# Carceral Machines: Algorithmic Risk Assessment and the Reshaping of Crime and Punishment

by

**Dasha Pruss**

BS, Computer Science, University of Utah, 2016

Submitted to the Graduate Faculty of

the Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

**Doctor of Philosophy**

University of Pittsburgh

2023

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Dasha Pruss

It was defended on May 22, 2023

and approved by

Colin Allen, Co-Advisor, Distinguished Professor, History & Philosophy of Science,

University of Pittsburgh

David Danks, Co-Advisor, Professor of Data Science and Philosophy,

University of California, San Diego

Edouard Machery, Distinguished Professor, History & Philosophy of Science,

University of Pittsburgh

Sandra Mitchell, Distinguished Professor, History & Philosophy of Science,

University of Pittsburgh

**Carceral Machines: Algorithmic Risk Assessment and the Reshaping of Crime and Punishment**

Dasha Pruss, PhD

University of Pittsburgh, 2023

Recidivism risk assessment instruments are used in high-stakes pre-trial, sentencing, or parole decisions in nearly every U.S. state. These algorithmic decision-making systems, which estimate a defendant's risk of rearrest or reconviction based on past data, are often presented as an 'evidence-based' strategy for criminal legal reform. In this dissertation, I critically examine how automated decision-making systems like these shape, and are shaped by, social values. I begin with an analysis of algorithmic bias and the limits of technical audits of algorithmic decision-making systems; the subsequent chapters invite readers to consider how social values can be expressed and reinforced by risk assessment instruments in ways that go beyond algorithmic bias. I present novel analyses of the impacts of the Sentence Risk Assessment Instrument in Pennsylvania and cybernetic models of crime in the 1960s Soviet Union. Drawing on methods from history and philosophy of science, sociology, and legal theory, I show not only how societal values about punishment and control shape (and are shaped by) the use of these algorithms – a phenomenon I term *domain distortion* – but also how the instruments interact with their users – judges – and existing institutional norms around measuring and sentencing crime. My empirical and theoretical findings illustrate the kinds of insidious algorithmic harms that rarely make headlines, and serve as a tonic for the exaggerated and speculative discourse around AI systems in the criminal legal system and beyond.

## Table of Contents

## List of Tables

## List of Figures

## Acknowledgments

Writing a PhD dissertation is not something done by one person in isolation. It truly takes a village, and I'd like to thank all of these co-inhabitants.

For giving me invaluable feedback and advice on the work in this dissertation, I would like to extend my deepest thanks to Marina DiMarco, Nedah Nemati, David Danks, Colin Allen, Camilo Ruiz, Michael Dietrich, Jessie Allen, Kevin Elliott, Maria Ryabova, John Norton, Jim Woodward, Katie Creel, Zina Ward, Sarah Brayne, Alex Albright, Cierra Robson, Rhys Hester, Hannah Pullen-Blasnik, Sam Plummer, Jack Samuel, Dorothea Anagnostopoulos, Gal Ben Porath, Dana Matthiessen, Nuhu Osman Attah, Dejan Makovec, JP Gamboa, Tom Wysocki, Jennifer Whyte, Seth Goldwasser, David Widder, Katrina Kish, Irene Pruss, Dmitry Pruss, Slava Gerovitch, Mario Small, Sandra Mitchell, Edouard Machery, and the HPS graduate student community.

I am deeply grateful to Jessie Allen for planting the seed for many of the ideas in this dissertation, for her incisive comments, and for our many stimulating discussions of jurisprudence and life over a glass of wine. I would like to give a special thanks to David Danks, Sophia Arbeiter, and Michael Dietrich for convincing me not to drop out of grad school. Thanks to Colin Allen for reading endless drafts of long grant applications and for his 'office hours on a bike', a model I hope to adopt with my own students someday.

