-scale problem-solving. In theory, the same basic computational methods might be used to develop a cure for malaria or to weigh in on important policy debates regarding public safety. Yet such interventions often ignore the nuances of local context and operate without concern for prior and ongoing grassroots efforts. Few computational practitioners take the time to cultivate relationships with directly impacted communities. And when they do, those interactions often lack the structures and commitments necessary to develop a shared sense of accountability and reciprocity in the encounter (Benjamin 2016b). As a result,

computational practitioners miss valuable opportunities to learn about the ways they might break cycles of domination in their work and forge new kinds of mutually enabling relationships.

A foundational first step toward cultivating common notions is to prioritize long-term relationships with grassroots collaborators situated in local contexts. As Graziani and Shi (2020, 409) explain, a commitment to long-term relationships means "moving at the speed, scale, and pace of community collaborators instead of according to the expectations of the academy." Embedding oneself in a local community over a prolonged period of time enables computational practitioners to become attuned to the contextual and temporal nuances of their work so that they can adjust to changing circumstances.

Such long-term commitments lay the foundation for building and continuously renegotiating a shared vocabulary for the work. A key aspect of building a shared vocabulary involves recognizing the importance of bi-directional communication—it is not simply about familiarizing lay people with technical jargon, but also about familiarizing technical people with the liberatory terms and counter-concepts communities use to break free of discursive violence. Yet all too often, the latter is neglected or completely overlooked. When bi-directional communication is prioritized, it creates space for multiple forms of expression and types of knowledge to be valued and heard. This lays the groundwork for productive conversations in which a set of common notions can emerge.

Conclusion

The goal of this article has been to explore the liberatory potential of refusal as a framework for ethical engagement in data science. While significant effort and resources are

spent on formalizing abstract notions of "fairness" in data, this work leaves a number of harms unexamined. Refusal expands the conversation beyond the "solution space," to grapple with the ways that computational work obscures violence and reproduces structures of domination. Rather than frame the expansion of data regimes and technological capacity as an inherent good to be maximized, the concept of refusal embraces boundaries as productive and beneficial.

But refusal is more than just an exit strategy. In its most potent form, refusal operates as a framework for renegotiating the terms of engagement, creating space for new kinds of relationships to emerge within radically different structures of commitment and accountability. Rather than adhering to a fixed set of rules, the goal of refusal is to cultivate an ethics of encounter that enables us to grow in our collective capacity to experiment, learn, and build new worlds together. In this way, refusal offers an entry point into a transformative process of becoming otherwise, to break free from overdetermined notions of the probable or the practical in order to enact the possible.

Works Referenced

Ahmed, Sara. 2017. “Institutional As Usual.” Feministkilljoys. October 24, 2017. https://feministkilljoys.com/2017/10/24/institutional-as-usual/.

Amrute, Sareeta. 2019. “Of Techno-Ethics and Techno-Affects.” *Feminist Review* 123 (1): 56–73. https://doi.org/10.1177/0141778919879744.

Bales, William, Karen Mann, Thomas Blomberg, Gerry Gaes, Kelle Barrick, Karla Dhungana, and Brian McManus. 2010. “A Quantitative and Qualitative Assessment of Electronic Monitoring.” *Bibliogov*, January, 208.

Ball, Molly. 2015. “Do the Koch Brothers Really Care About Criminal-Justice Reform?” *The Atlantic*, March 3, 2015. https://www.theatlantic.com/politics/archive/2015/03/do-the-koch-brothers-really-care-about-criminal-justice-reform/386615/.

Barabas, Chelsea. 2020. “To Build a Better Future, Designers Need to Start Saying ‘No.’” Medium. October 20, 2020. https://onezero.medium.com/refusal-a-beginning-that-starts-with-an-end-2b055bfc14be.

Barabas, Chelsea, Audrey Beard, Theodora Dryer, Beth Semel, and Sonja Solomun. 2020. “Abolish the #TechToPrisonPipeline.” Medium. June 29, 2020. https://medium.com/@CoalitionForCriticalTechnology/abolish-the-techtoprisonpipeline-9b5b14366b16.

Barabas, Chelsea, Colin Doyle, JB Rubinovitz, and Karthik Dinakar. 2020. “Studying Up: Reorienting the Study of Algorithmic Fairness around Issues of Power,” 10.

Benjamin, Ruha. 2016a. “Racial Fictions, Biological Facts: Expanding the Sociological Imagination through Speculative Methods.” *Catalyst: Feminism, Theory, Technoscience* 2 (2): 1–28. https://doi.org/10.28968/cftt.v2i2.28798.

———. 2016b. “Informed Refusal: Toward a Justice-Based Bioethics.” *Science, Technology, & Human Values* 41 (6): 967–90. https://doi.org/10.1177/0162243916656059.

———. 2019. *Race after Technology: Abolitionist Tools for the New Jim Code*. Medford, MA: Polity.

Bergman, Carla, and Nick Montgomery. 2017. *Joyful Militancy*. Chico, CA: AK Press.

Brayne, Sarah. 2014. “Surveillance and System Avoidance: Criminal Justice Contact and Institutional Attachment.” *American Sociological Review* 79 (3): 367–91.

Cech, Erin A. 2013. “The (Mis)Framing of Social Justice: Why Ideologies of Depoliticization and Meritocracy Hinder Engineers’ Ability to Think About Social Injustices.” In *Engineering Education for Social Justice*, edited by Juan Lucena, 10:67–84. Philosophy of Engineering and Technology. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-6350-0_4.

Cech, Erin A., and Heidi M. Sherick. 2015. “Depoliticization and the Structure on Engineering Education.” In *The Depoliticization Of*, 1st ed. 2015, 203–16. Philosophy of Engineering and Technology 20. Cham: Springer International Publishing : Imprint: Springer. https://doi.org/10.1007/978-3-319-16169-3.

Chouldechova, Alexandra. 2017. “Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments.” *Big Data* 5 (2): 153–63.

Christin, Angèle. 2017. “Algorithms in Practice: Comparing Web Journalism and Criminal Justice.” *Big Data & Society* 4 (2): 205395171771885. https://doi.org/10.1177/2053951717718855.

Cifor, Marika, P. Garcia, T.L. Cowan, J. Rault, Tonia Sutherland, A. Chan, J. Rode, Anna Lauren Hoffmann, N. Salehi, and Lisa Nakamura. 2019. “Feminist Data Manifest-No.” 2019. https://www.manifestno.com/home.

Comey, James. 2015. “Law Enforcement and Race Relations.” C-SPAN. 2015. https://www.c-span.org/video/?324342-1/fbi-director-james-comey-law-enforcement-race-relations.

Courtwatch MA. 2018. “First 100 Days.” *CourtWatch MA* (blog). 2018. https://www.courtwatchma.org/first-100-days.html.

Crenshaw, Kimberlé. 2012. “From Private Violence to Mass Incarceration: Thinking Intersectionally about Women, Race, and Social Control.” *UCLA Law Review* 59 (6): 1418–73.

D’Ignazio, Catherine, and Lauren F. Klein. 2020. *Data Feminism*. The MIT Press. https://doi.org/10.7551/mitpress/11805.001.0001.

Dolovich, Sharon. 2011. “Exclusion and Control in the Carceral State.” *Berkeley J. Crim. L.* 16: 259.

Dutta, Mohan, and Mahuya Pal. 2010. “Dialog Theory in Marginalized Settings: A Subaltern Studies Approach.” *Communication Theory* 20 (4): 363–86. https://doi.org/10.1111/j.1468-2885.2010.01367.x.

Eubanks, Virginia. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin’s Publishing Group.

Friedman, Brittany. 2021. “Toward a Critical Race Theory of Prison Order in the Wake of COVID-19 and Its Afterlives: When Disaster Collides with Institutional Death by Design.” *Sociological Perspectives*, April, 073112142110054. https://doi.org/10.1177/07311214211005485.

Fryer, Roland G. 2016. “An Empirical Analysis of Racial Differences in Police Use of Force.” *National Bureau of Economic Research* 22399: 63.

Gangadharan, Seeta. 2020. “Context, Research, Refusal: Perspectives on Abstract Problem-Solving.” April 30, 2020. https://www.odbproject.org/2020/04/30/context-research-refusal-perspectives-on-abstract-problem-solving/.

Gilmore, Ruth Wilson. 2002. “Fatal Couplings of Power and Difference: Notes on Racism and Geography.” *The Professional Geographer* 54 (1): 15–24.

———. 2007. *Golden Gulag: Prisons, Surplus, Crisis, and Opposition in Globalizing California*. American Crossroads. Berkeley: University of California Press.

———. 2014. *Forward to: Struggle within: Prisons, Political Prisoners, and Mass Movements in the United States*. Oakland, CA: PM Press.

———. 2017. “Abolition Geography and the Problem of Innocence.” In *Futures of Black Radicalism*, 224–41. Brooklyn, New York: Verso.

Graeff, Erhardt. 2020. “The Responsibility to Not Design and the Need for Citizen Professionalism.” *Tech Otherwise*, May. https://doi.org/10.21428/93b2c832.c8387014.

Graziani, Terra, and Mary Shi. 2020. “Data for Justice: Tensions and Lessons from the Anti-Eviction Mapping Project’s Work Between Academia and Activism.” *ACME: An International Journal for Critical Geographies* 19 (1): 397–412.

Harris, David A. 2003. “The Reality of Racial Disparity in Criminal Justice: The Significance of Data Collection.” *Law & Contemp. Probs.* 66: 71.

Hartman, Saidiya V. 1997. *Scenes of Subjection: Terror, Slavery, and Self-Making in Nineteenth-Century America*. Oxford University Press.

Hoffmann, Anna Lauren. 2020. “Terms of Inclusion: Data, Discourse, Violence.” *New Media & Society*, September, 146144482095872. https://doi.org/10.1177/1461444820958725.
———. 2021. “Even When You Are a Solution You Are a Problem: An Uncomfortable Reflection on Feminist Data Ethics.” *Global Perspectives* 2 (1): 21335. https://doi.org/10.1525/gp.2021.21335.
Honig, Bonnie. 2021. *A Feminist Theory of Refusal*. *A Feminist Theory of Refusal*. Harvard University Press. https://www.degruyter.com/document/doi/10.4159/9780674259249/html.
hooks, bell. 1989. “CHOOSING THE MARGIN AS A SPACE OF RADICAL OPENNESS.” *Framework: The Journal of Cinema and Media*, no. 36: 15–23.
———. 1990. “Marginality as a Site of Resistance.” In *Out There: Marginalization and Contemporary Cultures*, 241–43. Cambridge, Mass: MIT Press.
Hope, Alexis, Catherine D’Ignazio, Josephine Hoy, Rebecca Michelson, Jennifer Roberts, Kate Krontiris, and Ethan Zuckerman. 2019. “Hackathons as Participatory Design: Iterating Feminist Utopias.” In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14. Glasgow Scotland Uk: ACM. https://doi.org/10.1145/3290605.3300291.
Hu, Lily. 2021. “Race, Policing, and the Limits of Social Science.” Text. Boston Review. May 6, 2021. http://bostonreview.net/science-nature-race/lily-hu-race-policing-and-limits-social-science.
Irani, Lilly, and Khalid Alexander. 2021. “The Oversight Bloc.” *Logic Magazine*, December 25, 2021. https://logicmag.io/beacons/the-oversight-bloc/.
Irani, Lilly, Janet Vertesi, Paul Dourish, Kavita Philip, and Rebecca E. Grinter. 2010. “Postcolonial Computing: A Lens on Design and Development.” In *Proceedings of the 28th International Conference on Human Factors in Computing Systems - CHI ’10*, 1311. Atlanta, Georgia, USA: ACM Press. https://doi.org/10.1145/1753326.1753522.
Kaba, Mariame. 2020. “Interrupting Criminalization - What’s Next? Safer and More Just Communities Without Policing - Page 1.” Interrupting Criminaization: Research in Action. Project NIA. https://view.publitas.com/interrupting-criminalization-byekyy37zyrk/whats-next-safer-and-more-just-communities-without-policing/page/1.
———. 2021. *We Do This ’Til We Free Us: Abolitionist Organizing and Transforming Justice*. Haymarket Books.
Kahn, Natalie, and Simon Levien. 2021. “SEAS Cancels Class on Controversial Policing Strategy After Student Petition | News | The Harvard Crimson.” January 26, 2021. https://www.thecrimson.com/article/2021/1/26/seas-cancels-policing-course/.
Katz, Yarden. 2020. *Artificial Whiteness: Politics and Ideology in Artificial Intelligence*. Columbia University Press.
Kilgore, James. 2017. “Electronic Monitoring: A Survey of the Research for Decarceration Activists.” Challenging E-Carceration. July 2017. https://www.challengingecarceration.org/wp-content/uploads/2018/07/Survey-of-EM-Research.pdf.
Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. 2018. “Algorithmic Fairness.” In , 108:22–27.
Kubiak, Sheryl Pimlott, Woo Jong Kim, Gina Fedock, and Deborah Bybee. 2015. “Testing a Violence-Prevention Intervention for Incarcerated Women Using a Randomized Control

Trial." *Research on Social Work Practice* 25 (3): 334–48. https://doi.org/10.1177/1049731514534300.
Lacy, Akela, Alice Speri, Jordan Smith, and Sam Biddle. 2020. "Prisons Launch 'Absurd' Attempt to Detect Coronavirus in Inmate Phone Calls." *The Intercept*, April 21, 2020. https://theintercept.com/2020/04/21/prisons-inmates-coronavirus-monitoring-surveillance-verus/.
LEO Technologies. 2020a. "LEO Technologies and Verus: Supporting Our Nation's Correctional Facilities During the COVID-19 Pandemic." *LEO Technologies* (blog). March 19, 2020. https://leotechnologies.com/leo-technologies-and-verus-supporting-our-nations-correctional-facilities-during-the-covid-19-pandemic/.
———. 2020b. "What Is LEO Technologies?" *LEO Technologies* (blog). June 29, 2020. https://leotechnologies.com/what-is-leo-technologies/.
———. n.d. "How Verus Works." LEO Technologies. Accessed July 12, 2021. https://leotechnologies.com/services/verus/.
Lorde, Audre. 1984. *Sister Outsider: Essays and Speeches*. Berkeley, Calif: Crossing Press, c2007.
McGranahan, Carole. 2016. "Theorizing Refusal: An Introduction." *Cultural Anthropology* 31 (3): 319–25. https://doi.org/10.14506/ca31.3.01.
Media Justice, dir. 2021. *Points of Connection: Mapping Electronic Monitoring to Challenge E-Carceration*. https://www.youtube.com/watch?v=w_j-r8xqIZM.
Melamed, Jodi. 2019. "Operationalizing Racial Capitalism: Administrative Power and Ordinary Violence." Yale University, November 19. https://www.youtube.com/watch?v=o3Z9sOGf6BA&t=2675s.
Muhammad, Khalil Gibran, dir. 2015. *"The Condemnation of Blackness" - Khalil Gibran Muhammad Book Talk, May 6 2015*. John Jay College of Criminal Justice. https://www.youtube.com/watch?v=STKb-ai6874.
———. 2019. *The Condemnation of Blackness: Race, Crime, and the Making of Modern Urban America, With a New Preface*. Harvard University Press.
Murakawa, Naomi. 2014. *The First Civil Right: How Liberals Built Prison America*. Oxford University Press.
Nagar, Richa, and Roozbeh Shirazi. 2019. "Radical Vulnerability." In *Keywords in Radical Geography: Antipode at 50*, edited by Antipode Editorial Collective, Tariq Jazeel, Andy Kent, Katherine McKittrick, Nik Theodore, Sharad Chari, Paul Chatterton, et al., 236–42. Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/9781119558071.ch44.
Naples-Mitchell, Katherine. 2021. "Fool's Gold: How RCT Research Harms Communities Impacted by Criminal Punishment." Charles Hamilton Houston Institute for Race & Justice. January 26, 2021. https://charleshamiltonhouston.org/news/2021/01/fools-gold-how-rct-research-harms-communities-impacted-by-criminal-punishment/.
Norris, Samuel, Matthew Pecenco, and Jeffrey Weaver. 2021. "The Effects of Parental and Sibling Incarceration: Evidence from Ohio." SSRN Scholarly Paper ID 3590735. Rochester, NY: Social Science Research Network. https://doi.org/10.2139/ssrn.3590735.
North, Sandy. 2020. "New Study! Evaluating Counsel at First Appearance in Hays County, TX." The Access to Justice Lab. July 7, 2020. https://a2jlab.org/new-study-evaluating-counsel-at-first-appearance-in-hays-county-tx/.

Onuoha, Mimi. 2020. “When Proof Is Not Enough.” *FiveThirtyEight* (blog). July 1, 2020. https://fivethirtyeight.com/features/when-proof-is-not-enough/.
Prins, Seth J, and Adam Reich. 2018. “Can We Avoid Reductionism in Risk Reduction?” *Theoretical Criminology* 22 (2): 258–78.
Raji, Inioluwa Deborah, Morgan Klaus Scheuerman, and Razvan Amironesei. 2021. “You Can’t Sit With Us: Exclusionary Pedagogy in AI Ethics Education.” In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 515–25. Virtual Event Canada: ACM. https://doi.org/10.1145/3442188.3445914.
Roberts, Dorothy. 2019. “Digitizing the Carceral State.” *Harvard Law Review* 132: 1695–1713.
Saloner, Brendan, Kalind Parish, Julie A. Ward, Grace DiLaura, and Sharon Dolovich. 2020. “COVID-19 Cases and Deaths in Federal and State Prisons.” *Journal of the American Medical Associatiion* 324 (6): 602. https://doi.org/10.1001/jama.2020.12528.
Schneider, Avie, and Laura Sydell. 2019. “Microsoft Workers Protest Army Contract With Tech ‘Designed To Help People Kill.’” *NPR*, February 22, 2019, sec. Business. https://www.npr.org/2019/02/22/697110641/microsoft-workers-protest-army-contract-with-tech-designed-to-help-people-kill.
Schnepel, Kevin T. 2020. “Covid-19 in U.S. State and Federal Prisons.” National Commission on Covid-19 and Criminal Justice.
Schrader, Stuart. 2019. *Badges without Borders: How Global Counterinsurgency Transformed American Policing*. American Crossroads 56. Oakland, California: University of California Press.
Shah, Nishant. 2015. “Networked Margins: Revisiting Inequality and Intersection.” Digitally Connected: Global Perspectives on Youth and Digital Media. http://www.ssrn.com/abstract=2585686.
Simpson, Audra. 2014. *Mohawk Interruptus: Political Life across the Borders of Settler States*. Durham: Duke University Press.
Spade, Dean. 2015. *Normal Life: Administrative Violence, Critical Trans Politics, and the Limits of Law*. Revised and Expanded edition. Durham, [North Carolina]: Duke University Press.
Stanford Computational Policy Lab. n.d. “Driving Social Impact through Technical Innovation.” Stanford Computational Policy Lab. Accessed April 2, 2019. https://policylab.stanford.edu/.
Stevenson, Megan T. 2017. “Assessing Risk Assessment in Action.” *Minnesota Law Review* 103: 303–71. https://doi.org/10.2139/ssrn.3016088.
Stoller, Kristin. 2019. “Texas Billionaire John Arnold Gives $39 Million To Reform America’s Broken Bail System.” *Forbes*, 2019. https://www.forbes.com/sites/kristinstoller/2019/03/19/texas-billionaire-john-arnold-gives-39-million-to-reform-americas-broken-bail-system/.
Stop LAPD Spying Coalition. 2018. “StopLAPDSpying_Before-the-Bullet-Hits-the-Body-May-8-2018.Pdf.” Stop LAPD Spying Coalition. https://stoplapdspying.org/wp-content/uploads/2018/05/Before-the-Bullet-Hits-the-Body-May-8-2018.pdf.
Sutherland, Tonia. 2019. “The Carceral Archive: Documentary Records, Narrative Construction, and Predictive Risk Assessment.” *Journal of Cultural Analytics*. https://doi.org/10.22148/16.039.
Thompson, Jonathan. 2020. “National Sheriffs’ Association Teams with LEO Technologies on COVID-19 Industry Action Group for Correctional Facilities | NATIONAL SHERIFFS’

ASSOCIATION." April 14, 2020. https://www.sheriffs.org/National-Sheriffs%E2%80%99-Association-Teams-LEO-Technologies-COVID-19-Industry-Action-Group-for.

Tuck, Eve, and K. Wayne Yang. 2014a. "R-Words: Refusing Research." In *R-Words: Refusing Research*. 1 Oliver's Yard, 55 City Road London EC1Y 1SP: SAGE Publications, Inc. https://doi.org/10.4135/9781544329611.

———. 2014b. "Unbecoming Claims: Pedagogies of Refusal in Qualitative Research." *Qualitative Inquiry* 20 (6): 811–18. https://doi.org/10.1177/1077800414530265.

Wakabayashi, Daisuke, and Scott Shane. 2018. "Google Will Not Renew Pentagon Contract That Upset Employees." *The New York Times*, June 1, 2018, sec. Technology. https://www.nytimes.com/2018/06/01/technology/google-pentagon-project-maven.html.

Wang, Jackie. 2018. *Carceral Capitalism*. Semiotext(e).

Weld, Kirsten. 2014. *Paper Cadavers: The Archives of Dictatorship in Guatemala*. American Encounters/Global Interactions. Durham: Duke University Press.

Wright, Sarah. 2018. "When Dialogue Means Refusal." *Dialogues in Human Geography* 8 (2): 128–32. https://doi.org/10.1177/2043820618780570.

Zahara, Alex. 2016. "Ethnographic Refusal: A How to Guide." *Discard Studies*, August, 5.

Zong, Jonathan. 2020. "From Individual Consent to Collective Refusal: Changing Attitudes toward (Mis)Use of Personal Data." *XRDS: Crossroads, The ACM Magazine for Students* 27 (2): 26–29. https://doi.org/10.1145/3433140.

Zong, Jonathan, and J. Nathan Matias. 2020. "Building Collective Power to Refuse Harmful Data Systems." *Citizens and Technology Lab* (blog). August 12, 2020. https://citizensandtech.org/2020/08/collective-refusal/.

Chapter 5

Open Letters: On the Value of Imperfect Experiments

Introduction

"On the Moral Collapse of AI Ethics" — that was the title of the essay that Khadijah Abdurahman penned in the days following the high-profile firing of Dr. Timnit Gebru, former co-lead of Google's Ethical AI team. Dr. Gebru's termination set off a wave of outrage within Google and the broader AI ethics community, who viewed the company's decision as a retaliatory act of censorship. Gebru had recently co-authored a paper outlining the environmental and social risks of large language models, a technology that is central to Google's business. The article sparked a heated debate within the company, in which senior managers demanded that Gebru withdraw the paper from consideration for a conference, or remove her name from its list of authors. Gebru pushed back, saying she would only remove her name if Google agreed to set up a more transparent process for reviewing future research. She was fired the following day.

As the details surrounding Dr. Gebru's termination came to light, hundreds of Google staff members and other supporters signed an open letter, highlighting the importance of Gebru's research contributions and the fact that she was one of very few Black women Research Scientists at the company. The letter framed her termination as "an act of retaliation against Dr. Gebru, and it heralds danger for people working for ethical and just AI — especially Black people and People of Color — across Google" (Google Walkout for Real Change 2020). It was in this context that Abdurahman sat down to write their essay. Rather than focus on the details of Gebru's termination, Abdurahman turned a critical eye toward the response that her firing had generated amongst computer scientists, technologists, and academics working on ethical AI.

"Out of the love I got for [Gebru] and this community… I feel the need to be harsh and keep it real about the moral collapse of AI Ethics," Abdurahman (2020) wrote, "If demands for corporate transparency crystalized in the Standing with Dr. Timnit Gebru Petition defines the horizon for tech worker resistance, we are doomed."

Abdurahman (2020) went on to question the ways that the story of Gebru's firing had been framed as a tale of "the lone researcher against Goliath," arguing that such narratives perpetuated an unhealthy culture of celebrity within the AI ethics community and distracted from more fundamental issues regarding Big Tech's role in propping up violent systems of extraction and exploitation, like the ones detailed in Gebru et al's paper (Bender et al. 2021).

In a follow-up essay, Abdurahman drew parallels between the open letter to Google and the Franck report, an internal memo in which scientists working on the Manhattan Project raised grave concerns about the development of nuclear weapons, just two months before atomic bombs were dropped on Hiroshima and Nagasaki (Abdurahman 2021). Abdurahman argued that the scientists who wrote the Franck report had gravely misunderstood the political landscape surrounding the powerful technologies that they'd helped to develop.

"Rather than siphon the scientific knowledge they had accrued in developing nuclear weapons out of the lab and into the commons in order to build a mass movement, they waited until the final hour to pen a letter, addressed to a government that would never heed their call" (Abdurahman 2021). Abdurahman went on to argue that open letters like the one penned in the wake of Dr. Gebru's firing similarly misunderstood the power dynamics that shaped the development and deployment of AI, and that such efforts were in no way proportional to the threats looming on the horizon, as our political and economic systems become increasingly shaped by artificial intelligence.

I really appreciated Abdurahman's line of thinking in these essays. They put words to a nagging feeling that I'd had as I watched the hashtags and twitter threads circulate in the wake of Dr. Gebru's firing. On the one hand, Gebru's termination had sparked some fruitful dialogue regarding the ways that Black women's work is systematically silenced and erased by powerful organizations. For example, in "The Abuse and Misogynoir Playbook," Katlyn Turner, Danielle Wood and Catherine D'Ignazio (2021) drew connections between the abusive tactics used by Google to silence Dr. Gebru, and a long history of misogynoir policies used to shame and silence Black women.

But generally speaking, much less energy had been spent discussing the kinds of fundamental critiques outlined in Gebru et al's article, regarding the ways that Big Tech perpetuated broader forms of exploitation and social exclusion through their products, partnerships, and private lobbying efforts. It felt like Dr. Gebru's termination had sparked yet another cycle of misplaced outrage that had generated a lot of energy, but ultimately failed to get at the heart of the matter. So I appreciated Abdurahman's attempt to push the conversation forward, by encouraging those who felt angered on Gebru's behalf to channel that energy inward, to examine the ways their own efforts to make AI more ethical had fallen short. Their words were intended to unsettle readers out of comfortable modes of ally-ship and virtue signaling and into a more abolitionist stance of self-examination. As Abdurahman (2020) explained in their first essay,

"We have agency, my call out isn't for some grand afropessimist nihilism, it's to shake you all and say how we going to build this movement to change the world if you can't hit up an engineer to say you going down the wrong path or integrate the humanities beyond evidence for an apriori argument about bias."

Such calls for self-critique are rarely comfortable experiences. I certainly felt called out by Abdurahman's essays, particularly by the ways that they had situated open letters as being in

opposition to movement building. But I have since learned to interpret those initial feelings of defensiveness as an invitation to take a closer look at my own work.

I understand why Abdurahman pointed to the Google petition as a symbol of shallow political engagement. By late 2020, open letters had become a mainstay of AI activism (Belfield 2020). For example, in April 2018, over three thousand Google employees signed an open letter demanding that the company pull out of Project Maven, a Pentagon pilot program that aimed to develop a customized AI surveillance engine for the U.S. Department of Defense (Google Employees 2018). Two months after the open letter was published, Google announced that it would not renew the Project Maven contract. It also released a set of guiding principles for ethical engagement in future AI projects. However, just weeks later, employees raised the alarm about another secret project within Google, called Project Dragonfly. The goal of Project Dragonfly was to build a censored search engine for Chinese Internet users, raising concerns about the company's role in fueling a global "splinternet" (Kolodny 2018). Google employees opposed this project in another open letter (Google Employees 2019) and the company eventually decided to cancel that project as well.

However, it was during the response to Project Dragonfly that Google employees began to express concerns that such open letters were not a sustainable long-term organizing strategy. As one anonymous employee wrote in an email regarding the internal resistance to Project Dragonfly, "Individual employees organizing against the latest dubious project cannot be our only safeguard against unethical decisions. This amounts to unsustainable ethics whack-a-mole" (Belfield 2020). As the author of this email points out, open letters take a lot of work and, even when they led to immediate wins, they didn't appear to fundamentally change the way companies like Google decide which projects to take on and which ones to decline.

