# Studying Up: Reorienting the study of algorithmic fairness around issues of power

Chelsea Barabas
Massachusetts Institute of Technology
cbarabas@mit.edu

Colin Doyle
Harvard Law School
cdoyle@law.harvard.edu

JB Rubinovitz
Massachusetts Institute of Technology
jrubinov@mit.edu

Karthik Dinakar
Massachusetts Institute of Technology
Harvard Law School
kdinakar@media.mit.edu

## 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 blindspot 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 [12] 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. 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.

**ACM Reference Format:**
Chelsea Barabas, Colin Doyle, JB Rubinovitz, and Karthik Dinakar. 2020. Studying Up: Reorienting the study of algorithmic fairness around issues of power. In *ACM Conference on Fairness, Accountability, and Transparency, January 27–30, 2020, Barcelona, Spain.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/1122445.1122456


*ACM FAT* 2020, January 27–30, 2020, Barcelona, Spain*

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

## 1 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" [9, 14, 22]. 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 [35], data collection and preparation [40], claims-making [2, 33] and everyday use in applied contexts [39]. 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 [2, 5, 12, 33, 43].

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 [7, 12, 19, 39]. Researchers have problematized commonplace characterizations of data science as an objective and neutral process [27, 44], arguing that data and their subsequent analyses are always the by-product of socially contingent processes of meaning making and knowledge production [5, 26, 35, 44]. 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 [39].

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 [5, 12, 13]. 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 [12] 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" [12]. 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 understanding their work as situated practice. In this paper, we expand on Elish and Boyd's [12] 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 [31] called for her fellow anthropologists to expand their field of inquiry to include the study of elite individuals and institutions, who remained significantly underexamined 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 argued, "If one's pivot point is around those who have responsibility by virtue of being delegated power, then the questions change"[31].

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 (Bhavani 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.

## 2 THE CALL TO "STUDY UP"

### 2.1 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 argue, "studying up" indicates a relative direction between the researcher and subject, to study someone who has what Edward Said (1978, 7) referred to as "the relative upper hand" in terms of the amount of agency and authority they have in a given context [36].

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 [32]. 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 [32, 41]. These movements pushed anthropologists to critically examine the political nature of their work and to reflect on the blindspots 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 [36] have discussed, scholars like Berreman (1968, 395) "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 [16]. 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 [17].

It was in this context that Nader 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" [31]. 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 [11, 15, 23, 29, 34, 38]. As Priyadharshini 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." [38].

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) 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" [31]. 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 micropolitics 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 default 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 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 fairness, 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.

## 2.2 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 [21, 24, 37]. These controversies have given rise to an influential community of researchers from both academia and industry who have formed a new regulatory science [26] under the rubric of "fair, accountable, and transparent algorithms" [8, 28]. 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 [5, 25, 44]. 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 [5, 25]. Discriminatory design does not require intentional malice or prejudice on the part of the data scientist. As Benjamin (2016) 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" [4]. Similarly, D'Ignazio and Klein (2019) 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" [10].

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 [25]. 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. 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.

# 3 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.

## 3.1 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. Increases in pretrial incarceration rates are "responsible for all of the net jail growth in the last twenty years." 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." 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. 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.

America's unique reliance on money bail has allowed for the spike in pretrial incarceration. Money bail has become a loophole that judges use to bypass due process protections designed to ensure that pretrial detention is only used in exceptional circumstances in which a defendant clearly poses a threat to the community or is a flight risk. Most states have procedures through which a judge can intentionally detain someone pretrial. But 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 recent 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 that proceeds as follows: 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, pretrial risk assessments have been called into question after 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 [1]. This discrepancy in accuracy 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 [2, 33]. As scholars have pointed out, there are numerous challenges to modeling and effectively intervening on risks such as pretrial violence and flight [2, 18]. 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 [2, 33].

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. Narrow technical conceptualizations of how to fairly predict a pretrial defendant's behavior elide more fundamental issues. 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.