Thank you to Nadia Narnor, Yusuf Jones, and the other members of the Coalition to Abolish Death by Incarceration (CADBI-West) for their invaluable feedback and for their powerful organizing work. Thank you also to Bonnie Fan, Maria Ryabova, Abhishek Viswanathan, and the other co-organizers of the Against Carceral Tech collective for your resistance, solidarity, and continuing inspiration.

Thank you to Slava Gerovitch for helpful discussion about the history of Soviet science, for bringing to my attention the existence of legal cybernetics, and for providing practical advice for my 2018 visit to Moscow. Thanks to the staff at the Pennsylvania Sentencing Commission for their cooperation and assistance with my research. Thank you to the Diana Volkar, Matt Ceraso, Joann McIntyre, Jennifer Berkebile, and the other staff who helped me

deal with the endless bureaucratic questions one encounters during a PhD. Thank you to Carol Cleland for her continued mentorship and support. And thank you to the National Science Foundation, the Horowitz Foundation for Social Policy, the University of Pittsburgh's Year for Data and Society grant, and the Wesley C. Salmon Fund for making the research in this dissertation possible.

To Annie Cherkaev, Camilo Ruiz, Gal Ben Porath, Paula Kupfer, Vivian Feldblyum, and my family: thank you all for your love and for keeping me sane these past years. Finally, I am grateful beyond words to Marina DiMarco and Nedah Nemati for reading so many drafts of my work and for unconditionally supporting me in the strange times we live in. This dissertation would not have been possible without their friendship.

## 0.0 Introduction

Since late 2022, the sheer amount of fascination and alarm about AI, especially generative deep learning models like GPT and DALL-E, has been jarring and oftentimes frustrating to witness for researchers studying the social impacts of technology. One only needs to open social media or listen to the news to see the latest projection of how AI will – in the near-to-long-term – fundamentally reshape every tenet of society, from automating white collar jobs to a litany of other algorithmic harms for middle class people.

During the same time period, studies have continuously emerged showing how public and private sector algorithmic decision-making systems have *already* caused substantive harms – for decades, in some cases – and have compounded structural inequalities faced by poor people and communities of color.[1] The Allegheny Family Screening Tool – an algorithm used to inform responses to alleged child neglect phone calls to Allegheny County, Pennsylvania's child welfare agency – flagged Black children and families with residents with disabilities as higher risk than comparable white children and families without disabilities (Cheng et al., 2022; Gerchick et al., 2022). Internal algorithms used by the IRS to determine who to audit led to disproportionately high audit rates for Black taxpayers (Miller, 2023). An insurance algorithm used by the Medicare Advantage program denied coverage in care to seniors in need, leading to massive medical bills and treatment delays (Ross, 2023). Study after study shows that these types of algorithmic harms are ubiquitous, insidiously chipping away at the quality of life of vulnerable groups and hardening systemic injustices (Eubanks, 2018; Benjamin, 2019).

That technology is "inescapably value-laden" (Mittelstadt et al., 2016, 1) has been recognized by academics for decades and is widely accepted in philosophy of technology today. Philosophers of science have questioned the value-free ideal in science – the idea that scientific reasoning can be free of non-epistemic (social, political, and moral) values – by illustrating the social values introduced by trade-offs in mitigating inductive and epistemic risk (Douglas,

[1]These examples are raised in a tweet by Logan Koepke:
https://twitter.com/jlkoepke/status/1641170906759266304

2009; Elliott, 2017).[2] Historians, legal scholars, and Science and Technology Studies scholars have argued for the inherent political nature and consequences of certain technologies (Winner, 1980; Feeley and Simon, 1994; Porter, 1995; Graham, 1987; Harcourt, 2007). Scholars in the field of fair machine learning have written extensively about formal ways to measure and remove bias in the predictions made by algorithms (Corbett-Davies and Goel, 2018; Chouldechova, 2016). Sociologists have shown that interactions between technologies and the people that use them – also known as a sociotechnical system – can amplify or subvert social values and goals in unexpected ways (Brayne, 2020; Christin, 2017). This dissertation builds on the insights of these disciplinary perspectives on the value-ladenness of technology and invites a broad perspective on the relationship between social values and algorithmic systems.