Moreover, not all open letters precipitate such immediate results as the examples outlined above. For example, there have been a number of open letters penned by employees of other major tech companies, such as Amazon, Microsoft and Salesforce, demanding that their companies withdraw from corporate partnerships with Immigration and Customs Enforcement (ICE) and Customs and Border Protection (CBP) (Employees of Microsoft 2018; Fight for the Future 2018; Amazon Employees for Climate Justice 2021). Yet these letters have not generated as clear or satisfying outcomes as the letters targeting Project Maven and Project Dragonfly.

This raises the question: beyond the articulation of specific demands for action from powerful actors and institutions, what good are open letters? By the time I was reading Abdurahman's essay, I'd collaborated on half a dozen or so open letters myself, arguing against the development and deployment of a wide range of carceral technologies. Abdulrahman's words pushed me to reflect more deeply on those efforts. I agree that open letters are not likely to change the minds of people in power, and that they can even be counterproductive when they become cathartic outlets for emotions that might be directed towards other long-term, collective interventions. But I also felt that the purpose of such letters could be more than just trying to change the minds of powerful people. Open letters have the potential to serve as vehicles for a deeper kind of refusal, one that changes the terms of debate and gives us an opportunity to enter into communities of struggle that extend far beyond the act of letter writing itself. At their best, open letters are a tip of the iceberg kind of activity, the most visible part of a much deeper set of ongoing efforts to resist and transform the status quo.

Enclosed in this chapter are two open letters that represent imperfect attempts at this second kind of letter writing practice. The first letter, entitled "Technical Flaws of Pretrial Risk

Assessment Raise Grave Concerns," was co-authored by me, Karthik Dinakar and Colin Doyle, during a time when pretrial risk assessments were spreading rapidly across the United States. While risk assessments were not new in the criminal legal system, they'd garnered renewed enthusiasm in recent years, as philanthropic organizations like Arnold Ventures and the MacArthur Foundation invested tens of millions of dollars to roll them out in counties and states across the United States (Stoller 2019). By 2019, jurisdictions in nearly every state in the country had adopted a pretrial risk assessment tool (Movement Alliance Project 2019).

It was during this period of time that a group of community organizers based out of Chicago approached Karthik, Colin, and me, requesting that we draft a statement outlining the fundamental flaws of these tools. By then, it was widely acknowledged that risk assessments weren't perfect, particularly when it came to their performance across different racial groups (Angwin et al. 2016). But many argued that they were better than the status quo and worth investing in as a vehicle for culture change in the courts (Sunstein 2019; Berk et al. 2018; Goel et al. 2018). Moreover, there was a growing body of literature within both academia and industry which aimed to assuage concerns about the technical limitations of such tools, by outlining standards and best practices for their ethical implementation (Partnership on AI 2019; Arnold Foundation n.d.; Bavitz et al. 2018; Berk 2019). Generally speaking, these frameworks aimed to minimize specific types of bias (sample bias, label bias, etc.) procedurally, through semi-regular validations, in order to maximize their purported accuracy and minimize well-established forms of statistical bias (Berk et al. 2018; Goel et al. 2018; Corbett-Davies and Goel 2018).

For community organizers who had been working on bail reform for many years, this discourse represented a grave distraction from more fundamental concerns about the ways that risk assessments were used to criminalize people through fundamentally flawed data and poorly

defined notions of risk (Community Justice Exchange 2019; Movement Alliance Project 2019; Community Justice Exchange 2022). They feared that such tools would give the appearance of introducing meaningful change, while simultaneously entrenching the very beliefs that had fueled mass pretrial incarceration in the first place (Community Justice Exchange 2019; Grace 2021).

So they asked my colleagues and me to draft a statement which challenged the premise that risk assessments could be ethically implemented, if only the right technical procedures were put into place. To do this, we needed to combine historical and legal perspectives on the flaws of criminal legal data with statistical notions of validity and bias, in order to make a decisive critique of the ways that risk assessments were fundamentally flawed and therefore not worth investing in. The hope was that such a critique could disrupt seemingly rigorous debates about the technical implementation of such tools, so that organizers could change the terms of the conversation surrounding bail reform.

Our critiques echoed many of the concerns that organizers were already making on the ground (Community Justice Exchange 2019). In fact, we developed many of the arguments in the letter through conversations with members of the Community Justice Exchange, who helped us to pinpoint specific ways that criminal legal data were used to criminalize marginalized populations. Our hope was that, by connecting their historicized arguments about the fundamental biases of criminal legal data to issues regarding the technical validity of such tools, we could give their concerns more weight in the broader discourse. In our letter, we emphasized that the limitations we raised were fundamental and could not be addressed through technical fixes. In doing so, we hoped to create space for organizers to propose alternative solutions to the problem of mass pretrial incarceration.

This open letter gave us an opportunity to redistribute some of the cultural capital that we had access to as a result of being scholars from prestigious universities like MIT and Harvard. We sought to further amplify this effect by soliciting the endorsement of other well-known scholars from prestigious universities, such as Martha Minow, Safiya Noble, and Ruha Benjamin. This approach is similar to other collaborations that community organizers have pursued with scholars, which aim to redistribute the epistemic privileges accumulated via the academy in order to amplify and reinforce arguments being made on the ground (Graziani and Shi 2020).

That being said, such tactics can be fraught — they run the risk of reinforcing knowledge hierarchies that effectively silence the voices of directly impacted people and maintain the scholar's privileged status as the primary arbiter of truth. We ultimately decided that a strategic engagement with prestige politics could be useful in this context, where concerns of organizers had been eclipsed by vigorous debates surrounding the technical implementation of risk assessments. Though this approach had its limitations, we agreed with our collaborators that it could still be useful. As Graziani and Shi (2020, 404) explain, "the world of actual politics is not as clear cut as the dualisms academic critique would imply, and working in direct coalition with community partners often requires using impure tools."

Once the letter was drafted, we worked with a network of organizers to identify jurisdictions where they were actively fighting against the adoption of pretrial risk assessments, starting with Illinois. To date, we've submitted the letter to lawmakers in five jurisdictions and circulated an abridged version of our arguments via an op-ed in *The New York Times* (Barabas, Dinakar, and Doyle 2019). It's unclear what the impact of this letter has been on the broader discourse surrounding pretrial risk assessment. We aren't able to draw a nice straight line

between the statement and any specific policy decisions. However, the letter was integral to a larger process of inter-coalition building between my academic collaborators and a network of abolitionist organizers working on bail reform around the country. As I discuss in Chapter 4, trust building is one of the main challenges that computational practitioners, and academics more generally, face when forging collaborations with directly impacted communities. There are long-standing structural reasons for this, which stem from the uneven distribution of material resources and symbolic power. The writing of this letter gave us the opportunity to earn the trust of some community advocates, by spending our time and effort to work on an intervention that they deemed worthwhile. For me, this process of relationship building was the primary value of this intervention.

The second letter enclosed in this chapter is entitled, "Abolish the #TechToPrisonPipeline: Crime prediction technology reproduces injustices and causes real harm," which I co-authored with Audrey Beard, Beth Semel, Theodora Dryer, and Sonja Solomun. NM Amado also played an integral role in helping us to disseminate the letter and pursue other potential campaigns in the following months. This collaboration formed online during the COVID-19 pandemic, after a press release circulated on Twitter, announcing the publication of a new research paper which claimed to have developed facial recognition software that was "capable of predicting whether someone is likely going to be a criminal" with eighty percent accuracy and no racial bias ("HU Facial Recognition Software Predicts Criminality | Harrisburg University" 2020).

In contrast to the first letter, I didn't know any of my collaborators ahead of time — we found each other online, as we discussed the need for some intervention that would put an end to the endless game of racist AI whack-a-mole that seemed to occur online every few months. Over

zoom, we discussed the ways that AI applications claiming to predict criminality based on physical characteristics were a part of a long legacy of discredited pseudosciences such as physiognomy and phrenology. This was not a new claim — by then, the connections between eighteenth and nineteenth century pseudoscience and facial recognition had been widely addressed (Dick 2019; Arcas, Mitchell, and Todorov 2017; Chinoy 2019). Yet, academics, law enforcement specialists, and politicians continued to repackage these widely debunked claims with shiny new branding in order to advocate for oppressive policing and prosecutorial tactics.

My collaborators and I discussed how frustrating it was to witness this. Such experiences bred feelings of cynicism and impotence — what was the point of debunking yet another problematic study if the same ideas were just going to be recycled a few months later, only with slightly different packaging? We decided that, in addition to debunking the claims made in this specific study, we would make a call for computer scientists to continuously interrogate the epistemic norms which enable the same tired and racist ideas to be recycled over time. Our small group came from an eclectic set of academic disciplines, ranging from history and anthropology to computer science and media studies. As a result, we were able to tie together disparate conversations from each of our respective fields, in order to forge a shared language that helped us to make sense of not only this particular study, but also the structural, epistemic, and political factors which made it possible for such studies to re-emerge over time.

The result is a letter that operates at two levels. The first argument is in the body of the text itself, where we focus on debunking claims that biological or criminal legal data can ever be used to measure or predict criminality. The second argument resides in the footnotes, which take up more space than the body of the text. In the footnotes, we emphasize that the claims we are

making are not new — specifically, people of color have been making such arguments for decades, yet the same harmful ideas continue to re-emerge, time and again.

One of the things I appreciated most about this collaboration was the way our group centered an ethic of care in our interactions, prioritizing one another's survival and well-being over any concerns for output or timeline (Hobart and Kneese 2020; Sevenhuijsen 2003; Spade 2020). This was especially important during the height of the COVID-19 pandemic, when we were all managing additional risks and responsibilities in our daily lives, just to keep ourselves and our loved ones healthy. We also understood that caring for ourselves and one another was an essential aspect of radical praxis, one that enabled us to transcend the competitive and productivity-obsessed mindset of academia in order to enact an "otherwise" set of relations in our work (Hobart and Kneese 2020). As a result, we were able to bring a level of generosity and vulnerability to our collaboration that I had not experienced in prior projects. I began to feel a sense of kinship with my collaborators that helped me to overcome feelings of cynicism and isolation that had been lurking in the background of my work for months.

During the process of drafting this open letter, George Floyd was killed by police officers in Minneapolis, Minnesota. His death sparked protests across the country and imbued our work with a deeper sense of urgency. We ratcheted up our effort to connect our arguments to the demands of Black-led social movements, such as the Movement for Black Lives. And we became even more intentional about situating our work within a long lineage of struggles led by people of color to dismantle the carceral state. As we prepared to circulate the letter, we reached out to a group of abolitionist organizers who were focused specifically on fighting against criminalizing technology, in order to get their feedback on the letter and guidance on how to best circulate it online. Our group was composed entirely of White, queer, non-binary and/or femme-

identified people with varying degrees of involvement in social movement work, so we wanted to make sure that we were careful about building structures of accountability for the arguments that we made and the way we used our platform to bolster broader movement work led by people of color. Through these conversations, we were able to take small but meaningful steps to ensure that we didn't center ourselves in the work or distract from ongoing efforts led by people of color. For example, we checked in with this group as we brainstormed names for our coalition, in order to make sure we didn't come up with a name that was too similar to groups that already existed, such as the Critical Tech Resistance Network.

The open letter ultimately dropped in June 2020 and received 2,425 signatures. The article that had provoked us to write the letter in the first place was rescinded from publication. In the wake of this early win, our group was eager to identify other interventions to counter harmful and racist research within the academy. A few months later, we decided to put together a multimedia project that included an interactive timeline, illustrating the ways that certain racist ideas kept coming back to life within the academy, even after they'd been debunked multiple times. Our plan was to dub such ideas as "zombie science," and release it on Halloween.

However, this time we weren't as careful about taking care of ourselves in the work or checking in with our accountability partners about the premise and rollout of the project. Given that we had a hard deadline that we were trying to meet, many of us pushed ourselves past the point of being healthy, pulling long hours after our day jobs to get the work done. Communication wasn't as thoughtful as it had been before. Grudges developed about the distribution of work in the group. And, most importantly, we didn't check in with our accountability partners until just a few days before we planned to release the project.

When we did, they expressed concerns with how we'd framed the project in terms of "zombie science." They pointed out that zombies were a racially charged trope, originally based on Haitian folklore that spoke to the lack of agency that enslaved people had over their own bodies, even in death (Moreman and Rushton 2011). That trope had been diluted and distorted through American pop culture in some troubling ways (Bishop 2008; Mariani 2015). This was an issue that a member of our coalition, Theodora Dryer, had actually brought up to us in the early stages of the project, but we did not take enough time to thoughtfully consider her concerns. As a result, that history was not well accounted for in our handling of zombie imagery, and our accountability partners asked that we not release the project as planned. We decided to pull the plug just a couple of days before the project was supposed to be released. This decision was a huge blow to the morale of our coalition. We were exhausted and disappointed, and the goodwill that we'd so carefully cultivated over the course of our first campaign was severely diminished.

Our group chat went quiet soon after that, as people prioritized self-care and moved on to other projects. Since then, we have not been able to muster the energy to pursue other interventions or substantially engage with offers to collaborate with other groups. But I've come to peace with the fact that, even if our work as a coalition was fleeting, there were bonds that formed during that process which will endure for a long time. I consider a couple of people from the coalition to be dear friends and generative thought partners. They send my kid care packages halfway across the world. And they offer valuable feedback on my writing and ideas, many of which have significantly impacted the writing of this dissertation.[10] There are other members of the coalition with whom I get the chance to speak only sporadically. But I continue to harbor immense respect for their work and the way that they live their lives. Every person in our

[10] Special shout-out to Dr. Beth Semel, who pointed me to Laurence Ralph's book, *The Torture Letters*, and the concept of "ethnographic lettering, which greatly influenced my writing in Chapter 6.

coalition is a role model to me, someone whose life I look to as a guiding example as I stumble my way through trying to live out my values in my work.

I've also come to appreciate how the failed zombie project enabled us to strengthen bonds with our outside partners. It's one thing to seek out accountability when it requires little sacrifice or risk. But it's a whole other thing to invest in such relationships when they can result in disappointment, embarrassment, or regret. I really appreciate the level of grace that our accountability partners gave us when they asked us to not publish the project. They helped us to see the limitations of our thinking, but they didn't dismiss us as people, and they encouraged us to keep putting in the work. It was clear that their feedback stemmed from a real commitment to help us learn and grow from our mistakes. Moreover, the experience gave us the opportunity to demonstrate that we took their perspective seriously, even if the outcome was disappointing for us. Looking back on it now, I can appreciate the bittersweet benefits of messing up — it helped me to understand the radical potential of embracing failure as an essential opportunity to learn and build enduring relationships across difference.

I bring up this story to emphasize that collective action is difficult and messy work. It requires a kind of "radical vulnerability," in which we are constantly "reminding ourselves and one another of the violent histories and geographies that we inherit and embody despite our desires to disown them" (Nagar and Shirazi 2019, 239). I think that this is the kind of work that Abdurahman was trying to center when they spoke out against the open letter written in the wake of Dr. Gebru's termination. In fact, it was the kind of work that they were modeling through their essays, speaking uncomfortable truths out loud in the hopes that the community of AI ethics researchers might move beyond performative ally-ship and cathartic cycles of outrage, in order to form more durable forms of collective action.

For me, open letters have served as an important anchor point in the formation of collectivities that are both alive to the present moment, as well as in conversation with the communities of struggle that came before us and those that will follow in our footsteps. Both letters represent attempts to change the terms of debate on important topics, by being attentive to the ways those debates are situated in a much longer history of struggle. In the case of the letter on risk assessment, we connected deep historical analysis regarding the ways that criminal legal data has long been used to conflate Blackness with criminality, in order to make claims about the fundamental technical flaws of such tools. In the case of the facial recognition algorithm, we situated our arguments within a much longer lineage of abolitionist critique pioneered by Black activist intellectuals, in an effort to disrupt processes of recycling debunked race science within the academy.

Both letters represent experiments in trying to enact what Nagar and Shirazi (2019) describe as a "'blended but fractured we' across multiple axes of power and difference (Sangtin Writers and Nagar 2006)." Those experiments weren't perfect. One might argue that our risk assessment letter reified academic prestige and entrenched knowledge hierarchies by framing our arguments in technical terms. The failed zombie project is proof that even coalitions forged in a thoughtful care-based approach can slip back into harmful modes of operation. But both letters gave my collaborators and me a chance to experiment with new and exciting forms of collective resistance. Although our letters may not have persuaded powerful people to change their views or policies on a particular technology, the act of crafting our arguments helped my collaborators and I to develop a shared language that clarified and bridged our thinking across a wide range of disciplines, as well as across academic and activist spaces. This, in turn, helped to breathe new life into ongoing efforts to counter such technologies, by helping us to overcome the feelings of

hopelessness and cynicism that can diffuse the momentum behind social justice work over the long haul.

These efforts also provided practical opportunities to repair trust and address some of the structural inequalities which so often undermine collective struggles for justice across diverse groups. And they gave us the opportunity to grow as a result of our mistakes. I learned a lot from each of these experiences, lessons which continue to inform my relationship-based approach to abolitionist worldbuilding. To that end, I offer up these letters as examples of imperfect but necessary experiments in trying to, as Abdurahman (2021) puts it, "nourish an otherwise set of relations to each other while we strategize on getting free."

Works Referenced

Abdurahman, Khadijah. 2020. “On the Moral Collapse of AI Ethics.” December 7, 2020. https://upfromthecracks.medium.com/on-the-moral-collapse-of-ai-ethics-791cbc7df872.

———. 2021. “A Body of Work That Cannot Be Ignored.” *Logic Magazine*, December 25, 2021. https://logicmag.io/beacons/a-body-of-work-that-cannot-be-ignored/.

Amazon Employees for Climate Justice. 2021. “Public Letter to Jeff Bezos and the Amazon Board of Directors.” *Medium* (blog). January 20, 2021. https://amazonemployees4climatejustice.medium.com/public-letter-to-jeff-bezos-and-the-amazon-board-of-directors-82a8405f5e38.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. “Machine Bias.” *ProPublica*, May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Arcas, Blaise Aguera y, Margaret Mitchell, and Alexander Todorov. 2017. “Physiognomy’s New Clothes.” *Medium* (blog). May 20, 2017. https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a.

Arnold Foundation. n.d. “Arnold Ventures Statement of Principles on Pretrial Justice.” Arnold Foundation. Accessed April 2, 2019. https://www.arnoldventures.org/work/pretrial-justice/.

Barabas, Chelsea, Karthik Dinakar, and Colin Doyle. 2019. “Opinion | The Problems With Risk Assessment Tools - The New York Times,” July 17, 2019. https://www.nytimes.com/2019/07/17/opinion/pretrial-ai.html.

Bavitz, Christopher, Sam Bookman, Jonathan Eubank, Kira Hessekiel, and Vivek Krishnamurthy. 2018. “Assessing the Assessments: Lessons from Early State Experiences In the Procurement and Implementation of Risk Assessment Tools.” *Berkman Klein Center Research Publication*, no. 2018–8.

Belfield, Haydn. 2020. “Activism by the AI Community: Analysing Recent Achievements and Future Prospects.” In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 15–21. New York NY USA: ACM. https://doi.org/10.1145/3375627.3375814.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜.” In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23. Virtual Event Canada: ACM. https://doi.org/10.1145/3442188.3445922.

Berk, Richard. 2019. *Machine Learning Risk Assessments in Criminal Justice Settings*. 115-126: Springer.

Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. “Fairness in Criminal Justice Risk Assessments: The State of the Art.” *Sociological Methods & Research*, 110.

Bishop, Kyle. 2008. “The Sub-Subaltern Monster: Imperialist Hegemony and the Cinematic Voodoo Zombie.” *The Journal of American Culture* 31 (2): 141–52. https://doi.org/10.1111/j.1542-734X.2008.00668.x.

Chinoy, Sahil. 2019. “Opinion | The Racist History Behind Facial Recognition.” *The New York Times*, July 10, 2019, sec. Opinion. https://www.nytimes.com/2019/07/10/opinion/facial-recognition-race.html.

Community Justice Exchange. 2019. “An Organizers Guide to Confronting Pretrial Risk Assessment Tools in Decarceration Campaigns.”

https://static1.squarespace.com/static/60db97fe88031352b829d032/t/60dcd6e6ac9dae69c2243313/1625085675244/CJE_PretrialRATGuide_FinalDec2019Version.pdf.

———. 2022. “Overview: The Invisible Machinery of Data Criminalization.” From Data Criminalization to Prison Abolition. 2022. https://abolishdatacrim.org/en/report/overview-invisible-machinery-data-criminalization.

Corbett-Davies, Sam, and Sharad Goel. 2018. “The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning.” *ArXiv:1808.00023 [Cs]*, August. http://arxiv.org/abs/1808.00023.

Dick, dir. 2019. *“The Standard Head” - Stephanie Dick (Pennsylvania)*. Columbia University. https://www.youtube.com/watch?v=OLFSh37gves.

Employees of Microsoft. 2018. “An Open Letter to Microsoft: Don’t Bid on the US Military’s Project JEDI.” *Medium* (blog). October 16, 2018. https://medium.com/s/story/an-open-letter-to-microsoft-dont-bid-on-the-us-military-s-project-jedi-7279338b7132.

Fight for the Future. 2018. “An Open Letter to Salesforce: Drop Your Contract with CBP.” *Medium* (blog). August 9, 2018. https://fightfortheftr.medium.com/an-open-letter-to-salesforce-drop-your-contract-with-cbp-a8260841b627.

Goel, Sharad, Ravi Shroff, Jennifer L Skeem, and Christopher Slobogin. 2018. “The Accuracy, Equity, and Jurisprudence of Criminal Risk Assessment.” *Equity, and Jurisprudence of Criminal Risk Assessment (December 26, 2018)*.

Google Employees. 2018. “Open Letter: Project Maven,” 2018.

———. 2019. “We Are Google Employees. Google Must Drop Dragonfly.” *Medium* (blog). January 2, 2019. https://medium.com/@googlersagainstdragonfly/we-are-google-employees-google-must-drop-dragonfly-4c8a30c5e5eb.

Google Walkout for Real Change, Google Walkout For Real. 2020. “Standing with Dr. Timnit Gebru — #ISupportTimnit #BelieveBlackWomen.” *Medium* (blog). December 15, 2020. https://googlewalkout.medium.com/standing-with-dr-timnit-gebru-isupporttimnit-believeblackwomen-6dadc300d382.

Grace, Sharlyn. 2021. “‘Organizers Change What’s Possible’ | Sharlyn Grace.” *Inquest*, September 23, 2021. https://inquest.org/organizers-change-whats-possible/.

Graziani, Terra, and Mary Shi. 2020. “Data for Justice: Tensions and Lessons from the Anti-Eviction Mapping Project’s Work Between Academia and Activism.” *ACME: An International Journal for Critical Geographies* 19 (1): 397–412.

Hobart, Hi‘ilei Julia Kawehipuaakahaopulani, and Tamara Kneese. 2020. “Radical Care.” *Social Text* 38 (1): 1–16. https://doi.org/10.1215/01642472-7971067.

“HU Facial Recognition Software Predicts Criminality | Harrisburg University.” 2020. May 6, 2020. https://web.archive.org/web/20200506013352/https:/harrisburgu.edu/hu-facial-recognition-software-identifies-potential-criminals/.

Kolodny, Lora. 2018. “Former Google CEO Predicts the Internet Will Split in Two — and One Part Will Be Led by China.” CNBC. September 20, 2018. https://www.cnbc.com/2018/09/20/eric-schmidt-ex-google-ceo-predicts-internet-split-china.html.

Mariani, Mike. 2015. “The Tragic, Forgotten History of Zombies.” *The Atlantic*, October 28, 2015. https://www.theatlantic.com/entertainment/archive/2015/10/how-america-erased-the-tragic-history-of-the-zombie/412264/.

Moreman, Christopher M., and Cory James Rushton. 2011. *Race, Oppression and the Zombie: Essays on Cross-Cultural Appropriations of the Caribbean Tradition*. McFarland.

Movement Alliance Project. 2019. “Mapping Pretrial Injustice: A Community-Driven Database.” Mapping Pretrial Risk. 2019. https://pretrialrisk.com/.

Nagar, Richa, and Roozbeh Shirazi. 2019. “Radical Vulnerability.” In *Keywords in Radical Geography: Antipode at 50*, edited by Antipode Editorial Collective, Tariq Jazeel, Andy Kent, Katherine McKittrick, Nik Theodore, Sharad Chari, Paul Chatterton, et al., 236–42. Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/9781119558071.ch44.

Partnership on AI. 2019. “Report on Algorithmic Risk Assessment Tools in the U.S. Criminal Justice System.” Partnership on AI. file:///Users/chelsea.barabas/Downloads/Report-on-Algorithmic-Risk-Assessment-Tools.pdf.

Sevenhuijsen, Selma. 2003. *Citizenship and the Ethics of Care*. 0 ed. Routledge. https://doi.org/10.4324/9780203169384.

Spade, Dean. 2020. *Mutual Aid: Building Solidarity during This Crisis (and the Next)*. London ; New York: Verso.

Stoller, Kristin. 2019. “Texas Billionaire John Arnold Gives $39 Million To Reform America’s Broken Bail System.” *Forbes*, 2019. https://www.forbes.com/sites/kristinstoller/2019/03/19/texas-billionaire-john-arnold-gives-39-million-to-reform-americas-broken-bail-system/.

Sunstein, Cass R. 2019. “Algorithms, Correcting Biases.” *Social Research: An International Quarterly* 86 (2): 499–511.

Turner, Katlyn, Danielle Wood, and Catherine D’Ignazio. 2021. “The Abuse and Misogynoir Playbook.” Montreal: Montreal AI Ethics Institute. https://montrealethics.ai/wp-content/uploads/2021/01/State-of-AI-Ethics-Report-January-2021.pdf.

TECHNICAL FLAWS OF PRETRIAL RISK ASSESSMENTS RAISE GRAVE CONCERNS

Summary

Actuarial pretrial risk assessments suffer from serious technical flaws that undermine their accuracy, validity, and effectiveness. They do not accurately measure the risks that judges are required by law to consider. When predicting flight and danger, many tools use inexact and overly broad definitions of those risks. When predicting violence, no tool available today can adequately distinguish one person's risk of violence from another. Misleading risk labels hide the uncertainty of these high-stakes predictions and can lead judges to overestimate the risk and prevalence of pretrial violence. To generate predictions, risk assessments rely on deeply flawed data, such as historical records of arrests, charges, convictions, and sentences. This data is neither a reliable nor a neutral measure of underlying criminal activity. Decades of research have shown that, for the same conduct, African-American and Latinx people are more likely to be arrested, prosecuted, convicted and sentenced to harsher punishments than their white counterparts. Risk assessments that incorporate this distorted data will produce distorted results. These problems cannot be resolved with technical fixes. We thus strongly recommend turning to other reforms.

Actuarial Risk Assessments Do Not Accurately Measure Pretrial Risks

When making pretrial release decisions, judges must impose the least restrictive conditions of release necessary to secure the presence of a person at trial and protect the safety of the community. To accomplish this task, judges must identify and mitigate specific pretrial risks, specifically of a person causing serious harm to the community or fleeing the jurisdiction prior to their trial. Today's pretrial risk assessments are ill-equipped to support judges in evaluating and effectively intervening on these specific risks, because the outcomes that these tools measure do not match the risks that judges are required by law to consider. For example, many risk assessments only provide a pretrial failure risk score, which is a combined outcome of missing a court appearance or being rearrested. Many scholars have warned that such a composite score could lead to an overestimation of both flight and danger, and can make it more, not less, difficult to identify effective interventions.[11]

Even when pretrial risk assessments break out risk scores into distinct categories, the data used to define and measure flight and danger are inexact and overly broad. For example, risk assessments frequently define public safety risk as the probability of arrest.[12] When tools conflate the likelihood of arrest for any reason with risk of violence, a large number of people will be labeled a threat to public safety without sufficient justification. Risk assessments that include minor offenses, such as driving on a suspended license, in their

[11] E.g., Lauryn P. Gouldin, *Disentangling Flight Risk from Dangerousness*, 2016 BYU L. Rev. 837, 88788 (2018). The interventions which improve an individuals likelihood of appearing in court (text reminders, transportation services, flexible scheduling) are often quite different from interventions designed to ensure community safety (stay-away orders, curfews, drug testing).