## 3.2 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 decoupling" whereby the everyday practices of the courtroom diverge from state and federal laws regarding pretrial release [6]. From this angle, it seems perplexing that interventions seeking to change judicial behavior would focus exclusively on modeling the behavior of people awaiting trial. 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 unnecessary 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 jurisdictions 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 person 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 [20, 42]. 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.

*3.2.1 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 rearrest. 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 abandon our study on 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 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 outcomes of interest predetermined the conclusion of incarceration as an effective intervention. Because we had no ability to measure other outcomes, we had to end our research on ethical grounds.

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.

*3.2.2 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 snapshots 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 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 [30]. 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.

*3.2.3 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 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," [5] we are developing a new risk assessment model. Rather than predict the behavior of pretrial defendants, this new model predicts the behavior of judges at bail hearings. This judicial risk assessment tool specifically predicts a judge's likelihood of "failing to adhere" to the U.S. Constitution by imposing unaffordable bail without due process of law. The algorithm – a stochastic gradient boosting machine – is more transparent and exceeds the accuracy and ROC scores of pretrial risk assessments currently in use. The model's training data is a combination of the court data given to us by the AOC and demographic information on the 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.

This algorithm is not intended for practical use. Indeed, many of the technical and ethical problems with defendant risk assessments also apply with a judicial risk assessment. But this 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 puts it, "[help us] better understand and expose the many forms of discrimination embedded in and enabled by technology" [5]. This imaginative work echoes the advice of the critical race scholar Derrick Bell, "To see things as they really are, you must imagine them for what they might be" [3]. Our goal in developing a judicial risk assessment was to render as intuitive the various ways risk assessment methods and discourse are limited

and stigmatizing, by subjecting those in positions of power to sociotechnical processes which are typically reserved for only the poor and marginalized.

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?

## 4 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. At the same time, this case study reveals many of the ethical and practical challenges to "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 not the powerful. Projects that question the ethics of existing data approaches will often require unearthing or assembling alternative datasets.

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 data that already exists, to imagine the data for what it might be.