In the following four chapters, I draw on methods from history and philosophy of science, sociology, and legal theory to critically examine how algorithmic decision-making systems shape, and are shaped by, societal values. My research focuses on algorithmic recidivism risk assessment instruments, which estimate a criminal defendant's risk of rearrest or reconviction and are used throughout the United States to guide decision-making at multiple parts of the criminal legal pipeline. I show not only how societal values about punishment shape (and are shaped by) the use of these algorithms, but also how the instruments interact with their users – judges – and existing institutional norms around sentencing. My empirical and theoretical findings explain why algorithm-centric reforms like these can fail to live up to their hype, serving instead as performative acts without redressing institutional inefficiencies or biases. This dissertation thus illustrates the kinds of insidious algorithmic harms that rarely make headlines, and serves as a tonic for the exaggerated and speculative discourse around AI systems in the criminal legal system and beyond.

In the following sections, I briefly provide some context for how and why risk assessment tools are developed, identify the main concerns about their shortcomings, and situate the significance of my dissertation's findings in relation to these issues. I also provide a brief roadmap of the dissertation.

[2] I follow Biddle and Kukla (2017) in using the term 'epistemic risk' to refer to the risk of error at any stage of knowledge production, including inductive risk (how type I/II errors are weighed).

## 0.1 Risk Assessment Instruments

The injustices of the US criminal legal system are rooted in a long history of oppression and structural racism. The US incarcerates more people, and at higher rates, than any other country, and disproportionately arrests and incarcerates Black and poor people (Western and Wildeman, 2009; Hinton et al., 2018). Black people are also given harsher, longer sentences than white people for the same crimes, and this disparity has grown worse over time (Lopez, 2017).

In response, the 'evidence-based sentencing' reform movement has advocated for using science, technology, and big data to replace subjective and idiosyncratic penal decision-making, with the aim of reducing prison populations, increasing consistency in sentencing, and saving money (Klingele, 2016; Stevenson, 2018). Basing decisions on individuals' statistical risk of recidivism, as estimated by algorithmic risk assessment tools, is the movement's core strategy for achieving these goals (Starr, 2014; Hannah-Moffat, 2013). Evidence-based sentencing promotes risk assessment instruments on the basis that they are a "rational, objective, and empirically sound technology for improving decisionmaking" and "better predictors of recidivism than clinical judgments" (Hannah-Moffat, 2013, 271; Ægisdóttir et al., 2006). This advocacy strategically positions the mechanical objectivity associated with algorithms – that is, the minimization of human bias via mechanical procedures (Daston and Galison, 2007) – as a partial solution to the crisis of mass incarceration.

Broadly, risk assessment instruments are actuarial tools, that is, statistical measures of risk commonly used in the insurance industry to estimate an individual's likelihood of some future (typically unwanted) outcome, such as defaulting on a loan or having a car accident. A car insurance company, for instance, may set insurance premiums based on an individual's estimated risk of a crash based on the frequency of yearly crashes among the subgroup of the population with whom they share demographic features.[3]

In criminal risk assessment, the outcome of interest is typically recidivism, which is

[3]The practitioner's guide for COMPAS, a commonly used risk assessment instrument, explicitly compares the algorithm to risk prediction approaches used by the auto insurance industry to estimate accident risk (NorthPointe, 2015, p. 29).

operationalized as rearrest or reconviction within some time period after release.[4] Recidivism risk assessment tools use demographic factors like criminal history, gender, and age, which have been statistically correlated with recidivism in samples of (typically white male) inmates, to assign individuals a numerical risk score (Starr, 2014; Werth, 2019). This score is supposed to reflect an individual's risk of recidivism – the higher the score, the higher the likelihood of future crime.