[12] For example, the Colorado Pretrial Assessment Tool (CPAT) defines a risk to "public safety" as any new criminal filing, including for traffic stops and municipal offenses. The Colorado Pretrial Risk Assessment Tool Revised Report 18 (2012).

definition of danger, run the risk of increasing pretrial incarceration rates and further exacerbating racial inequalities in pretrial outcomes.[13]

Some risk assessments define public safety risk more narrowly as the risk that a person will be arrested for a violent crime while on pretrial release. But because pretrial violence is exceedingly rare, it is challenging to statistically predict. Risk assessments cannot identify people who are more likely than not to commit a violent crime. The fact is, the vast majority of even the highest risk individuals will not go on to be arrested for a violent crime while awaiting trial. Consider the dataset used to build the Public Safety Assessment (PSA): 92% of the people who were flagged for pretrial violence did not get arrested for a violent crime and 98% of the people who were not flagged did not get arrested for a violent crime.[14] If these tools were calibrated to be as accurate as possible, then they would predict that every person was unlikely to commit a violent crime while on pretrial release. Instead, risk assessments sacrifice accuracy and generate substantially more false positives (people who are flagged for violence but do not go on to commit a violent crime) than true positives (people who are flagged for violence and do go on to be arrested for a violent crime).[15] Consequently, violence risk assessments could easily lead judges to overestimate the risk of pretrial violence and detain more people than is justified.[16]

Finally, current risk assessment instruments are unable to distinguish one person's risk of violence from another's. In statistics, predictions are made within a range of likelihood, rather than as a single point estimate. For example, a predictive algorithm might confidently estimate a person's risk of arrest as somewhere between a range of five and fifteen percent. Studies have demonstrated that predictive models can only make reliable predictions about a person's risk of violence within very large ranges of likelihood, such as twenty to sixty percent.[17] As a result, virtually everyone's range of likelihood overlaps. When everyone is similar, it becomes impossible to differentiate people with low and high risks of violence. At present, there is no statistical remedy to this challenge.

Data Used to Build Pretrial Risk Assessments Are Distorted

Risk assessments are frequently posited as a solution to judges' implicit biases. Yet the data used to build pretrial risk assessments are deeply flawed and racially biased. Pretrial risk assessments rely on historical records of arrests, charges, convictions, and sentences to generate predictions about an individuals

---

[13] For decades, communities of color have been arrested at higher rates than their white counterparts, even for crimes that these racial groups engage in at comparable rates. As a result, people of color are more likely to be labeled as dangerous than their white counterparts when arrest data is used to measure public safety risk. Thus, they will bear a disproportionate amount of the burdens that stems from these harmful conflations between arrest and danger.

[14] Public Safety Assessment, PSA Results (2019).

[15] Julia Angwin et al., *Machine Bias*, Propublica (May 23, 2016), https://www.propublica.org/article/machinebias-risk-assessments-in-criminal-sentencing. These inaccuracies are very much mediated by race African Americans were twice as likely to be mislabeled as high risk than their white counterparts.

[16] For example, a recent study found that people significantly overestimate the recidivism rate for individuals who are labeled as moderate-high or "high" risk on a risk assessment. Daniel A., Krauss, Gabriel I. Cook & Lukas Klapatch, *Risk Assessment Communication Difficulties: An Empirical Examination of the Effects of Categorical Versus Probabilistic Risk Communication in Sexually Violent Predator Decisions*, Behav. Sci. & L. (2018). (Participants greatly overestimated the true recidivism rate for those assessed as moderate-high risk category – the true rate was less than fifty percent of what participants predicted.)

[17] Stephen D. Hart & David J. Cooke, *Another Look at the (Im–)Precision of Individual Risk Estimates Made Using Actuarial Risk Assessment Instruments*, 31 Behav. Sci. Law 81, (2013).

propensity for pretrial failure. These tools assume that criminal history data are a reliable and neutral measure of underlying criminal activity, but such records cannot be relied upon for this purpose. Arrest records are both under- and over-inclusive of the true crime rate. Arrest records are under-inclusive because they only chart law enforcement activity, and many crimes do not result in arrest.[18]8 Less than half of all reported violent crimes result in an arrest, and less than a quarter of reported property crimes result in an arrest. Arrest records are also over-inclusive because people are wrongly arrested and arrested for minor violations, including those that cannot result in jail time. Moreover, decades of research have shown that, for the same conduct, African-American and Latinx people are more likely to be arrested, prosecuted, convicted and sentenced to harsher punishments than their white counterparts.[19] People of color are treated more harshly than similarly situated white people at each stage of the legal system, which results in serious distortions in the data used to develop risk assessment tools:

• **Arrests**: For decades, communities of color have been arrested at higher rates than their white counterparts, even for crimes that these racial groups engage in at comparable rates.[20] For example, African-Americans are 83% more likely to be arrested for marijuana compared to whites at age 22 and 235% more likely to be arrested at age 27, in spite of similar marijuana usage rates across racial groups.[21] Similarly, African-American drivers are three times as likely as whites to be searched during routine traffic stops, even though police officers generally have a lower "hit rate" for contraband when they search drivers of color.[22] This leads to an overrepresentation of people of color in arrest data. Predictive algorithms that rely on this data overestimate pretrial risk for people of color.

• **Charges**: Empirical research has found that African-American defendants face significantly more severe charges than white defendants, even after controlling for a multitude of factors.[23] Persistent patterns of differential charging make prior charges an unreliable variable for building risk assessments.

---

[18] FBI, 2017 Crime in the United States: Clearances, https://ucr.fbi.gov/crime-in-the-u.s/2017/crime-in-the-u.s.-2017/topic-pages/clearances (last visited June 28, 2019).

[19] See generally The Sentencing Project, Report of the Sentencing Project to the United Nations Special Rapporteur on Contemporary Forms of Racism, Racial Discrimination, Xenophobia, and Related Intolerance Regarding Racial Disparities in the United States Criminal Justice System (2018); Lynn Langton & Matthew Durose, U.S. Dep't of Justice, Police Behavior During Traffic and Street Stops, 2011 (2013); Stephen Demuth & Darrell Steffensmeier, *The Impact of Gender and Race-Ethnicity in the Pretrial Release Process*, 51 Soc. Probs. 222 (2004); Jessica Eaglin & Danyelle Solomon, Brennan Center for Justice, Reducing Racial and Ethnic Disparities in Jails: Recommendations for Local Practice (2015); Sonja B. Starr & M. Marit Rehavi, *Racial Disparity in Federal Criminal Sentences*, J. Pol. Econ. 1320 (2014); Marc Mauer, Justice for All? Challenging Racial Disparities in the Criminal Justice System (2010).

[20] Megan Stevenson & Sandra G. Mayson, *The Scale of Misdemeanor Justice*, 98 B.U. L. Rev. 731, 769-770 (2018). This comprehensive national review of misdemeanor arrest data has shown systemic and persistent racial disparities for most misdemeanor offences. The study shows that "black arrest rate is at least twice as high as the white arrest rate for disorderly conduct, drug possession, simple assault, theft, vagrancy, and vandalism." Id. at 759. This study shows that "many misdemeanor offenses criminalize activities that are not universally considered wrongful, and are often symptoms of poverty, mental illness, or addiction." Id. at 766.

[21] "[R]acial disparity in drug arrests between black and whites cannot be explained by race differences in the extent of drug offending, nor the nature of drug offending." Ojmarrh Mitchell & Michael S. Caudy, *Examining Racial Disparities in Drug Arrests*, Just. Q., Jan. 2013, at 22.

[22] Ending Racial Profiling in America: Hearing Before the Subcomm. on the Constitution, Civil Rights and Human Rights of the Comm. on the Judiciary, 112th Cong. 8 (2012) (statement of David A. Harris).

[23] Sonja B. Starr M. Marit Rehavi, *Racial Disparity in Federal Criminal Charging and its Sentencing Consequences*, U. (Mich. L. Econ. Working Paper Series, Working Paper No. 12-002, 2012).

• **Convictions & Sentences**: Compared to similarly situated white people, African-Americans are more likely to be convicted[24] and more likely to be sentenced to incarceration.[25]

Risk assessments that incorporate this distorted data will produce distorted results.[26] There are no technical fixes for these distortions.

Conclusion

Pretrial risk assessments do not guarantee or even increase the likelihood of better pretrial outcomes. Risk assessment tools can simply shift or obscure problems with current pretrial practices. Some jurisdictions that have adopted risk assessment tools have seen positive trends in pretrial outcomes, but other jurisdictions have experienced the opposite. Within jurisdictions that have achieved positive outcomes, it is uncertain whether the risk assessment tools were responsible for that success or whether that success is due to other reforms or changes that happened at the same time. Given these mixed outcomes, it is impossible to predict the impact of pretrial risk assessments in any jurisdiction.

Beyond the technical flaws outlined in this statement, a broader and growing body of research questions the validity, ethics, and efficacy of actuarial pretrial risk assessments. For example, most risk assessments are proprietary technology, and defendants assessed by these tools are not allowed the opportunity to inspect and critique the algorithms or their underlying data. Poor implementation and lack of judicial training and buy-in can undermine reforms. Validity and fairness questions arise when tools are trained on data from one jurisdiction but deployed in a jurisdiction with different demographics, judicial culture, and policing practices.

This letter specifically addresses fundamental, technical problems with actuarial risk assessment instruments. These technical problems cannot be resolved. We strongly recommend turning to other reforms.

[24] Shamena Anwar, Patrick Bayer & Randi Hjalmarsson, *The Impact of Jury Race in Criminal Trials*, 127 **Q. J. Econ**. 1017, 1019 (2012).
[25] David S. Abrams, Marianne Bertrand & Sendhil Mullainathan, *Do Judges Vary in Their Treatment of Race*, 41 **J. L. Stud**. 347, 350 (2012)

[26] There have been attempts to solve this problem on the back end by mitigating outcome disparities in risk assessment predictions, but they overlook and do not address the fundamental distortions outlined above.

Signatories

Chelsea Barabas Research Scientist MIT Media Lab

John Bowers Research Associate Harvard Law School

Joy Buolamwini Founder Algorithmic Justice League

Ruha Benjamin, PhD Associate Professor Princeton University

Meredith Broussard, PhD Associate Professor New York University

Kate Crawford, PhD; Distinguished Professor, Co-Founder, Co-Director AI Now Institute

Karthik Dinakar, PhD Research Scientist MIT Media Lab

Timnit Gebru, PhD Co-founder Black in AI

Brook Hopkins, Executive Director Criminal Justice Policy Program Harvard Law School

Martha Minow, 300th Anniversary University Professor Harvard University

Cathy O'Neil, PhD Author Weapons of Math Destruction

Venerable Tenzin Priyadarshi Director, Ethics Initiative MIT Media Lab

Jason Schultz Professor NYU School of Law

Jordi Weinstock Lecturer on Law Harvard Law School

Ethan Zuckerman Director, Center for Civic Media MIT Media Lab

Colin Doyle Staff Attorney, CJPP Harvard Law School

Stefan Helmrich, PhD; Professor & Elting E. Morison Chair Anthropology, MIT

Joichi Ito, PhD, Director, MIT Media Lab Visiting Professor, Harvard Law School

Rodrigo Ochigame Doctoral Candidate, HASTS MIT

Heather Paxson, PhD Professor of Anthropology MIT

Rashida Richardson Director of Policy Research AI Now Institute, NYU

Vincent M. Southerland, Executive Director Center on Race, Inequality, & the Law NYU School of Law

Jonathan Zittrain George Bemis Professor of International Law Harvard Law School

Springer Publishing
Berlin, Germany
+49 (0) 6221 487 0
customerservice@springernature.com

RE: A Deep Neural Network Model to Predict Criminality Using Image Processing

June 22, 2020

Dear Springer Editorial Committee,

We write to you as expert researchers and practitioners across a variety of technical, scientific, and humanistic fields (including statistics, machine learning and artificial intelligence, law, sociology, history, communication studies and anthropology). Together, we share grave concerns regarding a forthcoming publication entitled "A Deep Neural Network Model to Predict Criminality Using Image Processing." According to a recent press release, this article will be published in your book series, "Springer Nature — Research Book Series: Transactions on Computational Science and Computational Intelligence."

We urge:

- **The review committee** to rescind the offer for publication of this specific study.
- **Springer** to issue a statement condemning the use of criminal justice statistics to predict criminality, and acknowledging their role in incentivising such harmful scholarship in the past.
- **All publishers** to refrain from publishing similar studies in the future.

This upcoming publication warrants a collective response because it is emblematic of a larger body of computational research that claims to identify or predict "criminality" using biometric and/or criminal legal data.[27] Such claims are based on unsound scientific premises, research, and

[27] Scholars use a variety of terms in reference to the prediction of criminal outcomes. Some researchers claim to predict "anti-social" or "impulsive" behavior. Others model "future recidivism" or an individual's "criminal tendencies." All of these terms frame criminal outcomes as the byproduct of highly individualized and proximate risk factors. As Prins and Reich (2018) argue, these predictive models neglect population drivers of crime and criminal justice involvement (Seth J. Prins, and Adam Reich. 2018. "Can we avoid reductionism in risk reduction?" *Theoretical criminology* 22 (2): 258-278). The hyper-focus on individualized notions of crime leads to myopic social reforms that intervene exclusively on the supposed cultural, biological and cognitive deficiencies of criminalized populations. This scholarship not only provides a mechanism for the confinement and control of the "dangerous classes," but also creates the very processes through which these populations are turned into deviants to be controlled and feared. As Robert Vargas (2020) argues, this type of scholarship "sees Black people and Black communities as in need of being fixed. This approach is not new but is rather the latest iteration in a series of efforts to improve cities by managing Black individuals instead of ending the police violence Black communities endure." Robert Vargas. 2020. "It's Time to Think Critically about the UChicago Crime Lab." *The Chicago Maroon* June 11. <https://www.chicagomaroon.com/article/2020/6/11/time-think-critically-uchicago-crime-lab/> (Accessed June 17, 2020). For examples of this type of criminalizing language see generally: Mahdi Hashemi and Margeret Hall. 2020. "Criminal tendency detection from facial images and the gender bias effect." *Journal of Big Data.* 7 (2):

methods, which numerous studies spanning our respective disciplines have debunked over the years.[28] Nevertheless, these discredited claims continue to resurface, often under the veneer of new and purportedly neutral statistical methods such as machine learning, the primary method of the publication in question.[29] In the past decade, government officials have embraced machine learning and artificial intelligence (AI) as a means of depoliticizing state violence and reasserting the legitimacy of the carceral state, often amid significant social upheaval.[30] Community

---

https://doi.org/10.1186/s40537-019-0282-4). Eyal Aharoni, et al. 2013. "Neuroprediction of future rearrest." *Proceedings of the National Academy of Sciences* 110 (15): 6223-6228. Xiaolin Wu and Xi Zhang. 2016. "Automated inference on criminality using face images." *arXiv preprint arXiv:1611.04135*: 4038-4052. Yaling Yang, Andrea L. Glenn, and Adrian Raine. 2008. "Brain abnormalities in antisocial individuals: implications for the law." *Behavioral sciences & the law* 26 (1): 65-83. Adrian Raine. 2014. *The anatomy of violence: The biological roots of crime*. Visalia: Vintage Press.

[28] AI applications that claim to predict criminality based on physical characteristics are a part of a legacy of long-discredited pseudosciences such as physiognomy and phrenology, which were and are used by academics, law enforcement specialists, and politicians to advocate for oppressive policing and prosecutorial tactics in poor and racialized communities. Indeed, in the opening pages of Hashemi and Hall (2020), the authors invoke the criminological studies of Cesare Lombroso, a dangerous proponent of social Darwinism whose studies the authors cited below overturn and debunk. In the late nineteenth and early twentieth century, police and other government officials relied on social scientists to create universalized measurements of who was "capable" of criminal behavior, based largely on a person's physical characteristics. This system is rooted in scientific racism and ultimately served to legitimize a regime of preemptive repression, harassment, and forced sterilization in racialized communities. The connections between eighteenth and nineteenth century pseudoscience and facial recognition have been widely addressed. For examples of the historical linkage between physiognomy, phrenology, and automated facial recognition, see Blaise Agüera y Arcas, Margaret Mitchell, and Alexander Todorov. 2017. "Physiognomy's New Clothes." *Medium*, May 6. <https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a>; on links between eugenics, race science, and facial recognition, see Sahil Chinoy. 2019. "The Racist History Behind Facial Recognition." *New York Times*, July 10. <https://www.nytimes.com/2019/07/10/opinion/facial-recognition-race.html>; Stephanie Dick, "The Standard Head," https://www.youtube.com/watch?v=OLFSh37gves.

[29] For example, Wu and Zhang (2016) bears a striking resemblance to the Harrisburg study and faced immense public and scientific critique, prompting the work to be rescinded from publication and the authors to issue a response (see Wu and Zhang, 2017. https://arxiv.org/pdf/1611.04135.pdf). Experts highlighted the utter lack of a causal relationship between visually observable identifiers on a face and the likelihood of a subject's participation in criminal behavior. In the absence of a plausible causal mechanism between the data and the target behavior, and indeed scientific rejection of a causal mechanism, the model is likely not doing what it claims to be doing. In this case, critics rightfully argued that the published model was not identifying criminality -- it was identifying historically disadvantaged ethnic subgroups, who are more likely to be targeted by police and arrested. For a summary of the critique see https://callingbullshit.org/case_studies/case_study_criminal_machine_learning.html. The fact that the current study claims its results have "no racial bias" is highly questionable, addressed further below in Sections I (for whether such a thing is possible) and II (whether metrics for bias really capture bias).

[30] As Jackie Wang (2018) argues, "'police science' is a way for police departments to rebrand themselves in the face of a crisis of legitimacy," pointing to internally generated data about arrests and incarcerations to justify their racially discriminatory practices. While these types of "evidence based" claims have been problematized and debunked numerous times throughout history, they continue to resurface under the guise of cutting-edge techno-reforms, such as "artificial intelligence." As Chelsea Barabas (2020, 41)

organizers and Black scholars have been at the forefront of the resistance against the use of AI technologies by law enforcement, with a particular focus on facial recognition.[31] Yet these voices continue to be marginalized, even as industry and the academy invests significant resources in building out "fair, accountable and transparent" practices for machine learning and AI.[32]

Part of the appeal of machine learning is that it is highly malleable — correlations useful for prediction or detection can be rationalized with any number of plausible causal mechanisms. Yet the way these studies are ultimately represented and interpreted is profoundly shaped by the political economy of data science[33] and

---

points out, "the term 'artificial intelligence' has been deployed as a means of justifying and de-politicizing the expansion of state and private surveillance amidst a growing crisis of legitimacy for the U.S. prison industrial complex." Sarah Brayne and Angèle Christin argue (2020, 1) that "predictive technologies do not replace, but rather *displace* discretion to less visible—and therefore less accountable—areas within organizations." Jackie Wang. 2018. *Carceral capitalism* (Vol. 21). MIT Press. Chelsea Barabas. 2020. "Beyond Bias: Reimagining the Terms of 'Ethical AI' in Criminal Law." 12 Geo. J. L. Mod. Critical Race Persp. 2 (forthcoming).<https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3377921#:~:text=Chelsea%20Barabas,-MIT%20Media%20Lab&text=The%20essay%20argues%20that%20attempts,the%20fairness%20of%20these%20tools>. Sarah Brayne and Angèle Christin. 2020. "Technologies of Crime Prediction: The Reception of Algorithms in Policing and Criminal Courts." *Social Problems*.

[31] The hard work of these organizers and scholars is beginning to gain public recognition. In recent weeks, major tech companies such as IBM and Amazon have announced commitments to stop collaborating with law enforcement to deploy facial recognition technologies. These political gains are the result of years of hard work by community organizations such as the Stop LAPD Spying Coalition, Media Mobilizing Project (renamed Movement Alliance Project), Mijente, The Carceral Tech Resistance Network, Media Justice, and AI For the People. This on-the-ground work has been bolstered by research led by Black scholars, such as Joy Buolamwini, Timnit Gebru, Mutale Nkonde and Inioluwa Deborah Raji. See: Joy Buolamwini and Timnit Gebru. 2018. "Gender shades: Intersectional accuracy disparities in commercial gender classification." *Conference on fairness, accountability and transparency*. Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. "Saving Face: Investigating the Ethical Concerns of Facial Recognition Auditing." ArXiv:2001.00964 [Cs], January 3, 2020. Mutale Nkonde. 2020. "Automated Anti-Blackness: Facial Recognition in Brooklyn, New York." Harvard Kennedy School Review: 30-36. https://hjaap.hkspublications.org/wp-content/uploads/sites/14/2020/05/Final-PDF-for-Website.pdf

[32] The Algorithmic Justice League has pointed out this blatant erasure of non-white and non-male voices in their public art project entitled, "Voicing Erasure" a project that was inspired in part by the work of Allison Koenecke, a woman researcher based at Stanford whose work uncovering biases in speech recognition software was recently covered in the New York Times. Koenecke was not cited in the original New York Times article, even though she was the lead author of the research. Instead, a number of her colleagues were named and given credit for the work, all of whom are men. In "Voicing Erasure" Joy Buolamwini pushes us to reflect on "Whose voice do you hear when you think of intelligence, innovation and ideas that shape our worlds?" https://www.youtube.com/watch?v=SdCPbyDJtK0&feature=youtu.be

[33] Timnit Gebru points out that "the dominance of those who are the most powerful race/ethnicity in their location…combined with the concentration of power in a few locations around the world, has resulted in a technology that can benefit humanity but also has been shown to (intentionally or unintentionally) systematically discriminate against those who are already marginalized." Timnit Gebru. 2020. "Race and Gender." In *Oxford Handbook on AI Ethics*. Oxford Handbooks. Oxford University Press. Facial recognition research (arguably a subset of AI) is no different - it has never been neutral nor unbiased. In addition to its deep connection with phrenology and physiognomy, it is entwined with the history of

their contexts of use.[34] Machine learning programs are not neutral; research agendas and the data sets they work with often inherit dominant cultural beliefs about the world. These research agendas reflect the incentives and perspectives of those in the privileged position of developing machine learning models, and the data on which they rely. The uncritical acceptance of default assumptions inevitably leads to discriminatory design in algorithmic systems, reproducing ideas which normalize social hierarchies and legitimize violence against marginalized groups.[35]

---

discriminatory police and surveillance programs. For example, Woody Bledsoe, the founder of computational facial recognition, was funded by the CIA to purportedly develop identify criminals and criminal behavior: Leon Harmon. 2020. "How LSD, Nuclear Weapons Led to the Development of Facial Recognition". *Observer.* Jan 29. https://observer.com/2020/01/facial-recognition-development-history-woody-bledsoe-cia/, Shaun Raviv. 2020. "The Secret History of Facial Recognition." *Wired.* Jan 21. https://www.wired.com/story/secret-history-facial-recognition/. See also: Inioluwa Deborah Raji and Genevieve Fried, "About Face: A Survey of Facial Recognition Datasets." Accepted to Evaluation Evaluation of AI Systems (Meta-Eval 2020) workshop at AAAI Conference on Artificial Intelligence 2020. Likewise, FERET, the NIST initiative and first large scale face dataset that launched the field of facial recognition in the US was funded by intelligence agencies, for the express purpose of use in identifying criminals in the war on drugs. This objective of criminal identification is core to the history of what motivated the development of the technology. Phillips Jonathon, Harry Wechsler, Jeffery Huang, and Patrick J Rauss. "The FERET Database and Evaluation Procedure for Face-Recognition Algorithms." *Image and Vision Computing* 16, no. 5 (1998): 295–306.

[34] As Safiya Umoja Noble (2018, 30) argues, the problems of data-driven technologies go beyond misrepresentation: "They include decision-making protocols that favor corporate elites and the powerful, and they are implicated in global economic and social inequality." Safiya Umoja Noble, 2018. *Algorithms of Oppression*. New York: New York University Press. D'Ignazio and Klein (2020) similarly argue that data collection environments for social issues such as femicide are often "characterized by extremely asymmetrical power relations, where those with power and privilege are the only ones who can actually collect the data but they have overwhelming incentives to ignore the problem, precisely because addressing it poses a threat to their dominance." Catherine D'Ignazio and Lauren F. Klein. 2020. *Data feminism*. Cambridge: MIT Press. On the long history of algorithms and political decision-making, see: Theodora Dryer. 2019. *Designing Certainty: The Rise of Algorithmic Computing in an Age of Anxiety.* PhD diss. University of California, San Diego. For an ethnographic study that traces the embedding of power relations into algorithmic systems for healthcare-related decisions, see Beth Semel. 2019. *Speech, Signal, Symptom: Machine Listening and the Remaking of Psychiatric Assessment.* PhD diss., Massachusetts Institute of Technology, Cambridge.

[35] As Roberts (2019, 1697) notes in her review of Eubanks (2018), "in the United States today, government digitization targets marginalized groups for tracking and containment in order to exclude them from full democratic participation. The key features of the technological transformation of government decision-making — big data, automation, and prediction — mark a new form of managing populations that reinforces existing social hierarchies. Without attending to the ways the new state technologies implement an unjust social order, proposed reforms that focus on making them more accurate, visible, or widespread will make oppression operate more efficiently and appear more benign." Dorothy Roberts. 2019. "Digitizing the Carceral State." *Harvard Law Review* 132: 1695-1728. <https://harvardlawreview.org/2019/04/digitizing-the-carceral-state/> Virginia Eubanks. 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press. Audrey Beard. 2020. "The Case for Care." *Medium* May 27, 2020. Accessed June 11, 2020. <https://medium.com/@audrey.s.beard/the-case-for-care-c5bcf5b78cbf>. See also: Ruha Benjamin, Troy Duster, Ron Eglash, Nettrice Gaskins, Anthony Ryan Hatch, Andrea Miller, Alondra Nelson, Tamara K. Nopper, Christopher Perreira, Winifred R Poster, et al. 2019. *Captivating Technology: Race, Carceral*

Such research does not require intentional malice or racial prejudice on the part of the researcher.[36] Rather, it is the expected by-product of any field which evaluates the quality of their research almost exclusively on the basis of "predictive performance."[37] In the following sections, we outline the specific ways crime prediction technology reproduces, naturalizes and amplifies discriminatory outcomes, and why exclusively technical criteria are insufficient for evaluating their risks.

I. Data generated by the criminal justice system cannot be used to "identify criminals" or predict criminal behavior. Ever.

In the original press release published by Harrisburg University, researchers claimed to "predict if someone is a criminal based solely on a picture of their face," with "80 percent accuracy and with no racial bias." Let's be clear: there is no way to develop a system that can predict or identify "criminality" that is not racially biased — because the category of "criminality" itself is racially biased.[38]

Techno-science, and Liberatory Imagination in Everyday Life. Durham: Duke University Press. Chelsea Barabas et al. 2020. "Studying up: reorienting the study of algorithmic fairness around issues of power." *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Anna Lauren Hoffmann. 2019. "Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse". *Information, Communication & Society* 22 (7): 900–915. Meredith Broussard. 2018. *Artificial Unintelligence: How Computers Misunderstand the World.* Cambridge, MA: MIT Press. Ben Green. 2019. "'Good' Isn't Good Enough." In *NeurIPS Joint Workshop on AI for Social Good*. AC. Sasha Costanza-Chock. 2020. *Design justice: Community-led practices to build the worlds we need*. Cambridge, MA: MIT Press.

[36] As Ruha Benjamin (2016, 148) argues, "One need not harbor any racial animus to exercise racism in this and so many other contexts: rather, when the default settings have been stipulated simply doing one's job — clocking in, punching out, turning the machine on and off — is enough to ensure the consistency of white domination over time." Ruha Benjamin. 2016. "Catching our breath: critical race STS and the carceral imagination." *Engaging Science, Technology, and Society* 2: 145-156.