## REFERENCES


[1] Julia Angwin et al. 2016. Machine Bias. (2016).
[2] Chelsea Barabas. 2019. Beyond Bias: Re-imagining the Terms of. *Criminal Law (April 25, 2019)* (2019).
[3] Derrick A Bell. 1995. Who's afraid of critical race theory. *U. Ill. L. Rev.* (1995), 893.
[4] Ruha Benjamin. 2016. Catching Our Breath: Critical Race STS and the Carceral Imagination. *Engaging Science, Technology, and Society* 2 (2016), 145–156.
[5] 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. [n. d.]. Captivating Technology: Race, Carceral Technoscience, and Liberatory Imagination in Everyday Life. ([n. d.]).
[6] Angéle Christin. 2017. Algorithms in practice: Comparing web journalism and criminal justice. *Big Data & Society* 4, 2 (2017).
[7] Sasha Costanza-Chock. 2018. Design Justice: towards an intersectional feminist framework for design theory and practice. *Proceedings of the Design Research Society* (2018).
[8] Nicholas Diakopoulos, Sorelle Friedler, Marcelo Arenas, Solon Barocas, Michael Hay, Bill Howe, HV Jagadish, Kris Unsworth, Arnaud Sahuguet, Suresh Venkatasubramanian, et al. 2017. Principles for accountable algorithms and a social impact statement for algorithms. *FAT/ML* (2017).
[9] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. ACM, 214–226.
[10] Catherine DâĂŹIgnazio and Lauren Klein. 2019. Data feminism. (2019).
[11] Mark Easterby-Smith, Richard Thorpe, and Paul R Jackson. 2012. *Management Research*. Sage.
[12] Madeleine Clare Elish and Danah Boyd. 2018. Situating methods in the magic of Big Data and AI. *Communication monographs* 85, 1 (2018), 57–80.
[13] Virginia Eubanks. 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.
[14] Michael Feldman, 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*. ACM, 259–268.
[15] J Fitz and D Halpin. 1994. Ministers and mandarins: Educational research in elite settings. Researching the powerful in education. G. Walford. (1994).
[16] Gutorm Gjessing. 1968. The social responsibility of the social scientist. *Current Anthropology* 9, 5, Part 1 (1968), 397–402.
[17] Kathleen Gough. 1968. New proposals for anthropologists. *Current anthropology* 9, 5, Part 1 (1968), 403–435.
[18] Lauryn P Gouldin. 2018. Defining Flight Risk. *U. Chi. L. Rev.* 85 (2018), 677.
[19] 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. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*.
[20] Hugh Gusterson. 1997. Studying up revisited. *PoLAR: Political and Legal Anthropology Review* 20, 1 (1997), 114–119.
[21] Jessica Guynn. 2015. Google photos labeled black people 'gorillas'. *USA Today 1* (2015).
[22] Eric Price Hardt, Moritz and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* (2016).
[23] Rosanna Hertz and Jonathan B Imber. 1995. *Studying elites using qualitative methods*. Vol. 175. Sage Publications.
[24] Kashmir Hill. 2012. How Target figured out a teen girl was pregnant before her father did. *Forbes* (2012).
[25] Anna Lauren Hoffmann. 2019. Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society* 22, 7 (2019), 900–915.
[26] Sheila Jasanoff. 1994. *Learning from disaster: risk management after Bhopal*. University of Pennsylvania Press.
[27] Niels Kerssens. [n. d.]. De-Agentializing Data Practices: The Shifting Power of Metaphor in 1990s Discourses on Data Mining. ([n. d.]).
[28] Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. 2018. Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology* 31, 4 (2018), 611–627.
[29] George E Marcus. 1983. *Elites, ethnographic issues*. Univ of New Mexico Pr.
[30] Khalil Gibran Muhammad. 2011. *The condemnation of blackness*. Harvard University Press.
[31] Laura Nader. 1972. Up the anthropologist: perspectives gained from studying up. (1972).
[32] Peter Novick. 1988. *That noble dream: The 'objectivity question' and the American historical profession*. Vol. 13. Cambridge University Press.
[33] Rodrigo Ochigame. 2019. The Illusion of algorithmic fairness. (2019).
[34] Teresa Odendahl, Aileen M Shaw, et al. 2002. Interviewing elites. *Handbook of interview research: Context and method* (2002), 299–316.
[35] Samir Passi and Solon Barocas. 2019. Problem formulation and fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 39–48.
[36] Rebecca Warne Peters and Claire Wendland. 2016. Up the Africanist: the possibilities and problems of 'studying up'in Africa. *Critical African Studies* 8, 3 (2016), 239–254.
[37] Rob Price. 2016. Microsoft is deleting its AI chatbotâĂŹs incredibly racist tweets. *Business Insider* (2016).
[38] Esther Priyadharshini. 2003. Coming unstuck: Thinking otherwise about âĂIJstudying upâĂİ. *Anthropology & education quarterly* 34, 4 (2003), 420–437.

[39] Andrew D Selbst and Solon Barocas. 2018. The intuitive appeal of explainable machines. *Fordham L. Rev.* 87 (2018), 1085.
[40] Sara Selwood. 2002. The politics of data collection: Gathering, analysing and using data about the subsidised cultural sector in England. *Cultural trends* 12, 47 (2002), 13–84.
[41] Mark Solovey. 2001. Science and the state during the Cold War: Blurred boundaries and a contested legacy. (2001).
[42] Daniel Souleles. 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 (2018), 51–68.
[43] Luke Stark and Anna Lauren Hoffmann. [n. d.]. Data Is the New What? Popular Metaphors & Professional Ethics in Emerging Data Culture. ([n. d.]).
[44] Luke Stark and Anna Lauren Hoffmann. [n. d.]. Data Is the New What? Popular Metaphors & Professional Ethics in Emerging Data Culture. *Journal ISSN* 2371 ([n. d.]), 4549.