[37] By predictive performance we mean strength of correlations found, as measured by e.g. classification accuracy, metric space similarity, true and false positive rates, and derivative metrics like receiver operator characteristic curves. This is discussed by several researchers, most recently Rachel Thomas and David Uminsky. 2020. "The Problem with Metrics is a Fundamental Problem for AI." arXiv preprint arXiv:2002.08512.

[38] Scholars have long argued that crime statistics are partial and biased, and their incompleteness is delineated clearly along power lines. Arrest statistics are best understood as measurements of law enforcement practices. These practices tend to focus on "street crimes" carried out in low income communities of color while neglecting other illegal activities that are carried out in more affluent and white contexts (Tony Platt. 1978. "'Street Crime'— A View From the Left." *Crime and Social Justice* 9: 26-34; Laura Nader. 2003. "Crime as a category—domestic and globalized." In *Crime's Power: Anthropologists and the Ethnography of Crime*, edited by Philip C. Parnell and Stephanie C. Kane, 55–76, London: Palgrave). Consider how loitering is treated compared to more socially harmful practices like wage theft and predatory lending. Similarly, conviction and incarceration data primarily reflect the decision-making habits of relevant actors, such as judges, prosecutors, and probation officers, rather than a defendant's criminal proclivities or guilt. These decision-making habits are inseparable from histories of race and criminality in the United States. As Ralph (2020, xii) writes, with reference to Muhammad (2019), "since the 1600s, and the dawn of American slavery, Black people have been viewed as potential criminal threats to U.S. society. As enslaved people were considered legal property, to run away was, by definition, a criminal act...Unlike other racial, religious, or ethnic groups, whose crime rates were commonly attributed to social conditions and structures, Black people were (and are) considered

Research of this nature — and its accompanying claims to accuracy — rest on the assumption that data regarding criminal arrest and conviction can serve as reliable, neutral indicators of underlying criminal activity. Yet these records are far from neutral. As numerous scholars have demonstrated, historical court and arrest data reflect the policies and practices of the criminal justice system. These data reflect who police choose to arrest, how judges choose to rule, and which people are granted longer or more lenient sentences.[39] Countless studies have shown that people of color are treated more harshly than similarly situated white people at every stage of the legal system, which results in serious distortions in the data.[40] Thus, any software built

---

inherently prone to criminality…Muhammad [thus] argues that equating Blackness and criminality is part of America's cultural DNA." Khalil Gibran Muhammad. 2011. *The Condemnation of Blackness: Race, Crime, and the Making of Modern Urban America*. Cambridge, MA: Harvard University Press; Laurence Ralph. 2020. *The Torture Letters: Reckoning with Police Violence.* Chicago: University of Chicago Press. See also: Victor M. Rios. 2011. *Punished: Policing the Lives of Black and Latino Boys.* New York: NYU Press.

[39] Yet, criminal justice data is rarely used to model the behaviors of these powerful system actors. As Harris (2003) points out, it is far more common for law enforcement agencies to use their records to justify racially discriminatory policies, such as stop and frisk. David A. Harris. 2003. "The reality of racial disparity in criminal justice: The significance of data collection." *Law and Contemporary Problems* 66 (3): 71-98. However, some data science projects have sought to reframe criminal legal data to center such powerful system actors. For example, the Judicial Risk Assessment project repurposes criminal court data to identify judges who are likely to use bail as a means of unlawfully detaining someone pretrial. Chelsea Barabas, Colin Doyle, JB Rubinovitz, and Karthik Dinakar. 2020. "Studying up: reorienting the study of algorithmic fairness around issues of power." In P*roceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (FAccT '20). Association for Computing Machinery, New York, NY, USA, 167–176. Similarly, the White Collar Crime project is a satirical data science project that reveals the glaring absence of financial crimes from predictive policing models, which tend to focus on "street crimes" that occur in low income communities of color. Brian Clifton, Sam Lavigne, and Francis Tseng. 2017. "White Collar Crime Risk Zones." *The New Inquiry* 59: ABOLISH (March). https://whitecollar.thenewinquiry.com/.

[40] Decades of research have shown that, for the same conduct, Black and Latinx people are more likely to be arrested, prosecuted, convicted and sentenced to harsher punishments than their white counterparts, even for crimes that these racial groups engage in at comparable rates. Megan Stevenson and Sandra G. Mayson. 2018. "The Scale of Misdemeanor Justice." B*oston University Law Review* 98 (731): 769-770. For example, Black people are 83% more likely to be arrested for marijuana compared to whites at age 22 and 235% more likely to be arrested at age 27, in spite of similar marijuana usage rates across racial groups. (Ojmarrh Mitchell and Michael S. Caudy. 2013. "Examining Racial Disparities in Drug Arrests." *Justice Quarterly* 2: 288-313.) Similarly, Black drivers are three times as likely as white drivers to be searched during routine traffic stops, even though police officers generally have a lower "hit rate" for contraband when they search drivers of color. "Ending Racial Profiling in America: Hearing Before the Subcomm." 2012. on the Constitution, Civil Rights and Human Rights of the Comm. on the Judiciary, 112th Cong. 8 (statement of David A. Harris). In the educational sector, Nance (2017) found that schools with a student body made up of primarily of people of color were two to eighteen times more likely to use security measures (metal detectors, school and police security guards, locked gates, "random sweeps") than schools with a majority (greater than 80%) white population. Jason P. Nance. 2017. "Student Surveillance, Racial Inequalities, and Implicit Racial Bias." *Emory Law Journal* 66 (4): 765-837). Systematic, racial disparities in the U.S. criminal justice system run historically deep as well. In as early as 1922, white Chicagoans who testified on a report that city officials commissioned following uprisings after the murder of 17-year-old Eugene Williams asserted that "the police are systematically engaging in racial bias when they're targeting Black suspects" (Khalil Gibran Muhammad, quoted in Anna North. 2020. "How racist policing took over American cities, explain by a historian."*Vox*, June 6,

within the existing criminal legal framework will inevitably echo those same prejudices and fundamental inaccuracies when it comes to determining if a person has the "face of a criminal."

These fundamental issues of data validity cannot be solved with better data cleaning or more data collection.[41] Rather, any effort to identify "criminal faces" is an application of machine learning to a problem domain it is not suited to investigate, a domain in which context and causality are essential and also fundamentally misinterpreted. In other problem domains where machine learning has made great progress, such as common object classification or facial verification, there is a "ground truth" that will validate learned models.[42] The causality underlying how different people perceive the content of images is still important, but for many tasks, the ability to demonstrate face validity is sufficient.[43] As Narayanan (2019) notes, "the fundamental reason for progress [in these areas] is that there is no uncertainty or ambiguity in these tasks —

<https://www.vox.com/2020/6/6/21280643/police-brutality-violence-protests-racism-khalil-muhammad> (Accessed June 18, 2020). These same inequities spurred William Patterson, then-president of the Civil Rights Congress, to testify to the United Nations in 1951 that "the killing of 'Negroes' has become police policy in the United States" (https://www.blackpast.org/global-african-history/primary-documents-global-african-history/we-charge-genocide-historic-petition-united-nations-relief-crime-united-states-government-against/). In addition, Benjamin (2018) notes that institutions in the U.S. tend toward the "wiping clean" of white criminal records, as in the case of a Tulsa, Oklahoma officer who had any evidence of her prosecution for the murder of Terrance Crutcher, a 43-year-old unarmed Black man, removed from her record altogether (Ruha Benjamin, 2018. "Black Afterlives Matter." *Boston Review* July 28. Accessed on Jun 1, 2020. <http://bostonreview.net/race/ruha-benjamin-black-afterlives-matter>) . All of these factors combined lead to an overrepresentation of people of color in arrest data.

[41] On the topic of doing "ethical" computing work, Abeba Birhane (2019) avers "the fact that computer science is intersecting with various social, cultural and political spheres means leaving the realm of the 'purely technical' and dealing with human culture, values, meaning, and questions of morality; questions that need more than technical 'solutions', if they can be solved at all." Abeba Birhane. "Integrating (Not 'Adding') Ethics and Critical Thinking into Data Science." *Abeba Birhane* (blog), April 29, 2019. <https://abebabirhane.wordpress.com/2019/04/29/ethics-should-not-be-an-afterthought-in-data-science/>. It is worth mentioning the large body of computer vision, machine learning, and data science research that acknowledges the gross ethical malfeasance of the work typified in the offending research, reveals the impotence of data "debiasing" efforts, and argues for deeper integration of critical and feminist theories in computer science. See, for instance: Michael Skirpan, and Tom Yeh. 2017. "Designing a Moral Compass for the Future of Computer Vision Using Speculative Analysis." In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1368–77. Honolulu, HI, USA: IEEE, 2017. Hila Gonen, and Yoav Goldberg. 2019. "Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But Do Not Remove Them." *ArXiv:1903.03862 [Cs]*, September 24. Green, Ben. 2019. "'Good' Isn't Good Enough." In *NeurIPS Joint Workshop on AI for Social Good*. ACM,. Rashida Richardson et. al.. 2019. "Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice" *NYU Law Review Online* 192, February 13. Available at SSRN: https://ssrn.com/abstract=3333423. Audrey Beard and James W. Malazita. "Greased Objects: How Concept Maintenance Undermines Feminist Pedagogy and Those Who Teach It in Computer Science." To be presented at the EASST/4S Panel on Teaching interdependent agency: Feminist STS approaches to STEM pedagogy, August 2020.

[42] In these applications, both groupings of pixels and human-given labels are directly observable, making such domains suitable for machine learning-based approaches. Criminality detection or prediction, on the other hand, are *not* because criminality has no stable empirical existence. See also: Momin M. Malik 2020. "A Hierarchy of Limitations in Machine Learning." arXiv preprint arXiv:2002.05193.

[43] Yarden Katz. 2017. "Manufacturing an Artificial Intelligence Revolution." SSRN: https://www.ssrn.com/abstract=3078224.

given two images of faces, there's ground truth about whether or not they represent the same person."[44] However, no such pattern exists for facial features and criminality, because having a face that looks a certain way does not *cause* an individual to commit a crime—there simply is no "physical features to criminality" function in nature.[45] Causality is tacitly implied by the language used to describe machine learning systems. An algorithm's so-called "predictions" are often not actually demonstrated or investigated in out-of-sample settings (outside the context of training, validation, and testing on an inherently limited subset of real data),

[44] Arvind Narayanan. 2019. "How to Recognize AI Snake Oil." Arthur Miller Lecture on Technology and Ethics, Massachusetts Institute of Technology, November 18, Cambridge, MA. <https://www.cs.princeton.edu/~arvindn/talks/MIT-STS-AI-snakeoil.pdf>

[45] By insisting that signs of criminality can be located in biological material (in this case, features of the face), this research perpetuates the process of "racialization", defined by Marta Maria Maldonado (2009: 1034) as "the production, reproduction of and contest over racial meanings and the social structures in which such meanings become embedded. Racial meanings involve essentializing on the basis of biology or culture." Race is a highly contingent, unstable construct, the meaning of which shifts and changes over time with no coherent biological correlate**.** To imply that criminality is eminent in biology and that certain kinds of bodies are marked as inherently more criminal than others lays the groundwork for arguing that certain categories of people are more likely to commit crimes *because* of their embodied physicality, a clearly false conclusion. This has motivated leading scholars to move beyond analysis of race *and* technology to race *as* technology. In Wendy Hui Kyong Chun's (2013, 7) words: "Could race be not simply an object of representation and portrayal, of knowledge or truth, but also a technique that one uses, even as one is used by it—a carefully crafted, historically inflected system of tools, mediation, or enframing that
builds history and identity?" See also; Simone Browne. 2010. "Digital Epidermalization: Race, Identity and Biometrics." *Critical Sociology* 36 (1): 131-150; Simone Browne. 2015. *Dark Matters: On the Surveillance of Blackness.* Durham: Duke University Press; Alondra Nelson. 2016. *The Social Life of DNA: Race, Reparations, and Reconciliation After the Genome.* Boston, MA: Beacon Press; Amande M'Charek. 2020. "Tentacular Faces: Race and the Return of the Phenotype in Forensic Identification." *American Anthropologist* doi:10.1111/aman.13385; Keith Wailoo, Alondra Nelson, and Catherine Lee, eds. 2012. New Brunswick: Rutgers University Press; Marta Maria Maldonado. 2009. "It is their nature to do menial labour': the racialization of 'Latino/a workers' by agricultural employers." *Ethnic and Racial Studies,* 32(6): 1017-1036; Wendy Hui Kyong Chun. 2013. "Race and/as Technology, or How to do Things to Race." In *Race after the Internet,* 44-66. Routledge; Beth Coleman. 2009. "Race as technology." *Camera Obscura: Feminism, Culture, and Media Studies,* 24 (70): 177-207. Fields across the natural sciences have long employed the construct of race to define and differentiate among groups and individuals. In 2018, a group of 67 scientists, geneticists, and researchers jointly dissented to the continuation of scientific discourse of race as a way to define differences between humans, and called attention to the inherently political work of classification. As they wrote, "there is a difference between finding genetic differences between individuals and *constructing* genetic differences across groups by making conscious choices about which types of group matter for your purposes. These sorts of groups do not exist 'in nature.' They are made by human choice. This is not to say that such groups have no biological attributes in common. Rather, it is to say that the *meaning and significance* of the groups is produced through social interventions." "How Not To Talk About Race And Genetics." 2018. *BuzzfeedNews* March 30. <https://www.buzzfeednews.com/article/bfopinion/race-genetics-david-reich> (Accessed June 18, 2020).

and so are more accurately characterized as "the strength of correlations, evaluated retrospectively,"[46] where real-world performance is almost always lower than advertised test performance for a variety of reasons.[47]

Because "criminality" operates as a proxy for race due to racially discriminatory practices in law enforcement and criminal justice, research of this nature creates dangerous feedback loops.[48] "Predictions" based on finding correlations between facial features and criminality are accepted as valid, interpreted as the product of intelligent and "objective" technical assessments.[49] In reality, these "predictions" materially conflate the shared, social circumstances of being unjustly overpoliced with criminality. Policing based on such algorithmic recommendations generates more data that is then fed back into the system, reproducing biased results.[50] Ultimately, any predictive algorithms that are based on these widespread mischaracterizations

---

[46] For further reading on why "strength of correlations, evaluated retrospectively," is a more accurate term for "prediction," see Momin M. Malik. 2020. "A Hierarchy of Limitations in Machine Learning." arXiv preprint arXiv:2002.05193; Daniel Gayo-Avello. 2012. "No, You Cannot Predict Elections with Twitter." *IEEE Internet Computing* November/December 2012. Arvind Narayanan 2019

[47] These reasons for real-world performance being less than test set performance include overfitting to the test set, publication bias, and distribution shift.

[48] Hinton (2016) follows the construction of Black criminality through the policies and biased statistical data that informed the Reagan administration's War on Drugs and the Clinton administration's War on Crime. She tracks how Black criminality, "when considered an objective truth and a statistically irrefutable fact...justified both structural and everyday racism. Taken to its extreme, these ideas sanctioned the lynching of black people in the southern states and the bombing of African American homes and institutions in the urban north before World War II…In the postwar period, social scientists increasingly rejected biological racism but created a new statistical discourse about black criminality that went on to have a far more direct impact on subsequent national policies and, eventually, serve as the intellectual foundation of mass incarceration" (19). Elizabeth Hinton, 2016. *From the War on Poverty to the War on Crime*. Cambridge, MA: Harvard University Press. See also: Charlton D. McIlwain. 2020. *Black Software: The Internet & Racial Justice, from the AfroNet to Black Lives Matter*. Oxford, UK: Oxford University Press. Data-gathering enterprises and research studies that uncritically incorporate criminal justice data into their analysis fuel stereotypes of African-Americans as "dangerous" or "risks to public safety," the history (and violent consequences) of which is reviewed in footnotes 12 and 14. The continued propagation of these stereotypes via academic discourse continues to foment anti-Black violence at the hands of the police. It is within this historically embedded, sociocultural construction of Black criminality and Blackness as inherently threatening that police often find their justification for lethal uses of force. Today, Black Americans are twice as likely as white Americans to be murdered at the hands of police. (Julie Tate, Jennifer Jenkins, and Steven Rich. 2020. "Fatal Force: Police Shootings Database." *Washington Post*, May 13). As of June 9, 2020, the *Mapping Police Violence* project found that 24% of the 1,098 people killed by the police in 2019 were Black, despite the fact that Black people make up only 13% of the population in the U.S. <https://mappingpoliceviolence.org/>.

[49] Wendy Hui Kyong Chun (2020) points to the performativity of predictive ML more broadly: "predictions are correct because they program the future [based on the past]. She offers a way to reimagine their use to work against an unwanted future: "In contrast, consider global climate-change models — they too make predictions. They offer the most probable outcome based on past interactions. The point, however, isn't to accept their predictions as truth but rather to work to make sure their predictions don't come true. The idea is to show us the most likely future so we will create a different future." Wendy Hui Kyong Chun and Jorge Cottemay. 2020. "Reimagining Networks An interview with Wendy Hui Kyong Chun." *The New Inquiry*. <https://thenewinquiry.com/reimagining-networks/>

[50] Barocas et al. 2019. "A 2016 paper analyzed a predictive policing algorithm by PredPol, one of the few to be published in a peer-reviewed journal. By applying it to data derived from Oakland police records, they found that Black people would be targeted for predictive policing of drug crimes at roughly twice the rate of whites, even though the two groups have roughly equal rates of drug use (Lum and Isaac 2016).

of criminal justice data justifies the exclusion and repression of marginalized populations through the construction of "risky" or "deviant" profiles.[51]

II. Technical measures of "fairness" distract from fundamental issues regarding an algorithm's validity.

Studies like the aforementioned reflect a growing crisis of validity in AI and machine learning research that's plagued the field for decades.[52] This crisis stems from the fact that machine learning scholars are rarely trained in the critical methods, frameworks, and language necessary to interrogate the cultural logics and implicit assumptions underlying their models. Nor are there ample incentives to conduct such interrogations, given the industrial incentives that are driving much machine learning research and development.[53] To date,

---

Their simulation showed that this initial bias would be amplified by a feedback loop, with policing increasingly concentrated on targeted areas. This is despite the fact that the PredPol algorithm does not explicitly take demographics into account." (Solon Barocas, Moritz Hardt and Arvind Narayanan. 2019. Fairness and Machine Learning. https://fairmlbook.org/; Kristian Lum, and William Isaac. 2016. "To predict and serve?" *Royal Statistical Society* 13 (5): 14-19).

[51] As Reginald Dwayne Betts (2015, 224) argues, "How does a system that critics, prisoners, and correctional officials all recognize as akin to torture remain intact today? The answer is simple: we justify prison policy based on our characterizations of those confined, not on any normative belief about what confinement in prison should look like." Reginald Dwayne Betts. 2015. "Only Once I Thought About Suicide." *Yale LJF* 125: 222. For more on the construction of deviant profiles as a means of justifying social exclusion, see: Sharon Dolovich. 2011. "Exclusion and control in the carceral state." *Berkeley Journal of Criminal Law* 16: 259. David A. Harris. 2003. "The reality of racial disparity in criminal justice: The significance of data collection." *Law and Contemporary Problems* 66 (3): 71-98. ; Michael J. Lynch. 2000. "The power of oppression: Understanding the history of criminology as a science of oppression." *Critical Criminology* 9: 144-152.

[52] This crisis is nothing new: Weizenbaum noted some of the epistemic biases of AI in 1985 (ben Aaron 1985), and Agre discussed the limits of AI methods in 1997 (Agre 1997). More recently, Elish and boyd directly interrogated practices and heritage of AI. Diana ben-Aaron. 1985. "Weizenbaum Examines Computers and Society." *The Tech*. April 9. Philip E. Agre. 1997. "Toward a Critical Technical Practice: Lessons Learned in Trying to Reform AI." In *Bridging the Great Divide: Social Science, Technical Systems, and Cooperative Work*, edited by Geof Bowker, Les Gasser, Leigh Star, and Bill Turner. Hillsdale, NJ: Erlbaum. M.C Elish and danah boyd. 2018. "Situating Methods in the Magic of Big Data and AI." *Communication Monographs* 85 (1): 57–80.

[53] This is perhaps unsurprising, given the conditions of such interventions, as Audre Lorde (1984) points out: "What does it mean when the tools of a racist patriarchy are used to examine the fruits of that same patriarchy? It means that only the most narrow parameters of change are possible and allowable." Audre Lorde. "The Master's Tools Will Never Dismantle the Master's House." 1984. The specific challenges of "ethical" AI practice (due to a lack of operational infrastructure, poorly-defined and incomplete ethics codes, and no legal or business incentives, among others) have been well documented in the past several years. Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. "Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI." In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 14, 2020. Stark, Luke, and Anna Lauren Hoffmann. 2019. "Data Is the New What? Popular Metaphors & Professional Ethics in Emerging Data Culture." *Journal of Cultural Analytics*. Daniel Greene, Anna Lauren Hoffmann, and Luke Stark. 2019. "Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning," 10. Maui, HI,. Hagendorff, Thilo. 2019. "The Ethics of AI Ethics - An Evaluation of Guidelines." *ArXiv Preprint ArXiv:1903.03425*, 15. Jess Whittlestone, Rune Nyrup, Anna Alexandrova, and Stephen Cave. 2019. "The Role and Limits of Principles in AI Ethics:

many efforts to deal with the ethical stakes of algorithmic systems have centered mathematical definitions of fairness that are grounded in narrow notions of bias and accuracy.[54] These efforts give the appearance of rigor, while distracting from more fundamental epistemic problems.

Designers of algorithmic systems need to embrace a historically grounded, process-driven approach to algorithmic justice, one that explicitly recognizes the active and crucial role that the data scientist (and the institution they're embedded in) plays in constructing meaning from data.[55] Computer scientists can benefit greatly from ongoing methodological debates and insights gleaned from fields such as anthropology, sociology, media and communication studies, and science and technology studies, disciplines in which scholars have been working for decades to develop more robust frameworks for understanding their work as situated practice, embedded in uncountably infinite[56] social and cultural contexts.[57] While many groups have

---

Towards a Focus on Tensions." In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 7,. Anna Jobin, Marcello Ienca, and Effy Vayena. "2019. The Global Landscape of AI Ethics Guidelines." *Nature Machine Intelligence*, September 2.

[54] Worthy of note are other discourses of "ethics" in AI, like transparency, accountability, ethics (with fairness, comprising the FATE framework), and trust. For discussion around fairness and bias, see Chelsea Barabas 2019. Beyond Bias: Reimagining the Terms of 'Ethical AI' in Criminal Law. <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3377921#:~:text=Chelsea%20Barabas,-MIT%20Media%20Lab&text=The%20essay%20argues%20that%20attempts,the%20fairness%20of%20these%20tools>. S.M. West, M. Whittaker, and K. Crawford 2019. "Discriminating Systems: Gender, Race and Power in AI. AI Now Institute". <https://ainowinstitute.org/discriminatingsystems.pdf>. However, many scholars have identified limitations of research and design within the Fairness, Accountability, Transparency and Ethics (FATE) streams of machine learning to over-simplify the "interlocking matrix" of data discrimination and algorithmic bias which are always differentially (and disproportionately) experienced (Costanza-Chock, 2018). Others have argued that the focus on fairness through antidiscrimination discourse from law, policy and cognate fields over-emphasizes a liberal framework of rights, opportunities and material resources (Hoffman, 2019: 908). Approaches which bring to bear the lived experience of those who stand to be most impacted into the design, development, audit, and oversight of such systems are urgently needed across tech ethics streams. As Joy Buolamwini notes, "Our individual encounters with bias embedded into coded systems – a phenomenon I call the 'coded gaze' – are only shadows of persistent problems with inclusion in tech and in machine learning." Joy Buolamwini. 2016. "Unmasking Bias". *Medium.* Dec 14. <https://medium.com/mit-media-lab/the-algorithmic-justice-league-3cc4131c5148>. In order for "tech ethics" to move beyond simply mapping discrimination, it must contend with the power and politics of technological systems and institutions more broadly. Sonja, Solomun. 2021. "Toward an Infrastructural Approach to Algorithmic Power" in Elizabeth Judge, Sonja Solomun and Drew Bush, eds. [Forthcoming]. *Power and Justice: Cities, Citizens and Locational Technology*. UBC Press.

[55] Greene, Hoffman, and Stark 2019. Chelsea Barabas, Colin Doyle, JB Rubinovitz, and Karthik Dinakar. 2020. Studying up: reorienting the study of algorithmic fairness around issues of power. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). Association for Computing Machinery, New York, NY, USA, 167–176. Sasha Costanza-Chock. 2018. Design Justice: towards an intersectional feminist framework for design theory and practice. Proceedings of the Design Research Society (2018).; Madeleine Clare Elish and danah boyd. 2018. Situating methods in the magic of Big Data and AI. Communication monographs 85, 1 (2018), 57–80.; Andrew D Selbst and Solon Barocas. 2018. The intuitive appeal of explainable machines. Fordham L. Rev. 87 (2018), 1085.

[56] We borrow verbiage from set theory here to illustrate the deep complexity of such contexts, and to illustrate the peril of attempting to discretize this space.

[57] In outlining parallels between archival work and data collection efforts for ML, Eun Seo Jo and Timnit Gebru (2020) bring forth a compelling interdisciplinary lens to the ML community, urging "that an interdisciplinary subfield should be formed focused on data gathering, sharing, annotation, ethics

made efforts to translate these insights to the field of computer science, it remains to be seen whether these critical approaches will be widely adopted by the computing community.[58]

Machine learning practitioners must move beyond the dominant epistemology of computer science, in which the most important details of a model are considered those that survive abstraction to "pure" technical problems, relegating social issues to "implementation details."[59] This way of regarding the world biases research outputs towards narrowly technical visions of progress: accuracy, precision and recall or sensitivity and specificity, F-score, Jaccard index, or other performance metric of choice, all applied to an ever-growing set of applications and domains. Machine learning does not have a built-in mechanism for investigating or discussing the social and political merits of its outputs. Nor does it have built-in mechanisms for critically exploring the relationship between the research they conduct and the researchers' own subject positions, group memberships, or the funding sources that make their research possible. In other words, reflexivity is not a part of machine learning's objective function.

If machine learning is to bring about the "social good" touted in grant proposals and press releases, researchers in this space must actively reflect on the power structures (and the attendant oppressions) that make their work possible. This self-critique must be integrated as a core design parameter, not a last-minute patch. The field of machine learning is in dire need of a critical reflexive practice.

---

monitoring, and record-keeping processes." Eun Seo Jo and Timnit Gebru. 2020. "Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning." *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. For other great examples of this kind of interdisciplinary scholarship, see: Chelsea Barabas, Colin Doyle, JB Rubinovitz, and Karthik Dinakar. 2020. Studying up: reorienting the study of algorithmic fairness around issues of power. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAccT '20). Association for Computing Machinery, New York, NY, USA, 167–176.

[58] Several key organizations are leading the charge in forwarding reflexive, critical, justice-focused, and anti-racist computing. Examples include Data For Black Lives, which is committed to "using the datafication of our society to make bold demands for racial justice" and "building the leadership of scientists and activists and empowering them with the skills, tools and empathy to create a new blueprint for the future" (Yeshimabeit Milner. 2020. "For Black people, Minneapolis is a metaphor for our world." *Medium* May 29. Accessed June 4, 2020 <https://medium.com/@YESHICAN/data-for-black-lives-statement-of-solidarity-with-black-minnesotans-74a0ef0fbcb5>). Another example is Our Data Bodies, which is "based in marginalized neighborhoods in Charlotte, North Carolina, Detroit, Michigan, and Los Angeles, California," and tracks "the ways [these] communities' digital information is collected, stored, and shared by government and corporations…[working] with local communities, community organizations, and social support networks, [to] show how different data systems impact re-entry, fair housing, public assistance, and community development" <https://www.odbproject.org/about/who-we-are/>. A third example is the Algorithmic Justice League, which combines "art, research, policy guidance and media advocacy" to build "a cultural movement towards *equitable* and *accountable* AI," which includes examining "how AI systems are developed and to actively prevent the harmful use of AI systems" and "[preventing] prevent AI from being used by those with power to increase their absolute level of control, particularly where it would automate long-standing patterns of injustice" (<https://www.ajlunited.org/>).

[59] Abstraction as epistemology in computer science was independently developed by Malazita and Resetar (2019) and Selbst et al. (2019). James W. Malazita and Korryn Resetar. 2019. "Infrastructures of Abstraction: How Computer Science Education Produces Anti-Political Subjects." *Digital Creativity* 30 (4): 300–312. Andrew D. Selbst, danah boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. "Fairness and Abstraction in Sociotechnical Systems." In *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*, 59–68. Atlanta, GA, USA: ACM Press.

III. Conclusion: Crime-prediction technology reproduces injustices and causes real harm

Recent instances of algorithmic bias across race, class, and gender have revealed a structural propensity of machine learning systems to amplify historic forms of discrimination, and have spawned renewed interest in the ethics of technology and its role in society. There are profound political implications when crime prediction technologies are integrated into real world applications, which go beyond the frame of "tech ethics" as currently defined.[60] At the forefront of this work are questions about power: who will be adversely impacted by the integration of machine learning within existing institutions and processes?[61] [62] How might the publication of this work and its potential uptake legitimize, incentivize, monetize, or otherwise enable discriminatory outcomes and real-world harm?[63] These questions aren't abstract. The authors of the Harrisburg

[60] Sareeta Amrute (2019, 58) argues that the standard, procedural approach of conventional tech ethics "provide[s] little guidance on how to know what problems the technology embodies or how to imagine technologies that organise life otherwise, in part because it fails to address who should be asked when it comes to defining ethical dilemmas" and "sidesteps discussions about how such things as 'worthy and practical knowledge' are evaluated and who gets to make these valuations." In so doing, it risks reinforcing "narrow definitions of who gets to make decisions about technologies and what counts as a technological problem." Alternatively, postcolonial and decolonising feminist theory offer a framework of ethics based on relationality rather than evaluative check-lists, in a way that can "move the discussion of ethics from establishing decontextualized rules to developing practices to train sociotechnical systems--algorithms and their human namkers--to being with the material and embodied situations in which these systems are entangled, which include from the start histories of race, gender, and dehumanisation" (ibid. Sareeta Amrute. 2019. "Of Techno-Ethics and Techno-Affects." *Feminist Review* 123 (1): 56-73). In other words, the conventional frame of "tech ethics" does not always acknowledge that the work of computer science is inherently political. As Ben Green (2019) states, "Whether or not the computer scientists behind [racist computational criminal prediction projects] recognize it, their decisions about what problems to work on, what data to use, and what solutions to propose involve normative stances that affect the distribution of power, status, and rights across society. They are, in other words, engaging in political activity. And although these efforts are intended to promote "social good," that does not guarantee that everyone will consider such projects beneficial." See also: Luke Stark. 2019. "Facial recognition is the plutonium of AI." *XRDS: Crossroads, The ACM Magazine for Students* 25 (3): 50–55. For efforts that exemplify the relational approach to ethics that Amrute endorses and includes the people most marginalized by technological interventions into the design process, see Sasha Costanza-Chock. 2020. *Design Justice: Community-Led Practices to Build the Worlds We Need.* Cambridge, MA: MIT Press. For an example of an alternative ethics based around relationality, see Jason Edward Lewis, Noelani Arista, Archer Pechawis, and Suzanne Kite. 2018. "Making kin with the machines." *Journal of Design and Science.*

[61] Power is here defined as the broader social, economic, and geopolitical conditions of any given technology's development and use. As Ruha Benjamin (2016; 2019), Safiya Umoja Noble (2018), Wendy Hui Kyong Chun (2013; 2019), Taina Bucher (2018), and others have argued, algorithmic power is productive; it maintains and participates in making certain forms of knowledge and identity more visible than others.

[62] Crime prediction technology is not simply a tool–it can never be divorced from the political context of its use. In the U.S., this context includes the striking racial dimension of the country's mass incarceration and criminalization of racial or ethnic minorities. Writing in 2020, acclaimed civil rights lawyer and legal scholar, Michelle Alexander (2020, 29) observes that "the United States imprisons a larger percentage of its black population than South Africa did at the height of apartheid". Michelle Alexander. 2020. *The new Jim Crow: Mass incarceration in the age of colorblindness*. The New Press.

[63] The Carceral Tech Resistance Network (2020) provides a useful set of guiding questions to evaluate new projects, procurements and programs related to law enforcement reform. These questions are

University study make explicit their desire to provide "a significant advantage for law enforcement agencies and other intelligence agencies to prevent crime" as a co-author and former NYPD police officer outlined in the original press release.[64]

At a time when the legitimacy of the carceral state, and policing in particular, is being challenged on fundamental grounds in the United States, there is high demand in law enforcement for research of this nature, research which erases historical violence and manufactures fear through the so-called prediction of criminality. Publishers and funding agencies serve a crucial role in feeding this ravenous maw by providing platforms and incentives for such research. The circulation of this work by a major publisher like Springer would represent a significant step towards the legitimation and application of repeatedly debunked, socially harmful research in the real world.

To reiterate our demands, the review committee must rescind the offer for publication of this specific study, Springer must issue a statement condemning the use of criminal justice statistics to predict criminality and acknowledging their role in incentivising such harmful scholarship in the past. Finally, all publishers must refrain from publishing similar studies in the future.

centered in an abolitionist understanding of the carceral state, which challenges the notion that researchers and private actors for profit can "fix" American policing through technocratic solutions that are largely motivated by profit and not community safety and reparations for historical harms. <https://static1.squarespace.com/static/5d7edafcd15c7f734412daf2/t/5ed57c887eecc1059f813c4b/1591049371228/2020-06-01+%2F+police+industrial+complex.pdf> For a comprehensive record of the risks law enforcement face recognition poses to privacy, civil liberties, and civil rights, see Clare Garvie, A. M. Bedoya, and J. Frankle. 2016. "The perpetual line-up. Unregulated police face recognition in America. Georgetown Law Center on Privacy & Technology." https://www.perpetuallineup.org/

[64] 2020. "HU facial recognition software predicts criminality." *Harrisburg University of Science and Technology*, May 5. https://web.archive.org/web/20200506013352/https://harrisburgu.edu/hu-facial-recognition-software-identifies-potential-criminals/ see also; Rose Janus. 2020. "University Deletes Press Release Promising 'Bias-Free' Criminal Detecting Algorithm." *Vice*, May 6 https://www.vice.com/en_ca/article/7kpzeg/university-deletes-press-release-promising-bias-free-criminal-detecting-algorithm

Chapter 6

Dear Laura: On Moral Debt and Fuzzy Logic in the Courts

Introduction

Enclosed in this chapter is a letter addressed to Laura,[65] a former collaborator of mine who used to be the Director of a Pretrial Services Agency in a mid-sized U.S. state. The letter reflects on a failed experiment to design a "judge dashboard" that used court data to surface trends in the ways that judges set bail. Our hope was that such a tool would enable pretrial workers to track judge responses to pretrial reforms that were designed to reduce pretrial jail populations. We were especially interested in developing a "compliance score" which measured the degree to which judge decisions adhered to statutory guidelines regarding the use of pretrial risk assessment. Our plan was to use such information to identify judges for follow-up conversations regarding their perspectives on risk assessment, and bail reform more broadly.

In those interviews, our aim was to surface key beliefs that shape the way judges make pretrial release decisions, as well as unpack judges' views on the role of algorithms and data in supporting their work. For example, how did risk assessments reinforce or challenge the habits, logics, and worldviews that informed judges' pretrial release decisions? Were judges comfortable thinking about and communicating pretrial risk in probabilistic terms? What were the perceived strengths and weaknesses of algorithmic risk assessment? We hoped that such questions would offer up a more in-depth perspective on the factors which shaped judge bail setting practices generally, and their engagement with statistical risk recommendations specifically.

[65] Laura is a pseudonym. All people and places in this chapter have been pseudonymized in order to preserve the anonymity of the people I write about.

However, after hundreds of hours of field visits, interviews, and data analysis, this project was ultimately shut down due to resistance from judges to having their decisions formally monitored by court administrators. I was initially very disappointed by this outcome. But over time, I've come to appreciate the many lessons that I learned in pursuing this project. These insights have greatly shaped my perspective on the opportunities and limitations of pursuing transformative social change from within carceral institutions.

This letter is my attempt to share those insights with my former collaborator, with whom I had not spoken for three years before writing this letter. The impetus for writing the letter stems from a growing interest that I have in experimenting with more relational forms of ethnographic writing, ones that might enable me to center specific relationships in my analysis and form arguments as part of an ongoing dialogue, rather than as a definitive endpoint to the discussion. I was inspired by Laurence Ralph's (2020) notion of "ethnographic lettering," which he describes as both a mode of ethnographic writing and a research methodology that enables him to participate in "an 'unfinished' and ever-evolving dialogue" with the people and groups involved in the social problems he studies (2020, 194).

In his book, *The Torture Letters*, Ralph structures his chapters as letters addressed to various people whom he encounters during the research process. These letters give him the chance to reframe his research subjects as interlocutors and the primary audience for the knowledge he produces. It also enables him to build in accountability for the claims he makes in his work. As Ralph (2020, 197) describes, "Letter writing forced me out of the typical, discipline-imposed shell of objectivity, making me account for myself as a speaking person, a person who helped fashion and develop a set of ideas that sometimes demanded a reply."

I was drawn to this idea of letter writing as a relational practice, one that invited ongoing conversation and accountability for the arguments that I made. So I decided to experiment with this mode of writing in Chapter 6, by drafting a letter to my main collaborator at the court administration. By frankly reflecting on the limitations of our experiment, I try to uncover some important ways that my thinking has changed as a result of our collaboration. I realize now that it was probably for the best that our project was shut down. Our work was constrained by powerful limitations, and, by accommodating those limitations, Laura and I ceded too much ground. As a result, it was virtually impossible for us to move judges beyond the distorted way of knowing that legal practitioners have used for so long to justify the punishment of thousands of legally innocent people.

In the following letter, I try to cultivate a set of "common notions" about why our project failed, and why similar projects are likely to fail in the future. Specifically, I outline the ways that expanded data collection in the criminal legal system fuels a culture of punishment, creating insurmountable mountains of moral debt for the people caught up in the system. I then outline the ways that judges use "fuzzy logic" to conjure a specter of violence that legitimates, and even mandates, the punishment of legally innocent people. When combined, I argue that data as moral debt and fuzzy logic conjure a boogie man that is hard to dispel from the minds of court officials.

It took me a few weeks to finally muster up the courage to call my collaborator once the letter was complete. I was unsure how she would receive the news that I'd divulged the nitty-gritty details of our collaboration in a public piece of writing. From what I could tell, she was still actively working on issues related to bail reform and I was worried that she might see the letter as a potential liability to her career, even if I'd done my best to anonymize her and her former agency. I was also not sure how she would receive the arguments that I made regarding

why it was inevitable, and even desirable, that our project was shut down. She'd always struck me as someone who genuinely cared about reducing jail populations. But her approach to pursuing decarceration varies significantly from my own. She is a competent and energetic bureaucrat, one who's spent decades working inside the administration of the courts. In contrast, I've opted to work primarily with outside stakeholders to create avenues for external accountability for decision-makers within the legal system. This difference in approach was part of why I wanted to write to Laura in the first place. She is someone whose perspective I value and respect, even though it varies quite a bit from my own.

When I finally did reach out to Laura, she received my request to review and offer feedback on the letter with an open mind and an immense amount of generosity. She spent nearly two hours on the phone with me, reviewing each line of my writing to ensure that it was as accurate as possible. To my relief, she didn't have any major points of contention with the way that I'd portrayed her or her agency. She's even expressed interest in writing a response, though I don't think that we'll have time to include that in this dissertation. In fact, this chapter introduction is the very last piece of writing that I am completing before handing everything over to my committee. I waited for as long as possible to write this introduction, because the letter enclosed in this chapter has sparked a conversation that is still active and ongoing. Moving forward, I plan to sit down with Laura and have another conversation that probes more deeply into what she thinks of the arguments made in the letter. I look forward to hearing what she has to say, as well as exploring the ways we might both expand our capacity to envision and enact freer futures through dialogues with people whose view on the world is different from our own.

Dear Laura,[66]

A few months ago, I looked up the organizational chart for your jurisdiction's Court Administration (CA) and I found a new, unfamiliar name written at the top of Pretrial Services. It took me by surprise — I never thought you'd actually leave the CA. You'd always struck me as a lifer, one of those tenacious public servants who refuses to jump ship, even as the water begins to pool around your knees. But everyone has a breaking point. Maybe you finally reached yours.

Your job was tough, but I can recall only one time that I ever saw you throw up your hands and wonder if maybe it was time to call it quits. There was a small group of us gathered in a conference room at the CA headquarters, brainstorming what could be salvaged from our collaboration. Our project had been on the rocks for months by then, after a judge called your boss to express concern that a couple of researchers from MIT were going around asking judges sensitive questions about their bail setting practices. The interviews were put on pause until the higher-ups could determine how viable such conversations were, given the current political climate.

Since that call, the debate surrounding bail reform had only grown more tense. A judge had recently invited a lobbyist from the American Bail Coalition[67] to come speak at a convening of judges about the importance of cash bail. You had managed to get the invitation revoked, arguing that the government couldn't entertain private interest lobbyists, but it had cost your agency a lot of political capital. Then, at the judge convening, apparently a majority of the circuit judges supported a vote of no confidence regarding your agency's pretrial risk assessment tool. I say apparently because you weren't in the room that day — you got the information secondhand,

[66] Laura is a pseudonym.

[67] The American Bail Coalition is a lobbying group for the for-profit bail bond industry.

after a small contingent of judges testified in front of local lawmakers, arguing that your team was withholding crucial information about defendants on their pretrial reports. These were serious allegations, ones which could have been easily refuted, had someone from your agency, or a more reform-friendly judge, been given the opportunity to testify. But they hadn't.

By the time we were meeting, I could tell you were beginning to fray at the edges. A liaison from the state Supreme Court's office had been sent to sit in on our discussion, and she was clearly getting on your nerves. She had a lot of ideas for how your team could win back the trust of skeptical judges. Like customizing the pretrial report so that each judge had access to the data they considered relevant for making pretrial release decisions. But your team already did that, you explained. In some places, they even used customized scripts, which catered to each judge's information preferences.

So maybe that was the problem, the liaison said, perhaps the level of inconsistency across courtrooms was causing confusion. She suggested that you poll judges on which data points they thought were most important, so that they could collectively decide what information everyone received. But you had done exactly that, you explained, just a few years prior. And your team had systematically tested every single data point that judges considered important, even the obviously biased and bullshit ones, such as whether or not the defendant was Hispanic.
The liaison opened her mouth to speak again, but you cut in.

"You know what, judges just shouldn't use the RAs[68] if they don't want to."

There was an awkward silence. It was hard to believe that someone like you, who'd invested thousands of hours overhauling your jurisdiction's data systems in order to develop an

[68] Risk assessments were frequently referred to as RAs.

up-to-date validated risk instrument, would just shrug their shoulders and say, 'Fine then, don't use it.' Were you for real?

It was as if you could hear the question in the air. We should just focus on restoring proper hearings for danger,[69] you argued, *that* would make judges think a little harder about who they decide to lock up, "Who cares if they are high risk."

But Pretrial's entire program was based on risk assessment, the liaison said. There were nods from others around the table. You ran one of the best-known pretrial agencies in the country, one which was considered a national leader in the implementation of risk assessment.

"Judge Smith's friend out in California is getting really tired of hearing about how great [our jurisdiction]'s risk assessment is," said a member of your team, trying to lighten the mood. But you dug in your heels. People were envious of your operational structure, you argued, not your actual pretrial system. That's when you pulled out your phone, flipping it around so that we could all see the photo on the screen. It portrayed four large stacks of paper, about a foot tall each. "That's how many people are in our jails pretrial," you said, "That's twelve thousand people."

You'd blown up two scanners and a printer producing all that paper, around four thousand five hundred pages in total. Those small mountains of files represented your jurisdiction's in-custody jail population on a single day. You planned to carry them with you to the state legislature in a couple of months, when you were slotted to testify on your jurisdiction's progress regarding bail reform. In the meantime, four members of your staff had been assigned

[69] If a prosecutor wants to detain a person pretrial on account of danger to the community, he must request a hearing, frequently referred to as the "dangerousness hearing." In this hearing, it is the prosecutor's burden to establish by clear and convincing evidence that no conditions of release will reasonably assure the safety of the community.

the monumental task of going through each page manually to check for errors in the data. That was a lot of manpower, especially since your department was understaffed. But you were willing to invest the time and resources in order to make your case.

"I'm always reactionary lately," you explained, "I try to dispel every single myth or reaction that the data is not good. Because that's what they'll say, 'Well I don't trust your data.'… Well, I have four staff members looking up all twelve thousand people, seeing what they're being held for, what holders they've got, what they've been charged with, so we can dispel the myths. And if they don't trust the data then we can have Trish and Bob and Lori and Trinity[70] say, 'Well we did this. You don't trust what we did?'"

Looking back on this now, I can't help but smile. It was such a classic move on your part, to try to persuade legislators of the dire need to decrease jail populations through brute force diligence. If only your data were cleaner, more accurate, then surely it would get the point across. Surely the numbers would speak for themselves.

At the time, I pretty much agreed with this sentiment. Our entire collaboration was based on the premise that we could shift the punitive culture of the courts through thoughtful, data-driven dialogues with judges. But over time, I've come to realize that we were wrong. Back then, I was optimistic about cultivating a shared sense of reality with judges by reflecting on court data. But now I understand the ways that the official record serves as an administrative alibi for harmful social policies. I used to believe that, if only we had the right kinds of data, the right quality assurance processes in place, then we could support more fair and humane decisions in the courts. But now I am wary of the many ways that expanded data collection fuels a culture of punishment, creating insurmountable mountains of moral debt for the people caught up in the

[70] All of these names are pseudonyms.

system. I used to have faith in the power of well-reasoned, good faith debate to shift the worldviews of people in positions of authority. But I no longer underestimate the power of fuzzy logic to fuel the specter of violence in the courts.

I realize now that it was probably for the best that our project was shut down. Our work was constrained by powerful limitations, and, by accommodating those limitations, we ceded too much ground. As a result, it was virtually impossible for us to move judges beyond the distorted way of knowing that legal practitioners have used for so long to justify the punishment of thousands of legally innocent people. These days, I'm more interested in expanding what's possible than I am in settling for what's practical.

We've never gotten the chance to talk about this kind of stuff. That's why I'm reaching out to you now. Because I know your commitment to bringing about a more just world is real. And I know that, in order to seed a better future, it is essential that we reflect on our failed experiments, that we learn from our mistakes.

In this letter, I don't want to mince words. You're someone who's always struck me as a straight talker, which stands out in a place like the Court Administration. The criminal legal system is overrun with euphemism, and the courts are no exception. When I first started studying the courts, I intuitively absorbed the acronyms and legal jargon. But over time I've grown wary of the ways that bland bureaucratic terms like "electronic monitoring" and "supervised conditions of release" dull the hard edges of violent social policies and transform harmful acts into socially sanctioned activities.

Take, for example, the death of Breonna Taylor, a 26-year old emergency medical technician who was killed by Louisville police officers in March 2020. The details of Breonna's death are gruesome — police officers entered her home without knocking, in the middle of the

night, and fired thirty-two shots into her bedroom while she slept. The justification that police offered for this killing was that they were looking for drugs, left over from a previous boyfriend who was already in custody. No drugs were ever found.

Breonna's story strikes a nerve because it lays bare the ways that criminalized communities are vulnerable to state-sanctioned police violence, even when resting in the relative safety of their own homes (Benjamin 2022). Yet, the unsettling facts of Breonna's death are smoothed over in official accounts, which describe the incident as an "officer-involved shooting" at a non-descript "residence" (Vogt 2020). Such language downplays the horror of being mowed down by bullets while resting in the comfort of one's own bed. And it distances the perpetrators of such wanton violence from culpability, transforming Breonna's murder into a perpetrator-less crime (Luu 2020).

There's a growing critical consciousness around the ways that euphemism is used to sanitize the violence of law enforcement (Frazier 2020). Yet, much less attention has been given to the role of such language in the courts. That's partially due to the fact that the criminal courts are widely considered an exception to the headline grabbing violence that is meted out in other spheres of the criminal legal system. The courts occupy an important symbolic role in our society, as the bedrock of justice, impartiality and fairness (Van Cleve 2022). They are a place where people's rights are supposed to be protected by the Constitution, due process, and official records, so that even the most marginalized among us benefit from the even hand of justice.

Yet, I've come to understand the criminal courts as sites of normalized humiliation and punishment, a critical cog in the machinery of state violence that is essential to the criminalization of massive numbers of people. The violence of the courts doesn't tend to make headlines, but it can be found lingering in the background of court records, if only you know

where to look. Take, for example, the judge in your jurisdiction who assigned random drug testing as a condition of release for every single person who stood before him in arraignment court. As I looked through the court records, this pattern of behavior stuck out — mandatory drug testing had ballooned very quickly in your jurisdiction,[71] but there weren't many judges who assigned drug testing to every single person in their courtroom. When we asked you about this pattern of behavior, you told us that this particular judge had a penchant for making people go pee in a cup the moment that they stepped into his courtroom, especially if he felt that they were not exhibiting the appropriate amount of respect for his authority. A defense lawyer had challenged this practice, arguing that the judge could not subject people to drug testing unless it was an official condition of their release. So the judge started to assign random drug testing to all of his defendants.

This story has stuck with me all these years, because it so clearly illustrates the extent to which the structure of the courts enables many judges to maintain a dehumanizing culture of authority. Mandatory drug testing is a deeply humiliating practice, one that leaves its mark on both the people subjected to it and the people enforcing it. I remember you telling me about the PTSD that you had from the time you'd spent watching women submit to mandatory drug tests. Such responsibilities were well below your pay grade by the time I'd met you, but you still avoided the drug testing bathrooms whenever you could, where there were mirrors flanking both sides of the toilet. They gave you flashbacks of women squatting, one hand between their legs and one hand in the air, trying their best to avoid your gaze. Eventually, the job was outsourced

[71] A few years prior, Laura had received a grant to pilot a drug testing program, in the hopes that it might reduce the number of people that judges chose to incarcerate pretrial. Some judges started to assign drug testing quite frequently, but pretrial incarceration numbers did not go down. The Pretrial Agency ultimately decided not to pursue the drug testing program after the pilot was finished, because it was costly, time-consuming and it hadn't significantly impacted incarceration rates. However, judges continued to assign mandatory drug testing, and they required defendants to pay the cost out of pocket.

to a third-party company, in part to save money and in part to preserve staff morale. But those experiences stuck with you, and you were constantly on the lookout for new strategies to curtail the casual use of such practices.

Unfortunately, your experience is not unique. Former court employees in other jurisdictions have recounted similar stories of officials engaging in ritual acts of intimidation and humiliation, such as dressing defendants in garbage bags to go to court or acting out mock executions when a defendant insists on taking their case to trial (Van Cleve 2022). As sociologist Nicole Van Cleve argues (2016, 62), "These practices transmit a powerful public service announcement to defendants and the public: submit to authority, be obedient, remain invisible, languish in silence, and agree that your failings are your fault."

This story also highlights the ways that officials manipulate the official record in order to maintain an informal culture of punishment in the courts, one that lingers just within the bounds of legality. By assigning random drug testing to all his defendants, your judge was able to justify his ongoing abuse of an official condition of release in order to assert his authority and maintain a culture of submission in his courtroom. While that judge was more the exception than the rule when it came to drug testing, there are other such practices which are much more widespread — such as the abuse of money bonds.

By now it's widely acknowledged that money bonds are used as a means of circumventing due process in order to detain people before their trial (Prison Policy Initiative 2016). More than 30 years ago, the Supreme Court affirmed that "liberty is the norm, and detention prior to trial or without trial is the carefully limited exception."[72] But these days the exception has become the rule, and money bonds are a key reason why. Money bonds were

[72] 1987. United States v. Salerno, 481 U.S. 739, 755. (1987).

originally introduced as an incentive mechanism for getting people to show up to court. But these days it's used primarily as a means of detaining legally innocent people without technically breaking the law. Such loopholes increase the opacity of the courts, by enabling judges to detain people without providing any explanation. Technically speaking, they aren't detaining anyone — defendants can get out, if they can pay.

As a result of this administrative loophole, the number of people who are incarcerated while awaiting a resolution to their case has skyrocketed over the last few decades.[73] Like the judge who assigned random drug testing to all his defendants, cash bail has become an administrative alibi for punishing people before their trials. Such practices serve an important purpose — to create an environment where defendants are incentivized to resolve their case as quickly as possible, in order to minimize the costs of ongoing engagement with a life-sucking system. Studies have shown that pretrial detention significantly increases the likelihood of a person accepting a plea bargain (Dobbie, Goldin, and Yang 2018). This is unsurprising, given that ongoing incarceration can jeopardize a person's employment, housing status, and custody of their kids, among other less tangible costs paid out in the currencies of stress, confusion and shame.

The cumulative result of such rule bending is that the vast majority of criminal cases (over ninety percent) never go to trial. When I first started this work, I thought that this was a relatively recent issue. But as I've dug into the history of bail reform, I realized that the daily practices of the courts have been decoupled from the letter of the law for over a century.[74] For

[73] The number of people incarcerated pretrial increased by 433 percent nationwide from 1970 to 2015, almost entirely due to increases in cash bail assignments (Preston and Eisenberg 2022).

[74] As Van Cleve (2022) argues, scholars were raising these issues as far back as the 1920's. For example, legal scholar Roscoe Pound described "the undignified offhand disposition of cases at high speed, [and] the frequent suggestion of something working behind the scenes" in urban criminal courts (Pound 2018). Fifty years later, sociologist Malcolm Feeley described an eerily similar situation, arguing that a recent increase in due process

decades, scholars have argued that this enduring discrepancy is due to the fact that court officials engage in "bureaucratically ordained or controlled 'work crimes,' shortcuts, deviations and outright rule violations," in order to dispose of cases quickly (Blumberg 1967, 22). As someone who's worked in the trenches of the courts for decades, it'll be no surprise to you that, even back then, this kind of rule breaking wasn't a secret. For example, in 1967, observers from the Department of Justice argued that it took very little effort to see the enormous gap between legal theory and the day-to-day practices of the courts (President's Commission on Law Enforcement and Administration of Justice 1967).

Today it's a widely acknowledged fact that judges use administrative records to by-pass rules and regulations that are intended to preserve the presumption of innocence and give every person their fair day in court. For example, when we asked judges to walk us through how they set bail, most of them frankly described the way they used money bonds to detain people pretrial, rather than as a tool for incentivizing people to come to court. Here are a couple of examples:

> **Judge C**: I am sort of against money bonds unless I wanna hold a guy in jail.
>
> **Judge E**: Ultimately, my decision is, do I want you out or do I want you in. If I want you out, I'm going to set a bond either ROR[75] or something I think you can post. If I don't want you out, I'll set a bond you can't post.

As someone who had been studying the issue of bail reform for a while, I knew that the use of money bail for pretrial detention was a widespread practice, but I was surprised by how forthright judges were in talking about it in this way. I'd been thinking of the abuse of money

protections for defendants had done little to change the status quo. As Feeley (1979, 290) argued, "these new-found rights and the opportunities may function largely as hollow symbols of fairness or at best as luxuries or reserves to be called upon only in big, intense, or particularly difficult cases."

[75] ROR is short for "release on own recognizance," which means that a person is released from jail without paying a bond or being subject to other conditions for release.

bail as an open secret, something that everyone knew happened, but nobody talked about, especially in reference to their own work. But the practice has become so normalized, that judges had no qualms in talking about it. I think a big reason for this was because the use of financial bonds as a means of detention was framed as an essential tool for maintaining public safety.

Although the judges we interviewed expressed a wide range of views, minimizing imminent danger to the community was unanimously characterized as the most pressing concern when evaluating someone for pretrial release. Throughout all of our interviews, judges repeatedly invoked fear of releasing someone who might go on to commit a "horrendous," "scary," or "tragic" crime. Here are some examples, taken from two judges who were pro-bail reform, one moderate, and one skeptic.

> **Judge E**: You don't want to drift into being punitive, but you're also trying to protect the community. You don't want to be the judge that releases someone who is very dangerous on my release, goes out and does something horrible to members of the community.
>
> **Judge B**: [Judges] err on the side of caution because we're afraid they're gonna hurt somebody.
>
> **Judge C**: So I think judges are intimidated or reluctant to perhaps take chances. Because all you need to do is let some idiot out of jail and they kill somebody and it's like, who is the judge that let them out of jail?
>
> **Judge D**: I'm all for incarcerating people post-conviction, if that sentence is what is appropriate. I'm not okay with incarcerating people presumed innocent, unless we need to keep the community safe…The only people we should be holding are the really dangerous people.

According to these judges, it's justifiable to use cash bail as a means of detaining someone, if that person poses a danger to the community.[76] The concept of "public

[76] In contrast, failure to appear (FTA) was a secondary concern, because the overarching belief was that those who did not come back to court would eventually be arrested again and brought back before the judge. For example:

safety" is a key organizing principle that judges use to distinguish those who deserve the full protection of the law from those who do not. This line of reasoning is quite common and is key to understanding why mass incarceration has been so challenging to reverse in recent decades. In an effort to pass new legislation, reform advocates frequently accept "carve-outs," which exclude people charged with or convicted of "serious" or "violent" crimes. In fact, every major criminal legal reform from the past two decades has focused exclusively on people convicted of "non-serious," "non-violent" or "non-sexual" offenses (Jones 2020).

Similarly, judges argue that the presumption of innocence and liberty at the pretrial stage cannot be extended to people who pose a risk to public safety. The concept of danger is used to divide people into categories of "deserving" and "undeserving" of full protection under the law. On its face, this line of reasoning seems hard to argue with. As one prominent legal scholar put it, "If defendants are incarcerated, the long-term consequences could be very severe. Their lives could be ruined…But if defendants are released… People might be assaulted, raped, or killed." (Sunstein 2019, 500).

However, there's a massive gap between this rhetoric and the reality of mass pretrial incarceration. Less than five percent of arrests in the United States are for crimes involving physical harm to another person (Vera Institute 2018). In fact, the incidence of arrest for such crimes during the pretrial stage is so low that it is statistically impossible to meaningfully differentiate people according to their level of risk (Dinakar, Barabas,

---

Judge E: If you don't return to court but you're not out there committing crimes against the community, yeah, it's a problem but I'm not overly concerned about it because we're going to get you.

Judge D: FTA, while important to me that a person come back to court, so if they don't. So a warrant goes out and they get arrested....everybody eventually comes back… but the real risk to me is in the danger to the community and to the person themselves.

and Doyle 2019). Given these realities, scholars have argued in unequivocal terms that essentially no one is dangerous enough to justify jail time before they've been convicted of a crime (S. Mayson and Stevenson 2021). Yet, judges in your jurisdiction continue to detain around a third of their pretrial population on any given day (Vera Institute of Justice 2015).

These incarceration rates are fueled by a psychological environment that makes pretrial detention *feel* necessary, even in the face of overwhelming evidence to the contrary. As a result, efforts to directly curtail pretrial detention by making cash bail more affordable have not worked. Take, for example, bail credits, which were designed to create an alternative pathway to release through credit for time served. Eligible defendants in your jurisdiction were presumed to receive one hundred dollars-worth of credit towards their bail for each day they spent in jail. You told me about the various tactics you'd tried to nudge judges into embracing this policy.[77] But in the years following the legislation, bail credits remained something that very few judges actually used, even though the vast majority of defendants were eligible. Judges had no qualms in explaining to me their reasons why. As one judge told me, "If I wanted them to get out, then I would just lower the bond." Judges didn't use bail credits because doing so would have undermined what they saw as *the point* of money bonds — to lock people up.

Such comments underscore the importance of investing in interventions that shift the way that judges perceive pretrial risk, rather than simply attempting to make money bonds more affordable. Only then might we have a decent shot at rolling back the punitive culture of the courts from the inside out. This came up in a conversation that I had with one of your more pro-

[77] For example, you printed out cheat sheets summarizing the eligibility criteria, so that judges wouldn't have to remember them off the top of their head. And you required that your staff write whether or not someone was eligible for bail credit on the front page of their pretrial report. That way it was hard to ignore.

reform judges, who discussed the frustrating knee-jerk reactions of his colleagues in response to efforts to decrease pretrial jail populations.

> **L**: When you say there has to be a culture shift, a shift in what?
>
> **Judge C**: Well attitude and approach and understanding. And you say, 'Well we've done it this way forever, so why should we change?' and if you change you're gonna get all sort of hysteria — "They're gonna be letting dangerous criminals out of jail!" and "Oh my goodness, the sky is falling!... oh my God it's fraught with disaster!"

In theory, it should have been easy to dispel such disaster-driven thinking, especially for someone like you, who was well positioned to wield data to facilitate good-faith dialogues with judges. But I was wrong to think that the numbers would simply speak for themselves, especially in a context where the official record has been weaponized for decades to uphold certain "fundamental fantasies"[78] (Zulaika 2018) regarding the role of incarceration in maintaining public safety.

More than "simple number data"

As part of our research, we drafted a survey that was designed to gauge the extent to which judges thought about pretrial risk in statistical terms, as well as how close their statistical estimates were to the forecasts generated by actuarial risk assessments. Before sending out the survey to the entire judge population, a staffer from the CA reached out to a group of judges who sat on the Pretrial Commission, requesting that they review our questions and offer feedback.

[78] This is a phrase coined by Joseba Zulaika (2018) to discuss the ways that official records are used to manufacture the specter of looming terroristic threats and the complementary fantasy of security. As Zulaika (2018, 216) argues, "In the waiting for terror defined by the imminence of a threat, what *could* happen is actually the case *now* as collective representation and fear . . . its future anticipation, its fantasy, is a critical component of the counterterrorism culture".

The judges were not impressed with our work. One said that he was opposed to giving out the survey because he thought the questions were worded in a way that was designed to expose judges who didn't follow statutory guidelines. He said that he'd taken enough surveys to know that they were never about the collection of "simple number data," warning that our questionnaire might cast judges in a negative light. Others agreed. In an email, one judge talked about the importance of knowing who had funded our research, insinuating that there might be ulterior motives behind the survey. He argued that good scientists would never blindly accept a result given out by an algorithm without scrutinizing the underlying data and ascertaining whether the predictions lined up with the facts. He suggested that we reword our questions to try and understand why judges did not trust the risk scores provided on the pretrial reports.

This commentary highlighted judges' deep skepticism regarding our attempts to collect data on their decision-making practices, as well as a keen awareness of the ways that such data might fuel political struggles regarding bail reform, in the courts and beyond. The judges did not view our survey questions as an effort to collect "simple number data," but rather as an attempt to gather malleable information that could be used to spin up narratives that served specific, if undefined, political interests. Such data had the potential to tarnish judges' reputations or pressure them to follow statutes that, according to members of the Pretrial Commission, judges had good reason to disagree with.

I found this feedback illuminating because it challenged my thinking on the ways that we might use data to nudge judges into making less punitive pretrial decisions. Through our judge interviews and surveys, we'd hoped to surface widely held beliefs and heuristics that could be tested against the CA's internal data. Statements like this:

> **Judge B**: For a very serious offense, and even trafficking charges, which can be very serious, once we have them they're gonna have an incentive to leave. And I am always surprised when they do post bond and they do come back.

Such statements could be reformulated into testable hypotheses that served as a springboard for more in-depth data-driven conversations with judges. For example, is it true that people facing serious charges (i.e. felonies) are more likely to fail to appear in court than someone charged with a misdemeanor? We hoped to use the CA's historical data to answer such targeted questions, and in doing so, subtly challenge the beliefs and rules of thumb that propped up carceral outcomes in the courts.

But our theory of change was naïve. It took for granted that judges would accept our numbers as straightforward and objective facts. In reality, judges had a strong critical consciousness about data — they knew that they "must look behind the mathematical result" to see if the data lined up with their sense of reality. Judges had their own ways of reasoning with data, which didn't always align with statistical notions of pretrial risk.

You took judges' views on risk seriously and, as a result, your team had a backlog of requests to test various judge theories regarding which factors mattered most for evaluating pretrial risk. For example, one judge wanted to know if a prior conviction for cruelty to animals was predictive of future arrest for a violent crime. Other judges were interested in using CA data to debate the effectiveness of specific pretrial reforms, such as the accuracy of risk assessments. For example, one of your hardcore risk assessment skeptics had requested data on the number of people who had been arrested after she'd released them from jail. She wanted this information broken down by risk level so that she could "prove" that the risk assessment was failing to accurately categorize people according to risk level.

"But the thing is, she's not letting the high-risk people out of jail," explained Mike, your lead data analyst. This judge was someone we'd labeled as high-yield, low release, meaning that

she saw a relatively large number of people, but released relatively few of them compared to her peers. “So I had to tell her, ‘Listen, the numbers I’m gonna give you, Judge, a lot of them are low risk but that’s because you kept the high risk people in jail.”

Judge W wanted to use her data in order to demonstrate how the risk assessment had consistently mislabeled the people whom she released, by highlighting the fact that the majority of them were labeled “low risk.” However, this argument doesn’t follow sound statistical reasoning. Because she released so few people in the high and moderate risk categories, people with low risk scores were overrepresented in her pretrial arrest data. The judge was essentially trying to make an argument about the false negative rate of your risk assessment tool. Yet she had not considered the fact that there were also false positives, or people she had incarcerated who would not have been arrested, had she released them from jail. These cases were not visible in the data, because there is no way of recording a counterfactual to incarceration. When I asked Mike if it might be helpful to explain how the data only showed one side of the story, he shrugged.

“The judges will take the data and twist it for their own use,” he explained.

When I asked him to elaborate, he told me about another time when that same judge had asked him to calculate the percentage of people who had been arrested on a violent charge after she’d released them. There were so few people who met that description, Mike explained, that it was too tiny to tabulate. But rather than let the numbers influence her perspective, the judge dug in her heels, highlighting anecdotes about specific cases that supported her argument.

“She has the data and she still doesn’t get it,” he said.

We could interpret these stories as examples of the ways judges use court data to justify their own flawed beliefs, rather than engage in empirically grounded dialogue. But I think there's

more to the story than just that. It was hard for your team to challenge such arguments because they made intuitive sense to judges, who had limited experience with statistical reasoning. Judges had a hard time expressing their pretrial decision-making processes in probabilistic terms. In our interviews, I'd ask them to explain to us what level of risk, in terms of percent likelihood, they thought warranted pretrial detention. Most judges struggled to answer our question in those terms. Here's an example of such an exchange with one of our moderate judges:

> **Chelsea**: What would be your percentage likelihood to detain someone for dangerousness?
>
> **Judge D**: I don't know. I don't have an answer to that. That has so many factors. I wouldn't have the same answer for a person charged with murder as with a person charged with theft. Even if a person is highly likely to go back and steal again, I'm more likely to let them out than someone who is highly likely to murder someone again. I couldn't really say what percentage...it depends on charge and history.
>
> **Chelsea**: So there's not a percentage likelihood? Do you have any indication of how high risk a high-risk person is on an RA in terms of percentage?
>
> **Judge D**: [Laughs] No.

Even for judges who placed a high value on risk assessments as a guide for making release decisions, they were not familiar with the probabilities used to categorize pretrial defendants into different risk groups. They preferred using the risk labels offered on the pretrial reports. For example:

> **L:** What's the purpose of risk assessment instruments?
>
> **Judge C**: To set good bonds, absolutely, all of us rely on them to various extents. But it's very, very important.
>
> **L:** How do they function?

> **Judge C**: Well, and again they give us probabilities, this is the algorithm thing. They give us the likelihood that a person will return to court and is likely to commit a future crime. And again, I don't get into 75% or 82%. I am comfortable with the high, medium and low and working off that.

Such comments highlight how unfamiliar judges were with using probabilities to reason about pretrial release. In fact, all of the judges we spoke with had a hard time assigning specific statistical ranges to the risk labels used in your jurisdiction's actuarial risk instrument, even though they were written at the top of every pretrial report. This disconnect between the tool's risk labels and their underlying probabilities sheds light on some of the friction that Pretrial experienced in getting judges to accept their risk-based recommendations. Risk assessments are designed to help judges to distinguish "signal from noise" when making pretrial release decisions. They are supposed to do this by identifying and appropriately weighing the factors that are most predictive of specific pretrial outcomes. Or put another way, risk assessments are meant to help eliminate factors which are not predictive of pretrial outcomes, and ensure that the factors which are predictive are not blown out of proportion.

Sometimes this process leads to unintuitive results, which can make it hard for judges to accept the tool's risk forecasts, especially when they conflict with deeply entrenched ways of thinking. We hoped that we could use data to take the boogie man out of the closet, challenging widely held myths about pretrial risk one by one, until judges no longer felt the need to lock people up. But we didn't fully appreciate how foundational court data was for sustaining the boogie man in the first place. We underestimated how hard it would be to invert the narrative lens used to interpret court data, away from evaluating the supposed risk of criminalized people and toward understanding the ways data are used to justify ongoing criminalization.

Creating Boogie Men thru Moral Debt and Fuzzy Logic

Not long after your jurisdiction adopted a new risk assessment tool, some local prosecutors started making pithy statements in the local newspaper, talking about how crime victims were not going to appreciate it if local judges started releasing dangerous felons back out onto the street. This kind of fear mongering is quite common in places where bail reform is hotly debated. For the sake of anonymity, I won't quote your local prosecutor here directly, but I can point to a similar example from a different jurisdiction, which captures the same general sentiment. For example, a prosecutor in a different jurisdiction started printing bright yellow bumper stickers that read, "Crime Victims Say: Catch & Release is for Fish, Not Felons" after their jurisdiction passed a new bail reform bill. The phrase gives off an air of down-to-earth folk wisdom, but it's flat-out wrong. In that county, along with many jurisdictions around the country, every person has a right to bail, except for in extreme circumstances like capital murder cases. As someone with a law degree, the prosecutor surely knew this. But his bumper sticker exemplified how legal practitioners justify diverging from the letter of the law, in ways that have widespread appeal.

That bumper sticker is a classic example of criminalization in action. By labeling a pretrial defendant a "felon," (presumably because they were either charged with or had previously been convicted of a felony), the prosecutor weaponized involvement with the criminal legal system in order to justify punishment in the present. This is in spite of the fact that the hypothetical defendant has either already served time for a prior conviction, or has only been charged with, not convicted of, a crime. Then, by referencing an unspecified group of "victims" in the tag line, the prosecutor subtly imbues the label of "felons" with a looming sense of danger. As the bumper sticker insinuates, felons have victims. Therefore, felons are dangerous.

This kind of slippery reasoning is extremely commonplace in the criminal legal system, and it reveals the ways that categories such as "danger" are constructed to include more and more people over time. The category of "violent crime" is neither natural nor self-evident (Gilmore 2015). In fact, state and federal laws apply the term to a wide range of offenses, many of which do not involve physical harm. For example, in some states, larceny is considered a violent crime, which means that offenses as minor as snatching a purse or stealing drugs fall under the rubric of violent offense (Sklansky 2021). In public discourse about crime, people often use terms like "violent" and "serious" crime interchangeably, which opens the floodgates for anyone charged with, say, a felony, to be labeled a danger to the community.

This two-part move — in which officials use prior interactions with the criminal legal system as a type of demerit system and then employ fuzzy logic to link those demerits to a looming sense of danger — is essential to understanding how court data are used to maintain a specter of violence, even in the face of good faith efforts to use data to foster a different understanding of reality.

Over the course of our collaboration, I grew to appreciate how committed you were to using data to engage in earnest debates with judges regarding various bail reforms. For example, back in 2011, you polled all the judges across the jurisdiction, asking them for input on what factors they thought were most relevant for evaluating a person's pretrial risk. Then you systematically tested each of those factors to identify which ones should be included in your jurisdiction's validated risk assessment instrument. Through that process, you were able to narrow down the set of factors which had a significant statistical correlation with the specific pretrial outcomes that judges were supposed to consider. In spite of these efforts, a number of

judges expressed a lack of trust in the data used in risk assessments. Take the below as an example:

> **Chelsea**: Do [risk assessments] make your job easier or harder?
>
> **Judge B**: [Sighs]. Are those my two choices? [Laughs]. I guess I'm gonna go with easier. Because I mean all the information is there, it's just not as user-friendly as it used to be... I can assure you I do not rely upon the words "low" and "low" at the top [of the pretrial report][79] to make my decision. Those words have very little impact on me.
>
> **Chelsea**: How come?
>
> **Judge B**: Because they don't give the same weight to things that I do.
>
> **Chelsea**: What is the difference between how RAs evaluate risk and how you evaluate risk?
>
> **Judge B**: Somebody can have a bail jumping conviction, they could have an escape conviction, they could have multiple failure to appears, they could be currently on probation or parole, and they're still assessed at a low risk assessment, and that makes me crazy. I can assure you if somebody has a prior escape, bail jumping or failure to appear on a significant charge, they are not getting out.
>
> **Chelsea**: So what level of risk for FTA or dangerousness merits detention?
>
> **Judge B**: All. [Pause] High. Low risk on FTA, if they have a prior bail jumping or escape or an FTA on another significant charge. Risk should be high level of risk for reappearing or reoffending, not low.

These comments came from a judge who was skeptical of risk assessment, but their views on the importance of escape charges appeared to be widely shared across the judge population. For example, on multiple occasions, judges and court officials recounted a story of a case in Furrow[80] County where a person had been brought into court with a prior escape charge, only to escape again by jumping through a courthouse window. As the story goes, the man leapt onto the courthouse roof and managed to cross the street before the police were able to arrest

[79] Risk assessment scores were written at top of every pretrial report that judges received. The scores for new criminal activity and failure to appear were communicated according to their risk label (i.e. low, moderate, or high).
[80] This is a pseudonym.

him. When Pretrial administered his risk assessment, he came back as low risk for Failure to Appear. I don't know if this story is true, but it's the kind of tale that was used repeatedly to illustrate how ridiculous it was that prior escapes were not factored into FTA scores.[81] Indeed, even judges who valued risk assessment recommendations shared this sentiment, such as the below:

> **Chelsea**: How frequently do you agree with the risk score? When you don't agree what's usually the reason?
>
> **Judge D**: Well I accept the risk scores and I accept them, again, because they are validated. Now I will reject them on cases like escape. So like, they are a low risk for FTA and low risk to reoffend, well, that doesn't add up if they were in a community treatment center or on work release and just walked out. So I will disregard them sometimes on the crimes, but generally I accept them as a validated document.

On their face, such statements make a lot of sense. It seems intuitively correct that anyone who has tried to abscond from law enforcement or a detention facility would pose a higher risk for not showing up to court. However, as our team learned more about the evolution of this data point, we began to realize that the issue was much more complicated than it appeared at first glance. As it turned out, the category of escape did actually correlate with future failures to appear, from 1995 until around 2010. After that, however, prior escape charges stopped being predictive of any pretrial outcomes. This was largely due to the fact that home incarceration

---

[81] As one moderate judge explained to us, the fact that escapes were not included in risk assessments had become a kind of running joke among judges.

L: Other than the probation, what other things do you think should be included [on risk assessments]?

Judge F: … if you're charged with escape, you can't ignore that and say that they're 90% likely to come back to court. No, they're not coming back to court, they're going to try to escape. That's the big one that all the judges make fun of [Pretrial] about.

(HIC) had become a much more prevalent practice in your jurisdiction. As the jail overcrowding crisis ballooned into unsustainable proportions, more judges began to incarcerate people in their homes, where they were monitored using a GPS tracking device. Whenever the home incarceration device registered someone as stepping outside of their prescribed area of confinement, or stopped working for some reason, then that event was classified as an "escape" in the court data.

Home incarceration devices are notoriously finicky.[82] In fact, false alerts are so prevalent that supervising officers in many jurisdictions often neglect to follow up (West and Karsten 2017). Such technical difficulties make it challenging for people to meet their basic needs. For example, there are multiple reports of people being fired from their jobs after taking breaks from work in order to walk around outside to reconnect their device to the network (Kilgore, Sanders, and Weisburd 2021). In your jurisdiction, many violations occurred when people momentarily stepped outside of their area of confinement to do things like grocery shopping. Such incidents are a classic example of the ways that technical violations are used to criminalize survival, by penalizing people for trying to meet their basic needs. They can hardly be equated with escaping from a jail or courthouse.

As home incarceration events became a larger proportion of the data points classified under "escape," this input became less and less predictive of future pretrial outcomes. As a result, a researcher who'd been hired to locally validate your tool recommended that it be omitted from the risk assessment. Yet, judges interpreted this as a willful omission of relevant facts and consistently listed it as a reason that they did not trust the scores offered up by the risk assessment tool.

[82] For example, in California, researchers found that the monitors used by the Department of Corrections failed their accuracy test in 46 of 102 tests. (West and Karsten 2017)

As I learned more about these views, I began to see the ways that the data-fication of certain events had become a kind of demerit system, a way of penalizing pretrial defendants by giving them dings on their record that were then used to justify their ongoing punishment further down the line. Prior to the launch of your validated risk assessment, you tried to minimize the prevalence of this practice. When your agency migrated its records over to a new database system, you did a full overhaul of your team's data entry processes. A key aspect of this work was to minimize the tendency for certain categories, like failure to appear and arrest, to become inflated by other events. For example, you implemented quality assurance processes in order to ensure that pretrial workers did not record a "failure to appear" on a proof of compliance event or an "arrest" on a failure to pay warrant. You told me that this kind of quality assurance was essential if you were to use administrative data to evaluate specific pretrial risks.

But it also curtailed the ability for penal officials to use data as a way of assigning what I've come to understand as "moral debt" to incarcerated people, debt that is virtually impossible to repay. At its core, data collection within the criminal legal system is geared towards capturing negative interactions with law enforcement, such as prior arrests, charges and convictions, as well as minor technical violations like failure to pay a fine. There are very few opportunities for people to demonstrate through data the ways that they are improving their lives or contributing positively to society. Take for example, a story that Fred, a member of your team, told me, about a man who had been subjected to an extreme amount of mandatory drug testing while awaiting a resolution to his case. For months, the man was steadfast in his commitment to meeting his conditions of release. He showed up two to three times a week to pee in a cup and every single one of his tests came back negative. Yet, after submitting to hundreds of drug tests and making all of his court appearances, he missed his court date for sentencing.

"We beat him down so much that he finally gave up," Fred explained.

When the man did eventually return to court, the judge rebuked him. Rather than consider all the ways the man had demonstrated his willingness to comply with court conditions, the judge focused primarily on his one misstep. Such examples illustrate the ways that the "data as demerit" paradigm stabilizes the punitive nature of the criminal legal system. When a person fails to jump through even a minor bureaucratic hoop, that failure ends up on their permanent record, which can be dredged up again and again to reinforce a narrative that this person poses a "risk" to the community.

As different spheres of the criminal legal system become more integrated through shared databases, then it becomes even harder to counteract the accumulation of moral debt over time. Data becomes a shared structural resource that creates a contagion of cultural stigma that spreads across jurisdictions and institutions (Van Cleve 2022). It is often impossible to challenge such "derogatory data" (Community Justice Exchange 2022b), which over time is used to justify the exclusion of people from their full legal rights. In this way, criminal legal data becomes not just an archive, but an arsenal (Community Justice Exchange 2022a).

Such practices are key to understanding not only how the courts manufacture moral debt through official recordkeeping, and also how such data drive growing racial disparities in the system. Decades of research have shown that, for the same conduct, Black and Latinx people are more likely to be arrested, prosecuted, convicted and sentenced to harsher punishments than their White counterparts (Mauer 2010; Rehavi and Starr 2014; Demuth and Steffensmeier 2004; The Sentencing Project 2018). People of color are treated more harshly than similarly situated white people at every stage of the criminal legal system, which results in serious racial biases in the data. Unsurprisingly, when judges incorporate this discriminatory data into their decision-making

processes, it produces discriminatory results. These discrepancies not only maintain, but expand existing disparities, by creating self-fulfilling prophecies through data that justify racial profiling and disparate treatment.

Criminal legal data operates as a kind of "negative credential" (Pager 2009) that bestows a risk status upon justice-involved people. In the context of pretrial, this risk status then serves as a means of justifying racially disparate access to liberty and due process according to flawed notions of merit. By translating historical discrimination into individualized measures of moral debt, "criminal history" data serve as a way of justifying the disproportionate exposure of marginalized groups to violent social policies, while maintaining an appearance of impartiality. Such practices have a powerful sanitizing effect, creating a "class of "innocents" (Freeman 1978) who no longer feel responsible for producing discriminatory outcomes, because such outcomes can be explained through technocratic notions of merit.

It was an uphill battle to fight against this demerit system, especially in the context of pretrial, because judges held strong views about what information was relevant for their decisions. Although certain data categories were not conflated with a broader set of events on your validated risk assessment, your agency was obliged to include additional information in your pretrial reports, according to what a given judge thought was important. For example, in some jurisdictions your staff included information about whether or not an individual was on probation or had previously violated a condition of his release. Judges could then use this information however they saw fit, which often involved spinning up their own "sincere fictions" (Van Cleve 2022) about the ways such data implicated a person in risky or dangerous behavior.

But this practice alone doesn't account for the visceral responses that judges expressed when discussing the prospect of reducing pretrial jail populations. Judges need to interpret data

through a lens that significantly raises the perceived stakes of such demerits, transforming minor technical violations into indicators of menacing future behavior. This is achieved through a process which I call "fuzzy logic," where judges develop their own homespun narratives about the relevance of certain data points in order to ratchet up the perceived risks of pretrial release. Your office was in a constant state of dialogue with judges over what should be counted in the system and how. It was one of those things that judges returned to over and over when they discussed reasons why they strayed from the recommendations offered in the pretrial risk assessment.

For example, in our interviews, some judges argued that non-compliance with a condition of release, such as a failure to pay fines, should be recorded as an FTA event, because a failure to pay fines indicated a general lack of respect for the courts and their authority. Similarly, judges grossly over-weighed the importance of prior involvement with the system in their estimates of danger. When I asked one judge to explain how she evaluated the risk of pretrial violence, she said, "If they have a prior serious offense, such as assault, I would give them 75%. I would say the same thing if they're currently on probation or parole. I would give all of those a great deal of weight." These estimates of risk are completely divorced from reality. Nobody has a seventy-five percent chance of failing to appear or being arrested for a violent crime while awaiting trial. In fact, the judge's estimates are more than ten times greater than the prevalence of arrest for a violent crime in the pretrial population.[83] Yet such patterns of thought persist with the help of

[83] For the vast majority of pretrial defendants (93%) the rate of arrest for a violent offense is less than three percent. For those few people remaining in the highest risk category, their rate of arrest for a violent crime is only marginally higher, at 7% (Arnold Foundation 2019).

fuzzy logic, wherein judges manufacture their own explanations for the relevance of certain data points in order to magnify the perceived riskiness of a person facing trial.

As someone who believed in quality assurance processes, you held strong to the conviction that specific data points should be tightly coupled with the outcomes they are supposed to measure. Your risk tool did not conflate minor technical violations, like failure to pay fines, with official pretrial outcomes like failure to appear. However, there was one major exception to this policy — the inclusion of new criminal activity (NCA) scores on your agency's pretrial risk assessment. Your jurisdiction uses one of the most widely used pretrial risk instruments in the country, on which foregrounds the idea of "public safety" in its name and marketing materials. This is unsurprising, given judges' preoccupation with safety, but it also means that they have to perform a feat of statistical gymnastics in order to make good on their branding.

There are three risk scores provided by the risk assessment — risk of failure to appear (FTA), risk of arrest for a violent crime (NVCA),[84] and risk of arrest for any crime (NCA). Whereas FTA and NCA scores are assigned risk labels of low, moderate, and high, the violence score is divided into a binary: flag or no flag for elevated risk of danger. The lowest risk people have about a two percent chance of being arrested for a violent crime and the highest risk people (about seven percent of the pretrial population) have about an eight percent chance of being arrested for a violent crime. That's a distribution of just six percentage points, a difference that is

[84] The NVCA score is customized across jurisdictions, according to what each jurisdiction classifies as a "violent crime." The term "violent criminal activity" is narrowly defined in this jurisdiction to include arrests for things such as assault, murder, rape, etc.

not large enough to make statistically meaningful distinctions amongst defendants.[85] In fact, incidence of arrest for a violent crime is so rare in your jurisdiction that it was actually not possible to validate the NVCA score in your localized assessment. But rather than abandon the quest to measure pretrial violence altogether, the developers of the risk assessment still required your agency to list the NVCA score on your pretrial report, as part of your contractual agreement. They also included an additional risk score for "New Criminal Activity" in the design of the tool, which evaluated a person's likelihood of being arrested for any crime, even though arrest for any crime is not a legally permissible reason to detain someone before their trial. Unsurprisingly, judges incorporated NCA scores into their evaluations of public safety risk, which perpetuated a looming sense of danger in the courts:

> **Chelsea:** So, should we distinguish between a violent offense or should we be able to take nonviolent offenses and count that as towards potential danger?
>
> **Judge D: …**When we go through the process okay so this person is a high risk to reoffend – he is a high moderate to reoffend. Well can we get some kind of a sense about what the future crime might be. And if it's a nonviolent crime, judges are more likely to give some kind of a break on a bond decision if we get a sense that the potential future crime is not going to be a violent one.

The above judge considers the likelihood that a person will be arrested for any crime, and then tries to discern whether or not that crime will be violent in nature. This results in a massive expansion in the perceived prevalence of pretrial violence. Consider

[85] In statistics, predictions are made within a range of likelihood, rather than as a single point estimate. For example, a predictive algorithm might confidently estimate a person's risk of arrest as somewhere between a range of five and fifteen percent. Studies have demonstrated that predictive models can only make reliable predictions about a person's risk of violence within very large ranges of likelihood, such as twenty to sixty percent (Hart and Cooke 2013). As a result, virtually everyone's range of likelihood overlaps. When everyone is similar, it becomes impossible to differentiate people with low and high risks of violence.

the following, where we asked a judge to reason through the percent likelihood for danger that they thought merited detention.

> **Chelsea:** What level of risk of failure to appear or danger merits detention in terms of percent likelihood?
>
> **Judge E:** Danger to the community, that percentage would be lower, I don't know how to quantitate that, if I said that, I feel that someone is 40% likely to reoffend that's a high enough number where I'm...that's gonna make me give consideration to how they're going to reoffend. Does that mean going back to the corner where they were selling drugs or does reoffend mean they're going to get a gun or DUI? Those latter things, that percentage, that 40%, could drop to 20 or 30 [percent], where I'm gonna say we need to hold them. If it's nonviolent, 40% to recommit nonviolent offense, that goes either way…

Let's do a quick gut check on how the above numbers line up with historical rates of arrest for violent crime. Absolutely nobody has a twenty to thirty percent chance of being arrested for a violent crime while awaiting trial, whereas a significant number of people have a twenty to thirty percent chance of being arrested for any crime.[86] Studies have shown that the people with the highest risk of arrest for a violent crime are not the same people with the highest risk for being arrested for any crime at all (S. G. Mayson 2017). As Mayson (2017, 562) argues, "If the goal is to avert violence, focusing on the likelihood of any future arrest targets the wrong individuals and fails to target the right ones." In other words, whether or not a person goes on to be arrested for a non-violent offense is just not relevant. But judges have their own ways of reasoning through arrest data, which fuels a looming sense of danger. Take, for example:

[86] Consider the dataset used to build the Public Safety Assessment (PSA), one of the most widely used risk assessments in the country: 92% of the people who were flagged for pretrial violence did not get arrested for a violent crime and 98% of the people who were not flagged did not get arrested for a violent crime. (Arnold Foundation 2019).

> **Chelsea**: Should we be using general re-arrest data for nonviolent offenses to evaluate risk of dangerousness?
>
> **Judge B**: Yes. I just think that generally all prior criminal history should be considered and all current charges. I don't think we should put on blinders one way or the other whether they're serious or not serious...damaging property...criminal mischief, first degree. That's a felony for damaging property. Is it the kind where it's stupid kids that are out spray painting the side of a building. Or is it gangsters out spray painting the side of the building that they are tagging...is it a [domestic violence] kind of situation where they've destroyed a car and are torturing these souls. That all sounds like a property offense.

As this judge argues, even seemingly minor charges, such as a property offense, could be masking the capacity for more serious violence in the future. The judge makes her point by distinguishing the behavior of "stupid kids" from "gangsters," where the latter group is implicitly understood to be more dangerous and threatening. Such language is loaded with racial undertones that enable the judge to evoke a generalized state of anxiety and fear of racialized subjects — the "gangster," the "thug," the "bad hombre" — without ever explicitly bringing race into the conversation.

Given the historical discrepancies in arrest rates across racial groups, this slippage between arrest and danger enables judges to maintain racist cognitive schemas that rationalize the uneven distribution of justice in the courts. This process is exacerbated when this slippage becomes an outright conflation, where protecting the public requires preventing any and all future arrests.

> **Judge B**: Obviously you want to avoid reoffending. Whatever the current charge is, you want to try to protect victims if there are any. Generally, protecting the public from any type of criminal offense.
>
> **Judge C**: For instance, when they present that he is a low/low or he is a low/high… or they may tell you he is a low with a 2 and a high with a 9,[87] so the defendant is a higher risk to reoffend. Now a person who is a higher risk to reoffend is less

[87] On pretrial reports, a person's NCA and FTA scores are printed at the top of the first page.

> likely to get the benefit of being let out of jail because if someone says they are likely to reoffend, it's best that they remain in jail.

In contrast to other battles you were fighting, the conflation of pretrial danger with arrest is widely accepted and normalized in the CA. When speaking to your staff, arrest for any new charge was used interchangeably with danger. These views are broadly reinforced through academic and institutional narratives, which habitually conflate generalized arrest with public safety threat (DeMichele et al. 2018; Arnold Ventures 2017; Sunstein 2019; Bureau of Justice Assistance 2022). In doing so, your staff, and the broader milieu of experts working on bail reform, institutionalized fuzzy logic, transforming shaky leaps in reasoning into common sense wisdom. Like the "Fish Not Felons" bumper sticker, such practices extend the power of the state to criminalize people, not only through laws, but through more symbolic processes that create categories of people deemed unworthy of protection under the law. In this way, racialized criminalization is able to thrive in the full light of day.

I've come to understand criminalization through data as a core competency of the criminal legal system, essential for maintaining its relentless pace of expansion over time. Far from providing an objective view on individual risk, official court records serve as a tool for manufacturing moral debt through a demerit system that often punishes people for simply trying to survive. With the help of fuzzy logic, court officials then use that moral debt to conjure a specter of violence that legitimates, and even mandates, the punishment of legally innocent people. When combined, data as moral debt and fuzzy logic conjure a boogie man that is hard to dispel from the minds of court officials. Such baseless fears then trap targeted subjects in "a kind of temporal penitentiary" (Benjamin 2016) that is used to justify their ongoing exposure to dehumanizing treatment. These

practices are deeply ingrained in court culture. Even in your agency, where there was top-down support for using data in a different way, these old habits had a way of creeping back in.

What now?

Technically speaking, our project was never completely shut down — it was put on an indefinite pause. For a couple of years, I held out hope that perhaps the political climate would cool down enough for us to resume the work. But these days I am no longer interested in pursuing our project as it was originally conceived. Our collaboration gave me a healthy dose of skepticism regarding the role of data in supporting transformative change in the courts. Data is often framed as a pragmatic first step toward criminal justice reform, a place for establishing common ground in order to support more fair outcomes. I was initially drawn to this idea because it seemed practical, but these days I'm more interested in expanding the horizon of what's considered possible. That's because I no longer accept the commonsense wisdom which has so profoundly circumscribed the "solution space" of bail reform for all these years. I wonder if you feel the same?

Whenever I start to speculate about what you're up to now, my mind goes back to that photo of those four stacks of paper in your office. It's clear to me now that, even back then, you were beginning to reckon with the limits of using data to internally transform the culture of the courts. That in-custody report represented a shift in strategy, away from nudging judges towards less punitive decisions, and toward building external support for limiting judges' ability to wield carceral tools like cash bail. Data-based argumentation was still an integral part of your approach, but your audience was different. I think the impulse to change audiences was a good one. By then it was clear that the punitive culture of the courts was so deeply ingrained that it

wasn't going to change without clear outside pressure. But I doubt that your lawmakers were any more receptive to your arguments for decarceration — after all, they often used data in ways that were similar to judges.

For example, in 2021, a state representative requested that the CA and other state agencies provide statistics on the number of people who had been arrested after the governor commuted their sentences during the early months of the COVID-19 pandemic. The resulting report revealed that nearly half of those who'd been released during the pandemic had been arrested again. This sparked sensational headlines across the state, many of which focused on the new arrests which had resulted in felony charges. The representative who'd requested the stats demanded an apology from the governor, arguing that there were people who had been victimized as a result of his irresponsible actions.

This kind of rhetoric is basically the "Fish not Felons" bumper sticker on repeat. In response, the governor pointed out that the numbers provided by the courts were for arrests, not convictions, and that many of the charges were ultimately dropped or pled down to a less serious charge. In fact, only twenty percent of those who had been arrested had actually been convicted and sentenced by a judge. But by then the stats had done their job, to maintain a culture of fear that sustains the demand for incarceration over time.

As we sat together in your office for the last time, I could tell you were trying to pre-empt this kind of argumentation. You'd already sifted through one hundred and forty-eight of the in-custody case files yourself, looking for examples that would be hard for lawmakers to justify. For example, you'd identified a person who had been in jail for over thirty days and was only

charged with a non-violent, class B misdemeanor.[88] There were multiple sighs of exasperation from around the table, as members of your team agreed that it was ridiculous for someone to be held in jail for so long on such a minor charge. Others chimed in with similar stories, of people who'd been locked up for weeks, even though they were facing drug possession charges and didn't have any holders.[89]

I get the impulse to surface these kinds of stories. By focusing on the low-level drug offenses or petty theft cases, it is easier to illustrate the injustices of punitive policies like pretrial incarceration. But by foregrounding these easy examples, we inadvertently reinforce the cultural division between those who deserve the full protection of the law from those who do not. As the political geographer Ruth Wilson Gilmore (2015) explains, "the danger of this approach should be clear: by campaigning for the relatively innocent, advocates reinforce the assumption that others are relatively or absolutely guilty and do not deserve political or policy intervention."

Many well-intentioned advocates embrace this approach in the hopes that it will make it easier to make the case for decarceration. But while such rhetoric might lead to some near-term wins, it cedes too much ground in the long run, because it leaves too many people behind. When I asked you how many of the cases you'd looked at were for non-violent misdemeanors, you said that, of the one hundred and forty-eight cases you'd looked at, there were only seven.

"But the fact that there are any is disturbing," you insisted, "There shouldn't be any in there on a misdemeanor, period. It just shouldn't happen."[90]

---

[88] This wasn't unusual in smaller counties, especially if the person had asked for a jury trial, because those only took place once in a month in some places. If there are too many trials on that day, then your case might be bumped to the following month.

[89] Holders are for things that do not directly involve pretrial, but prevent people from being released from jail, such as probation detainers, immigration holds, or serving a sentence on another charge or civil matter.

[90] Laura has since clarified that she thinks there are some instances in which a person charged with a misdemeanor might warrant detention, if they pose an imminent threat to someone in the community, such as in the case of domestic violence.

This sentiment is echoed in reform debates across the country. For example, in a widely touted "bipartisan" op-ed in the *Los Angeles Times,* Newt Gingrich and B. Wayne Hughes Jr. argued that "California has been overusing incarceration. Prisons are for people we are afraid of, but we have been filling them with many folks we are just mad at" (Gingrich and Hughs 2014). This line of reasoning is catchy, but it doesn't chip away at the heart of the problem of mass incarceration. Over forty percent of our prison and jail populations are locked up for "violent" offenses, which, as discussed earlier, often include crimes that don't involve physical harm (Jones 2020). What we need to be attacking head on is the culture of fear which justifies the incarceration of so many people, not debating over the fine line between those who we are "scared of" versus those who we are "just mad at."

In the case of pretrial reform, that starts by refusing to entertain goals that are actually impossible to achieve — like predicting pretrial violence. We must challenge the notion that pretrial violence is a predictable and manageable risk. It is not, and efforts to forecast individualized measures of pretrial violence only perpetuate such illusions. I know this might sound like an unreasonable proposition. But it's important for us to remember that pretrial danger has not always been a legally valid reason for detaining someone pretrial. It only entered the picture during the first wave of bail reform in the 1970's, after the District of Columbia passed a law which made community safety an equal consideration to future court appearance in bail bond setting (Schnacke, Jones, and Brooker 2010). Fourteen years later, public safety considerations were mainstreamed in the federal courts with the passing of The Bail Reform Act of 1984. But before that, judges were only allowed to consider a person's likelihood of appearing in court. And the country wasn't overrun with violent pretrial crime.

We can no longer allow fear of the hypothetical murderers and rapists serve as a justification for maintaining a system that creates real and significant violence in the lives of millions of people (Kaba and Nagao 2021). Such practices make our communities less safe by contributing to the systemic factors which drive cycles of violence in the first place. Most key risk factors for violence are related to social and community conditions, not individual attributes (Centers for Disease Control and Prevention 2020). As the Square One Project (2019) explains, "Rather than violence being a behavioral tendency among a guilty few who harm the innocent, people convicted of violent crimes have lived in social contexts in which violence is likely." Promoting public safety means investing in neighborhood infrastructure and basic social services, like access to affordable healthcare and quality education. This vision of justice is endorsed by survivors of violence, who when polled, overwhelmingly support preventative interventions rather than incarceration (Jones 2020).

If we are going to take public safety seriously, then we have to eliminate the loopholes that drive mass incarceration, like cash bail. Insiders like you are in a critical position to identify and advocate for the removal of these kinds of administrative alibis for pretrial punishment. A few years ago, this might have seemed like an unachievable goal. But advocates are already pushing the agenda of pretrial reform beyond what was once considered do-able. For example, in 2021, Illinois became the first state in the country to pass a bill that completely eliminated the use of cash bail. The bill was the culmination of several years of community organizing, in which local advocates worked to build a broader movement to support decarceration.

Data also played an important role in achieving this goal, but not in the way one might think. Rather than build tools to forecast pretrial risk, community volunteers observed court proceedings and collected data about judge decisions. They then used that data to fuel public

accountability campaigns that pressured judges to follow the law. This work is based on the assumption that the core challenge of bail reform was not informational, but political in nature.

The process of data collection is a process of rendering some things visible and other things invisible. In the criminal legal system, data regarding the decisions and actions of key officials, such as judges, prosecutors, and police officers, are scant. As a result, key decision-makers continue to operate with impunity, locking up more and more people, in spite of laws and policies that are designed to bring their behavior back into line with the law. As Sharlyn Grace (2021), a community advocate from Chicago, explains, "Policymaking is about *power*, not just textual changes to the law." She describes how the chief judge in Cook County drafted a rule that formalized, at least on paper, a process for judges to follow existing law regarding the proper use of money bail. When organizers realized that this rule was written as a hedge against an ongoing civil rights lawsuit, they decided to launch a court watching initiative to monitor the implementation of the rule. According to Grace (2021), this led to a transformation in Cook County's bail setting practices. As she explains, "Ultimately, the court rule changed judicial behavior *only* because pressure from community organizing, litigation, and the larger Black Lives Matter movement succeeded in creating the political will needed to enforce existing law."

Court watching initiatives like the one in Cook County aim to create new data sets which enable communities to build stronger systems of accountability around decision makers and ensure that reform efforts translate into real change. But they're also extremely labor intensive and challenging. Volunteers' requests for access to public data, such as docket information, are frequently denied without further explanation. At other times, government offices rescind access to data that was initially made available to the public, after that data was used to catalyze social movements. As a result, some researchers and community groups have undertaken labor

intensive data collection initiatives in order to record information that the courts already possess but will not provide.

Court insiders like yourself are in a powerful position to advocate for the removal of such barriers. Rather than expand data collection about the people targeted by the criminal legal system, you could improve access to basic information about how many people are behind bars. You could help shift the way we interpret court records, away from evaluating the supposed risk of criminalized people and toward understanding the ways data are used to justify ongoing criminalization.

None of these propositions will be easy to achieve. For example, even after the passing of the Pretrial Fairness Act in Illinois, cash bail is still an option for judges— it was scheduled to be eliminated on January 1, 2023, but the Illinois Supreme Court issued a stay to consider lawsuits by prosecutors and sheriffs who claim that the law is unconstitutional (Berlatsky 2023; AP News 2023). But the good news is that you don't have to do it alone. There is a growing movement of people who are already reframing the debate with regard to bail reform, renegotiating the terms of engagement in productive ways. Such people are committed to experimenting towards a freer future beyond the mainstream solution space. The pursuit of such goals will inevitably involve frustration and failure, but even our struggles can serve as stepping stones to success, if only we can learn from our mistakes. Transformative change does not hinge on a series of clever political tactics, after all, but a process that can and must transform us from the inside out (Kelley 2002, xii). Our collaboration was an integral part of that journey for me. Our time together helped me to see powerful ideas and institutions in a new light. It helped me think through political possibilities and strategies that, for me, were previously unthinkable. This letter is my humble attempt to share those learnings with you, to keep the conversation going.

Take care,
Chelsea

Works Referenced

AP News. 2023. “Illinois High Court Halts Elimination of Cash Bail.” AP NEWS. January 1, 2023. https://apnews.com/article/politics-illinois-state-government-legal-proceedings-lawsuits-b310c3f68a1abed94fd71a0709d34df2.

Arnold Foundation. 2019. “PSA Results; For Reference When Creating a Release Conditions Matrix.” psapretrial.org.

Arnold Ventures. 2017. “Public Safety Assessment: A Risk Tool That Promotes….” 2017. https://www.arnoldventures.org/stories/public-safety-assessment-risk-tool-promotes-safety-equity-justice.

Benjamin, Ruha. 2016. “Catching Our Breath: Critical Race STS and the Carceral Imagination.” *Engaging Science, Technology, and Society* 2: 145–56.

———. 2022. *Viral Justice: How We Grow the World We Want*. 1st ed. Princeton: Princeton University Press.

Berlatsky, Noah. 2023. “Illinois Organizers Push to End Cash Bail as They Await Court Ruling.” Prism. May 19, 2023. http://prismreports.org/2023/05/19/illinois-fight-to-end-cash-bail/.

Blumberg, Abraham S. 1967. “The Practice of Law as Confidence Game: Organizational Cooptation of a Profession.” *Law & Society Review* 1 (2): 15. https://doi.org/10.2307/3052933.

Bureau of Justice Assistance. 2022. “Justice Counts: Actionable Data to Bolster Public Safety.” Bureau of Justice Assistance.

Centers for Disease Control and Prevention. 2020. “Risk and Protective Factors |Violence Prevention|Injury Center|CDC.” Centers for Disease Control and Prevention. https://www.cdc.gov/violenceprevention/youthviolence/riskprotectivefactors.html.

Community Justice Exchange. 2022a. “Challenging Information-Sharing Environments.” 2022. https://abolishdatacrim.org/en/resources/activity-3-challenging-information-sharing-environments.

———. 2022b. “Overview: The Invisible Machinery of Data Criminalization.” From Data Criminalization to Prison Abolition. 2022. https://abolishdatacrim.org/en/report/overview-invisible-machinery-data-criminalization.

DeMichele, Matthew, Megan Comfort, Shilpi Misra, Kelle Barrick, and Peter Baumgartner. 2018. “The Intuitive-Override Model: Nudging Judges Toward Pretrial Risk Assessment Instruments.” *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3168500.

Demuth, Stephen, and Darrell Steffensmeier. 2004. “The Impact of Gender and Race-Ethnicity in the Pretrial Release Process.” *Social Problems* 51 (2): 222–42. https://doi.org/10.1525/sp.2004.51.2.222.

Dinakar, Karthik, Chelsea Barabas, and Colin Doyle. 2019. “Technical Flaws of Pretrial Risk Assessments Raise Grave Concerns,” 2019. https://www.media.mit.edu/posts/algorithmic-risk-assessment/.

Dobbie, Will, Jacob Goldin, and Crystal S Yang. 2018. “The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges.” *American Economic Review* 108 (2): 201–40.

Feeley, Malcolm M. 1979. *The Process Is the Punishment: Handling Cases in a Lower Criminal Court*. New York: Russell Sage Foundation.

Frazier, Mya. 2020. “Stop Using ‘Officer-Involved Shooting.’” *Columbia Journalism Review*, August 7, 2020. https://www.cjr.org/analysis/officer-involved-shooting.php.

Freeman, Alan David. 1978. “Legitimizing Racial Discrimination through Antidiscrimination Law: A Critical Review of Supreme Court Doctrine.” *Minnesota Law Review* 62: 1049.
Gilmore, Ruth Wilson. 2015. “The Worrying State of the Anti-Prison Movement | Social Justice.” *Social Justice* (blog). February 23, 2015. http://www.socialjusticejournal.org/the-worrying-state-of-the-anti-prison-movement/.
Gingrich, Newt, and B. Wayne Hughs. 2014. “Op-Ed: What California Can Learn from the Red States on Crime and Punishment.” *Los Angeles Times*, September 17, 2014, sec. Opinion. https://www.latimes.com/opinion/op-ed/la-oe-0917-gingrich-prop--47-criminal-justice-20140917-story.html.
Grace, Sharlyn. 2021. “‘Organizers Change What’s Possible’ | Sharlyn Grace.” *Inquest*, September 23, 2021. https://inquest.org/organizers-change-whats-possible/.
Hart, Stephen D, and David J Cooke. 2013. “Another Look at the (Im-) Precision of Individual Risk Estimates Made Using Actuarial Risk Assessment Instruments.” *Behavioral Sciences & the Law* 31 (1): 81–102.
Jones, Alexi. 2020. “Reforms Without Results: Why States Should Stop Excluding Violent Offenses from Criminal Justice Reforms | Prison Policy Initiative.” Prison Policy Initiative. https://www.prisonpolicy.org/reports/violence.html.
Kaba, Mariame, and Eva Nagao. 2021. “What About The Rapists.” https://www.interruptingcriminalization.com/what-about-the-rapists.
Kelley, Robin D. G. 2002. *Freedom Dreams: The Black Radical Imagination*. Boston: Beacon Press.
Kilgore, James, Emmett Sanders, and Kate Weisburd. 2021. “The Case Against E-Carceration.” Inquest. July 30, 2021. https://inquest.org/the-case-against-e-carceration/.
Luu, Chi. 2020. “The Ethical Life of Euphemisms.” *JSTOR Daily*, September 30, 2020. https://daily.jstor.org/the-ethical-life-of-euphemisms/.
Mauer, Marc. 2010. “Justice for All? Challenging Racial Disparities in the Criminal Justice System.”
Mayson, Sandra G. 2017. “Dangerous Defendants.” *Yale LJ* 127: 490.
Mayson, Sandra, and Megan Stevenson. 2021. “Virtually No One Is Dangerous Enough to Justify Jail.” *The Appeal* (blog). March 15, 2021. https://theappeal.org/virtually-no-one-is-dangerous-enough-to-justify-jail/.
Pager, Devah. 2009. *Marked: Race, Crime, and Finding Work in an Era of Mass Incarceration*. Chicago, Illinois: University of Chicago Press.
Pound, Roscoe. 2018. *Criminal Justice in America*. First issued in hardback, [New edition]. London New York: Routledge.
President’s Commission on Law Enforcement and Administration of Justice. 1967. “Challenge of Crime in a Free Society.” Department of Justice. https://www.ojp.gov/ncjrs/virtual-library/abstracts/challenge-crime-free-society.
Preston, Allie, and Rachael Eisenberg. 2022. “Cash Bail Reform Is Not a Threat to Public Safety.” Center for American Progress. https://www.americanprogress.org/article/cash-bail-reform-is-not-a-threat-to-public-safety/.
Prison Policy Initiative. 2016. “Detaining the Poor: How Money Bail Perpetuates an Endless Cycle of Poverty and Jail Time.” Prison Policy Initiative. https://www.prisonpolicy.org/reports/incomejails.html.
Rehavi, M. Marit, and Sonja B. Starr. 2014. “Racial Disparity in Federal Criminal Sentences.” *Journal of Political Economy* 122 (6): 1320–54. https://doi.org/10.1086/677255.

Schnacke, Timothy R, Michael R Jones, and Claire M B Brooker. 2010. “The History of Bail and Pretrial Release.” Pretrial Justice Institute.

Sklansky, David A. 2021. *A Pattern of Violence: How the Law Classifies Crimes and What It Means for Justice*. Cambridge, Massachusetts: The Belknap Press of Harvard University Press.

Sunstein, Cass R. 2019. “Algorithms, Correcting Biases.” *Social Research: An International Quarterly* 86 (2): 499–511.

The Sentencing Project. 2018. “Report to the United Nations on Racial Disparities in the U.S. Criminal Justice System.” The Sentencing Project. https://www.sentencingproject.org/reports/report-to-the-united-nations-on-racial-disparities-in-the-u-s-criminal-justice-system/.

The Square One Project. 2019. “Executive-Session-Pdf-Reconsidering-the-Violent-Offender-Report-ONLINE_FINAL.Pdf.” Square One Project. https://squareonejustice.org/wp-content/uploads/2019/09/executive-session-pdf-Reconsidering-the-violent-offender-report-ONLINE_FINAL.pdf.

“United States v Salerno.” 1987. https://www.law.cornell.edu/supremecourt/text/481/739.

Van Cleve, Nicole Gonzalez. 2016. *Crook County: Racism and Injustice in America’s Largest Criminal Court*. Stanford, California: Stanford Law Books, an imprint of Stanford University Press.

———. 2022. “Due Process & the Theater of Racial Degradation: The Evolving Notion of Pretrial Punishment in the Criminal Courts.” *Daedalus* 151 (1): 135–52. https://doi.org/10.1162/daed_a_01894.

Vera Institute. 2018. “Arrest Trends.” Vera Institute of Justice. https://communityresourcehub.org/resources/the-vera-institute-of-justice-arrest-trends/.

Vera Institute of Justice. 2015. “Incarceration Trends in Kentucky.” 2015. https://www.vera.org/downloads/pdfdownloads/state-incarceration-trends-kentucky.pdf.

Vogt, Dustin. 2020. “Coroner Identifies Woman Shot and Killed in Officer-Involved Shooting,” May 16, 2020. https://www.wave3.com/2020/03/15/coroner-identifies-woman-shot-killed-officer-involved-shooting/.

West, Darrell M., and Jack Karsten. 2017. “Decades Later, Electronic Monitoring of Offenders Is Still Prone to Failure.” *Brookings* (blog). September 21, 2017. https://www.brookings.edu/blog/techtank/2017/09/21/decades-later-electronic-monitoring-of-offenders-is-still-prone-to-failure/.

Zulaika, Joseba, ed. 2018. “What Do You Want? Evidence and Fantasy in the War on Terror.” In *Bodies as Evidence: Security, Knowledge, and Power*. Global Insecurities. Durham: Duke University Press.

Chapter 7

The World We Deserve

"The world we deserve requires us to be freer than we ever imagined."

— Alexis Pauline Gumbs, 2013

What does it mean to imagine the world more rigorously?[91]

This is not the question that I started out with when I began my PhD nearly four years ago, but it's the question that guides much of my work today. That's because I've grown to understand the ways that technology design and implementation are deeply entangled with struggles for justice, through the maintenance of sociotechnical imaginaries that imbue our unjust social order with coherence and legitimacy. Sociotechnical imaginaries are "collectively held, institutionally stabilized, and publicly performed visions of desirable futures," that are enabled through techno-scientific innovation (Jasanoff and Kim 2015, 19). Such imaginaries are a key terrain of struggle for social change, particularly in times when the status quo is in flux. They are also an important site of inquiry for computational practitioners who care about developing an ethical approach to their work. By attending to the ways that data practices and sociotechnical imaginaries are co-produced (Jasanoff and Kim 2015), we can get a clearer sense

[91] This question is inspired by the work of Alexis Pauline Gumbs, who has written extensively on the role of the radical imagination in igniting abolitionist struggles for justice. In an essay reflecting on the legacy of Harriet Tubman and the Combahee River Uprising, Gumbs asks, "What [does] our prison abolitionist movement need to turn its back on and use as kindling in order to earn the world that we deserve, which is beyond what all of us who have lived our lives in a police state can even imagine… Do we need to give up our belief that the state can evolve when it was founded on forced labor and captivity? Do we need to give up our need to get individual rewards for our action like Harriet Tubman did when she worked underground under the name Moses for so many years?... Do we need to imagine the future more rigorously?" (Gumbs 2013, 13).

of the mechanisms through which certain ways of ordering the world are rendered natural, or even inevitable. Building from Ruha Benjamin's (2019) concept of "captivating technology," I explore the ways that feel-good narratives about technological innovation can lock us into a "changing same," (Alagraa 2021) by giving the appearance of progress while simultaneously reproducing unjust social dynamics that pervade across the structural, interpersonal and personal dimensions of our lives.

I've come to respect how seductive captivating technologies can be, as well as how their widespread appeal helps to maintain discriminatory and racist systems over time. This dissertation brings together a collection of eclectic experiments in trying to imagine the world more rigorously, through the cultivation of a radical sociotechnical imagination in the context of the U.S. carceral state. At its core, a radical imagination refers to the ability to uncover the root causes of oppression and then devise strategies for overcoming them (Haiven and Khasnabish 2014; Davis 1989). The radical imagination is a catalyzing force — it fuels the work of abolitionist world-building by providing the language needed to make sense of the present, as well as envision and enact freer futures.

This work is anchored in an abolitionist approach to justice through and beyond data-driven technology. Rather than strive to create a framework for ethical data practice that completely prevents complicity with dominant social structures, an abolitionist approach to ethical data practice requires computational practitioners to constantly examine the ways that they reproduce harmful social dynamics in their work, so that they can recognize and nurture processes of transformational social change that are already in the making (bergman and Montgomery 2017). The core preoccupation of this dissertation, then, is to identify practical ways of breaking free from dominant modes of knowing and being through computational work,

in order to "un-invent" carceral technology and increase our capacity to envision and cultivate life-giving alternatives.

Over time, I've come to appreciate the ways that these two processes are deeply intertwined and continuously inform one another. For example, the starting place for my work on "studying up" was to identify a specific underexamined orientation in data-driven research and practice — the tendency to cast the algorithmic gaze on the relatively less powerful. My collaborators and I then used that default setting of computation as the starting point for thinking through an alternative orientation for our engagements in the criminal legal system, one which centered the accountability of state officials, rather than the measurement and management of risk in criminalized populations. By pursuing projects that hinged on this flip of the script, we stumbled upon a method for revealing the often-hidden contours of the political economy underlying data science. As challenges emerged in our pursuit of "studying up," we were able to identify specific ways that the material conditions of data science powerfully shaped how we framed the problems we aim to solve and what questions we chose to answer with data. This experience helped me to understand how the imagination is not only integral to conjuring alternative futures, but also to making sense of the present. As Derrick Bell (1994, 893) explains, "[t]o see things as they really are, you must imagine them for what they might be."

Such experiences have pushed me to think about not only the importance of forging liberatory relationships in the present, but also across time. Throughout the chapters in this dissertation, I reflect on what it means to pursue a relationship-based approach to ethics in computation, one that de-centers the *data* in "data ethics" in favor of experimenting with new ways of building our collective capacity to learn and build new worlds together. When I first started writing this dissertation, I thought about this approach primarily in terms of fostering

more equitable collaborations in the present. For example, in Chapter 4, I introduce refusal as a framework that computational practitioners can use to intentionally "center the margins" in their work, by identifying practical ways of participating in and supporting social transformation rather than control it. In Chapter 5, I explore the value and limitations of open letters as a strategy for forging solidarity within and across diverse groups. I also reflect on the cyclical nature of un-learning toxic habits which so frequently undermine collective struggles for justice. Chapter 6 is an experiment in trying to cultivate a set of "common notions" with a former collaborator. By frankly reflecting on the limitations of a failed experiment that we worked on together, I try to uncover some ways that my thinking has changed as a result of our collaboration.

All of these chapters represent attempts to think through and experiment with forming more joyful collectivities in my work. Joy is the growth of people's capacity to perceive the world more expansively and live in ways that were previously unimaginable (bergman and Montgomery 2017). Cultivating joy is a collective pursuit — it's about being in relationship with people who push us to perceive the world more expansively and hold us accountable to living in ways that manifest the latent potential for justice in our everyday lives (Gumbs 2014). As such, I initially thought about the work of forging joyful relationships as something that we do in the present. But over time, my notion of a relationship-based approach to data ethics has grown more expansive. I've realized that it's not just about forging solidarities in the present, but also about cultivating a sense of kinship with those who were engaged in struggles for justice in the past and those who will pick up the mantle long after we're gone. This is a key aspect of more rigorously imagining our way into a more just world (Gumbs 2013).

By drawing from prior histories of struggle, we can form counter-memories that contextualize and re-interpret present day debates about "ethical AI" within a more nuanced understanding of the ways data has been used to entrench social hierarchies over time. This strategy significantly informed the writing presented in Chapters 3 and 4, where I revisit two powerful analytics from other disciplines, "studying up" and "refusal," in order to make sense of contemporary challenges in ethical data practice. Rather than coin a new academic term or develop a novel framework for ethical data practice, the goal of these chapters is to reanimate, concepts that have been around for a while, by putting them into conversation with the present. In doing so, I cast contemporary debates on ethical data practice in a new light, by situating them in larger conversations regarding the role of authoritative knowledge production in maintaining oppressive social structures.

Moreover, the open letters shared in Chapter 5 represent attempts to change the terms of debate regarding two contemporary data tools, by being attentive to the ways that those debates are situated in a much longer history of criminalizing technology. These experiments helped me to appreciate how the act of remembering can serve as a powerful method of illuminating and transforming the present (hooks 1992, 174), just as the act of imagining alternative futures can be used to attune ourselves to the often-invisible forces which shape our present reality. As bell hooks (1992, 6) explains, "we are always in the process of both remembering the past even as we create new ways to imagine and make the future."

Moving forward, I am interested in exploring more opportunities for putting contemporary debates regarding data ethics in conversation with critical concepts from the past, similar to the way I engaged with the analytics of "studying up" and "refusal" in Chapters 3 and 4. I am particularly interested in drawing from disciplines which might help us to reframe data

science as a form of "memory work" (Hughes-Watkins 2018), by interrogating the ways that history is made and re-made through data. One fruitful entry point into this line of inquiry might be critical archival studies. There's a growing body of scholarship which draws from this field in order to pose interesting questions regarding the ways that contemporary data regimes are used to construct competing narratives about the past, present, and future (Agostinho et al. 2019; Dixon-Román 2017; Sutherland 2019). Scholars of critical archival studies insist that meaning and interpretation of data is always bound up with relationships of inequality, and specifically attend to the ways that asymmetrical power dynamics shape the interaction between narrative construction and "hard facts." Critical archivists challenge the notion that history is fixed or independent from the present by interrogating the conditions of historical production that enable certain narratives to emerge while others are silenced (Caswell, Punzalan, and Sangwand 2017). Such an approach rejects the idea that empirical data provide an objective view on reality, and focuses instead on the uneven contribution of different groups who have unequal access to the means of creating empirically grounded narratives of history (Trouillot 2011).

Critical archival scholars have already begun to outline the various ways that the demand for punitive policies is continuously reinforced by the state's ability to appropriate emerging data cultures into "the carceral archive" (Sutherland 2019). For example, Sutherland (2019) outlines the ways that state records are used to manufacture criminalizing narratives that legitimize the state's power to punish racialized groups. In doing so, she challenges the naturalness of categories such as "safety" and "crime," by shedding light on the ways archival practices are highly contingent, political acts. The interpretive lens brought to archival materials defines what questions can be asked of the data and shapes how the records are used and by whom (Sutherland 2019, 6). Such interventions help to challenge the cultural dominance of hegemonic

data narratives and could offer a starting point for data scientists looking to engage differently with data produced by the carceral state. As M'charek (2013, 436) argues, "the factness of facts depends on their ability to disconnect themselves from the practices that helped produce them."

Other scholars have explored the ways that subaltern histories are suppressed through the production and interpretation of archival data. As Trouillot (2011, 28) argues, "Silences are produced not so much through an absence of facts or interpretations, as through conflicting appropriations." Such statements challenge the notion that we can get a more accurate view on the world by simply gathering more data, and shifts attention toward thinking about the ways prevailing narratives might be challenged by creating alternative interpretations of existing data. We can learn a lot by looking at the ways critical archivists have asserted alternative histories, as well as the struggles they have faced in getting those narratives "to gain their right to existence" (Trouillot 2011, 49). For example, Trouillot (2011, 26) traces four key moments in which silences occur during the production of historical narratives:

> "the moment of fact creation (the making of sources); the moment of fact assembly (the making of archives); the moment of fact retrieval (the making of narratives); and the moment of retrospective significance (the making of history in the last instance)."

We might re-read these moments as: the making of data, the making of data sets, the making of data analyses, and the making of "data science" which enables certain kinds of claims to be made and others not. When thinking strategically about how to posit counter-narratives to prevailing data discourses, these "moments of silencing" might be viewed as potential sites of intervention, where we challenge the political economy surrounding data at specific stages of the process in order to construct more liberatory narratives. Such an approach is exciting because it frames data not as an inherently negative force, but as sites of negotiation and therefore possibility, where competing narratives are always struggling for influence (Hatfield 2020).

Future research on ethical data practice could build from this work, by attending to the ways certain narratives are enabled and others are silenced through the process of constructing and interpreting contemporary data archives.

By exploring the ways that data are used to simultaneously silence and amplify competing narratives, we might also develop more expansive ways of thinking about data ethics in relational terms. Throughout this dissertation, I've highlighted the various ways that data production and analysis perpetuate complementary structures of privilege and harm. For example, in Chapter 1, I discuss the ways that the PATTERN risk assessment was used to both re-cast vulnerable incarcerated populations as dangerous and therefore disposable, while also projecting an image of penal institutions as sites of compassion and care. These twin processes of demonization and sanitization are key to understanding the enduring presence of the carceral state over time. By emphasizing the inextricable link between the "winners" and the "losers" of data intensive systems, we are better positioned to disrupt pathologizing narratives that disconnect the plights of marginalized populations from structural forms of oppression via statistical discourse. In future work, I'd like to draw from critical archival studies in order to flesh out the relationship between processes of silencing and processes of narrative amplification through data archives, in order to shed more light on the relational aspect of enduring inequality.

Moreover, critical archival studies may provide some useful concepts for thinking about the connections between data narratives and broader abolitionist struggles to cultivate radical sociotechnical imaginations. I am particularly interested in drawing from scholarship which connects archival practice to the formation and maintenance of "unthinkable facts." Bourdieu (1980, 14) defines unthinkable facts as "that which one cannot conceive within the range of possible alternatives… because it defies the terms under which the questions were framed." As

Trouillot (2011) explains, certain key moments of liberation throughout history, such as the Haitian Revolution, were rendered unthinkable, because the reality of such events defied prevailing racialized narratives used to justify colonialism and slavery in the Americas. In order to make sense of the Haitian Revolution, the dominant class of planters and colonial officials had to shove the facts back into their proper order of discourse in order to maintain a sense of coherence to their reality. We could draw some fruitful parallels between these discussions and current-day debates regarding captivating technology. In what ways do contemporary data practices render liberatory social formations as "unthinkable" and what strategies from the past might we draw from in order to increase our capacity to not only think but enact alternatives to the status quo?

In addition to drawing from history as a resource for imagining the world more rigorously, we must remain vigilant about ensuring that today's "wins" don't create more obstacles in larger struggles for freedom, by maintaining a sense of responsibility for and accountability to the people who will inherit the world long after we are gone. As Ruth Wilson Gilmore (2015) explains, *"Prepare to win* means *be ready for the morning after*." This requires developing our capacity to foresee the ways that proposed data interventions stabilize structural violence by strengthening dominant systems of meaning and control. Our ability to see the world is socially trained in ways that powerfully shape what is recognizable and what goes unnoticed (Jasanoff and Kim 2015; Tuck and Yang 2014). One of the key functions of dominant social imaginaries is to render certain forms of injustice invisible or justifiable. This is exacerbated by the depoliticized nature of engineering culture, which often trains computational practitioners to overlook or dismiss epistemic questions regarding how data are interpreted and operationalized as "non-technical" concerns, irrelevant to the "real" work of engineering (Cech 2013). This

ideology assumes that political and social contexts can and should be kept separate from the purely technical concerns of engineering (Cech and Sherick 2015). As a result, dominant logics and discourses are easily embedded into these systems, because the interpretation and operationalization of data analysis is considered outside the purview of system engineers.

In spite of these challenges, it is possible for computational practitioners to develop their capacity to foresee the ways that the tools they build will be weaponized. One of the best ways of doing this is by engaging with concrete examples that illustrate the various mechanisms through which computational work upholds social inequality. In this dissertation, I offer up a number of examples that have emerged from my own work. Through these case studies, I describe a diverse set of processes through which harmful social dynamics are reproduced through computation, ranging from the structural dimensions of data science (i.e. data as "moral debt") to the interpersonal (i.e. reinforcing knowledge hierarchies through data) and personal dynamics (i.e. dealing with discomfort) which make it challenging for computational practitioners to participate in collective struggles for justice.

Given the highly adaptive nature of racism, the need for this kind of work is ever-present, and often involves uncovering the same truths in novel ways, under new circumstances. As Audre Lorde (1984, 71) says, "[T]here are no new ideas. There are only new ways of making them felt..." Over the course of this dissertation I have grown to appreciate the essential role of affect in helping us to break from hegemonic ways of knowing in order to let alternative imaginations take root and flourish. This requires interrogating the various ways that "epistemologies of ignorance" (Mills 1997), or structured ways of un-knowing, shield those steeped in dominant cultures and institutions from harsh social realities (Gravlee 2021). In this dissertation, I build from the work of scholars such as Amrute (2019) and Keys and Austin

(2022), who have explored the ways we might hone our affective sensibilities in order to perceive unjust social configurations in sociotechnical systems and attune ourselves to different ways of living. As Amrute (2019) explains, "attention to affect—how subjects and technologies are aligned and realigned, attached and reattached to one another—is a method for practising ethics that critically assesses a situation, imagines different ways of living and builds the structures that make those lives possible."

Such a practice requires that people become "traitors" (Hill Collins 1990, 406) to the various forms of privilege that they inhabit and inherit. This can be a messy and uncomfortable process. But there is a particular need for this kind of work in fields such as computation and data science, whose cultures are deeply shaped by, as well as integral to maintaining dominant systems of meaning (Katz 2020). Rather than try to internalize the messiness of the world in computational models, concepts such as "attunement" from feminist affect theory can enable us to be responsive to changing circumstances and experiment with forming different kinds of relationships through sociotechnical systems (Amrute 2019).

In this dissertation, I've examined the ways that carceral technologies are framed as vehicles for compassion and care, in order to unpack the mechanisms through which dominant ideologies make computational practitioners feel good about sustaining violent systems. Over time, I have developed a healthy skepticism toward feel-good narratives about technological progress, as well as my own desires and ambitions to make the world a better place through data. That doesn't mean that I've completely opted out of doing work that involves data in favor of some purer form of radical praxis — the desire for purity can be as problematic as anything else (bergman and Montgomery 2017). But I've managed to reorient the ways that I process and act

on my desires, in order to increase my ability to perceive everyday injustices hidden in plain sight.

For example, in the early stages of my research, I saw myself as a potential bridge-figure, someone who might facilitate productive dialogues across diverse groups of people with competing agendas. I sought out collaborations with a wide range of social actors, including government insiders and community organizers, in the hopes that I might help to establish common ground and foster mutual understanding across these groups. But over time, I began to question the little voice inside my head that told me that I was the lynchpin to social change in this context. Or that, if I didn't agree to conduct some problematic study, then someone else would come along and do it worse than me. I started to understand the need for me to not only create space for dialogue across diverse groups, but also to cede space so that others could engage on their own terms.

I have also explored the ways that computational practitioners might embrace the productive potential of attuning ourselves to the "ugly feelings" (Ngai 2007) that emerge in the process of unlearning harmful practices. By cultivating a sense of curiosity towards feelings of discomfort, defensiveness, and disappointment, we can get better at transforming seemingly negative feelings into joyful opportunities for growth (Ahmed 2010). For example, in Chapter 3, I write about the ethical grey areas that emerged when my collaborators and I shifted the algorithmic lens upward, to study people and institutions of power. One of the motivations behind this work was a sense of anxiety that I felt as we experimented with repurposing data that we'd acquired from the courts in order to make our speculative judge risk assessment. Although we weren't technically breaking any of the terms of our data sharing agreement, I knew that our collaborators from the state agency probably wouldn't have granted us access to that data if

they'd known we were going to use it in that way. This led me to think about the unarticulated assumptions which shape prevailing frameworks for harm mitigation and ethical engagement with research subjects and collaborators. I realized that my anxiety stemmed from the fact that I didn't have a clear framework for making sense of important questions regarding trust and access to data when pursuing projects that "study up." My anxiety was the catalyst for me to explore these ideas further in the article enclosed in Chapter 3.

Similarly, my writing on refusal started from a nagging sense of disappointment that I felt after turning down opportunities to participate in potentially harmful studies. It felt like I was missing out on chances to engage in important policy conversations. And it wasn't clear to me how declining to participate in such research would make a difference. That's because I didn't have a framework for processing the insights that I was learning from my decisions to not participate. The concept of refusal helped me to make sense of these feelings and distill them into kernels of wisdom that have informed my work moving forward.

I've also learned to use affect as a gut check when thinking through how to put my values into practice. For example, one of the main reasons that I decided to write a letter to a former collaborator in Chapter 6 was because I wanted to experiment with building different forms of accountability into my writing practice. But when it came time to actually reach out to my collaborator to ask for feedback, I was really nervous. I procrastinated for as long as I could before giving her a call. The conversation went better than I'd hoped, and I can see now how my writing and thinking was sharpened by not only her feedback, but also by the knowledge that I was going to ask for her feedback in the first place. This experience has taught me to embrace feelings of nervousness as a good sign — by making myself vulnerable to criticism and rejection, I am able to improve the quality of my work. Rather than try to avoid these feelings in the future,

I will embrace them as a compass to guide my interactions with collaborators and accountability partners.

This dissertation has taught me that history and affect are essential tools that we can draw on in larger processes of "un-inventing" (MacKenzie 1993) harmful algorithmic systems. By un-inventing, I mean reconfiguring the structural, interpersonal, and personal aspects of computation to the point that it no longer makes sense to create harmful algorithmic systems (Keyes and Austin 2022). The experiments enclosed in this dissertation represent various attempts at trying to imagine the world more rigorously, in order to move beyond the allure of captivating technology and enact alternative relationships through and beyond technology. All of these experiments are deeply relational at their core. Chapters 3 and 4 represent experiments at trying to become "attentive descendants and good ancestors" in debates regarding captivating technology (Shevin, Shabazz, and Pfefferkorn 2022), by forging a sense of kinship with the people who were engaged in struggles for justice long before us, as well as a sense of accountability to those who will follow in our footsteps.

Chapters 5 and 6 represent messy and ongoing experiments in trying to figure out how to enact more joyful collectivities in the present. In Chapter 5, I grapple with the lessons learned from pursuing a particular strategy for collective action — open letters — in order to explore the value of consistently reflecting on what's worked and what hasn't in our imperfect experiments at abolitionist worldbuilding. Chapter 6 is very much a living document, one that has sparked ongoing reflection and debate with a former collaborator. I don't know how this experiment will end, but I hope that it results in both of us being able to perceive and align ourselves with the latent potential for justice that exists in our daily lives. In fact, that is the hope seeded throughout

this entire dissertation — that it might, in its own small way, contribute to our ability to imagine our way into the world that we deserve.

Works Referenced

Agostinho, Daniela, Catherine D'Ignazio, Annie Ring, Nanna Bonde Thylstrup, and Kristin Veel. 2019. "Uncertain Archives: Approaching the Unknowns, Errors, and Vulnerabilities of Big Data through Cultural Theories of the Archive." *Surveillance & Society* 17 (3/4): 422–41. https://doi.org/10.24908/ss.v17i3/4.12330.

Ahmed, Sara. 2010. *The Cultural Politics of Emotion*. Reprinted. Edinburgh: Edinburgh Univ. Press.

Alagraa, Bedour. 2021. "What Will Be the Cure?: A Conversation with Sylvia Wynter." *Offshoot*, January. https://offshootjournal.org/what-will-be-the-cure-a-conversation-with-sylvia-wynter/.

Amrute, Sareeta. 2019. "Of Techno-Ethics and Techno-Affects." *Feminist Review* 123 (1): 56–73. https://doi.org/10.1177/0141778919879744.

Bell, Derrick. 1994. "Who's Afraid of Critical Race Theory?" *UNIVERSITY OF ILLINOIS LAW REVIEW* 1994 (4): 893–910.

Benjamin, Ruha. 2016. "Racial Fictions, Biological Facts: Expanding the Sociological Imagination through Speculative Methods." *Catalyst: Feminism, Theory, Technoscience* 2 (2): 1–28. https://doi.org/10.28968/cftt.v2i2.28798.

———, ed. 2019. *Captivating Technology: Race, Carceral Technoscience, and Liberatory Imagination in Everyday Life*. Durham: Duke University Press.

bergman, carla, and Nick Montgomery. 2017. *Joyful Militancy: Building Thriving Resistance in Toxic Times*. AK Press.

Bourdieu, Pierre. 1980. *Le sens pratique*. Paris: Editions de Minuit : Maison des sciences de l'homme.

Caswell, Michelle, Ricardo Punzalan, and T-Kay Sangwand. 2017. "Critical Archival Studies: An Introduction." *Journal of Critical Library and Information Studies* 1 (2). https://doi.org/10.24242/jclis.v1i2.50.

Cech, Erin A. 2013. "The (Mis)Framing of Social Justice: Why Ideologies of Depoliticization and Meritocracy Hinder Engineers' Ability to Think About Social Injustices." In *Engineering Education for Social Justice*, edited by Juan Lucena, 10:67–84. Philosophy of Engineering and Technology. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-6350-0_4.

Cech, Erin A., and Heidi M. Sherick. 2015. "Depoliticization and the Structure on Engineering Education." In *The Depoliticization Of*, 1st ed. 2015, 203–16. Philosophy of Engineering and Technology 20. Cham: Springer International Publishing : Imprint: Springer. https://doi.org/10.1007/978-3-319-16169-3.

Davis, Angela Y. 1989. *Women, Culture & Politics*. 1st ed. New York: Random House.

Dixon-Román, Ezekiel. 2017. "Toward a Hauntology on Data: On the Sociopolitical Forces of Data Assemblages." *Research in Education* 98 (1): 44–58. https://doi.org/10.1177/0034523717723387.

Gilmore, Ruth Wilson. 2015. "The Worrying State of the Anti-Prison Movement | Social Justice." *Social Justice* (blog). February 23, 2015. http://www.socialjusticejournal.org/the-worrying-state-of-the-anti-prison-movement/.

Gravlee, Clarence. 2021. "How Whiteness Works: JAMA and the Refusals of White Supremacy." *Somatosphere* (blog). March 27, 2021. http://somatosphere.net/2021/how-whiteness-works.html/.

Gumbs, Alexis Pauline. 2013. “Sometimes Freedom Means You Have To Burn It Down: Harriet Tubman and Abolitionist Vision That Don’t Quite.” *The Abolitionist*, 2013, Fall edition.
———. 2014. “Prophecy in the Present Tense.” *Meridians* 12 (2): 142–52. https://doi.org/10.2979/meridians.12.2.142.
Haiven, Max, and Alex Khasnabish. 2014. *The Radical Imagination: Social Movement Research in the Age of Austerity*. Halifax ; Winnipeg : London: Fernwood Publishing ; Zed Books.
Hatfield, Joe Edward. 2020. “Book Review: How Data Haunt.” *New Media & Society* 22 (1): 177–83. https://doi.org/10.1177/1461444819880585.
Hill Collins, Patricia. 1990. *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment*. Perspectives on Gender, v. 2. Boston: Unwin Hyman.
hooks, bell. 1992. *Black Looks: Race and Representation*. Boston, MA: South End Press.
Hughes-Watkins, Lae’l. 2018. “Moving Toward a Reparative Archive: A Roadmap for a Holistic Approach to Disrupting Homogenous Histories in Academic Repositories and Creating Inclusive Spaces for Marginalized Voices.” *Journal of Contemporary Archival Studies* 5 (6).
Jasanoff, Sheila, and Sang-Hyun Kim. 2015. *Dreamscapes of Modernity: Sociotechnical Imaginaries and the Fabrication of Power*. University of Chicago Press.
Katz, Yarden. 2020. *Artificial Whiteness: Politics and Ideology in Artificial Intelligence*. Columbia University Press.
Keyes, Os, and Jeanie Austin. 2022. “Feeling Fixes: Mess and Emotion in Algorithmic Audits.” *Big Data & Society* 9 (2): 205395172211137. https://doi.org/10.1177/20539517221113772.
Lorde, Audre. 1984. *Sister Outsider: Essays and Speeches*. Berkeley, Calif: Crossing Press, c2007.
MacKenzie, Donald A. 1993. *Inventing Accuracy: A Historical Sociology of Nuclear Missile Guidance*. 4. print. Inside Technology. Cambridge, Mass.: MIT Press.
M’Charek, Amade. 2013. “BEYOND FACT OR FICTION: On the Materiality of Race in Practice: BEYOND FACT OR FICTION.” *Cultural Anthropology* 28 (3): 420–42. https://doi.org/10.1111/cuan.12012.
Mills, Charles W. 1997. *The Racial Contract*. Ithaca: Cornell University Press.
Ngai, Sianne. 2007. *Ugly Feelings*. First Harvard University Press paperback. Cambridge, Mass. London: Harvard University Press.
Onuoha, Mimi. 2020. “When Proof Is Not Enough.” *FiveThirtyEight* (blog). July 1, 2020. https://fivethirtyeight.com/features/when-proof-is-not-enough/.
Shevin, Michelle, Aasim Shabazz, and Marika Pfefferkorn. 2022. “Copowering an Emergent Horizon.” *Stanford Social Innovation Review*, July 14, 2022. https://ssir.org/articles/entry/co_powering_an_emergent_horizon#.
Sutherland, Tonia. 2019. “The Carceral Archive: Documentary Records, Narrative Construction, and Predictive Risk Assessment.” *Journal of Cultural Analytics*. https://doi.org/10.22148/16.039.
Trouillot, Michel-Rolph. 2011. *Silencing the Past: Power and the Production of History*. Nachdr. Boston, Mass: Beacon Press.
Tuck, Eve, and K. Wayne Yang. 2014. “Unbecoming Claims: Pedagogies of Refusal in Qualitative Research.” *Qualitative Inquiry* 20 (6): 811–18. https://doi.org/10.1177/1077800414530265.

Zulaika, Joseba. 2018. “What Do You Want?: Evidence and Fantasy in the War on Terror.” In *Bodies as Evidence*, edited by Mark Maguire, Ursula Rao, and Nils Zurawski, 201–27. Duke University Press. https://doi.org/10.1215/9781478004301-010.