An individual's risk score is used by judges or probation officers to inform pre-trial detention, sentencing, and parole decisions, depending on the jurisdiction. Often, different risk categorizations come with different recommended actions. A pretrial risk assessment, for instance, may be paired with a policy to release low-risk defendants from jail, assign medium-risk defendants bail, and detain high-risk defendants without the option of bail. Different risk assessment instruments can also be used at multiple phases of the criminal legal pipeline within the same jurisdiction. These algorithmic decision-making tools are developed by private companies (e.g., COMPAS developed by Northpointe), non-profit organizations (e.g., the Public Safety Assessment (PSA) developed by Arnold Ventures), and state agencies (often in partnership with universities – e.g., the Ohio Risk Assessment System (ORAS) developed by the University of Cincinnati and the Ohio Department of Rehabilitation and Correction).

Risk assessment is split into several 'generations'. Prison wardens started to classify and predict the behavior of their subjects on the basis of clinical judgment as early as the 19$^{\text{th}}$ century (Harcourt, 2007); the clinical evaluation of recidivism risk by trained individuals on a case-by-case basis is known as '1$^{\text{st}}$ generation' risk assessment (Hannah-Moffat et al., 2009). Beginning in the 1920s, '2$^{\text{nd}}$ generation' actuarial techniques began to be used to predict recidivism risk using static factors such as age at first arrest, prior convictions, and gender (O'Malley, 2010). The first such actuarial instrument was developed in the 1920s on a dataset of 3,000 parolees in Chicago; it used an individual's marital status, criminal history, and employment information to estimate their likelihood of reoffense (Burgess, 1936). Risk assessment instruments like these began to be widely used in the US criminal legal system in the 1980s (Feeley and Simon, 1992), and actuarial risk assessment has rapidly grown in its

[4]For pre-trial detention decisions, the predicted outcome is sometimes an individual's flight risk.

uptake and importance in the decades since (Werth, 2019). '3rd generation' risk assessments additionally incorporate dynamic, or changeable, risk factors, such as employment status, substance abuse, or education level (Andrews et al., 2016). Such factors are sometimes referred to as 'criminogenic needs' because they can be treated as possible loci for intervention (Werth, 2019), though in practice the 'needs' component of these assessments is ignored by practitioners (Bonta et al., 2008; Hannah-Moffat, 2005). '4th generation' risk assessments aim to more explicitly link assessment of risk with case planning and "risk management" (Werth, 2019; Public Safety Risk Assessment Clearinghouse, 2023).

The most commonly used risk assessment instruments use fairly simply algorithms. Most are 2nd generation (static factors only) or 3rd generation (static and dynamic factors) actuarial risk assessments and calculate a risk score in a manner similar to a check-list – each feature is assigned a number of points, based on how statistically associated it is with recidivism and its ability to discriminate high- and low-risk classes (Silver and Miller, 2002), and the sum of these points is typically categorized into low, medium, and high risk categories (Stevenson, 2018). Machine learning methods have been proposed to supplant traditional actuarial approaches in recidivism risk assessment but are not yet widely in use (Berk, 2012, 2019). Although actuarial risk assessments are computationally simpler than machine learning models, the development process of both is similar: data is collected on the outcome of interest in past cases (e.g., rearrest data in the state of Utah); a model is developed based on these data to optimize prediction success ("risk factors" statistically correlated with the outcome are identified and risk cutoffs for 'low', 'medium', and 'high' categories are set); and the model is used to inform decision-making (e.g., high-risk defendants are given longer sentences) (Fazelpour and Danks, 2021).[5] The main differences among mainstream recidivism risk assessment instruments are which population samples, risk factors, and risk thresholds they use (Hannah-Moffat,

[5]Hannah-Moffat (2019) draws a starker distinction between algorithmic risk assessment that uses machine learning and actuarial risk assessment. She argues that actuarial risk assessment is typically grounded in psychological theory, whereas "big data technologies are not constrained by preconceived theoretical or methodological disciplinary norms or necessarily administered and interpreted by certified assessors" (459). However, insofar as machine learning tools rely on the same datasets and measured qualities that more traditional actuarial risk assessment instruments do, the challenges I identify in the dissertation are reflected by both approaches, and the users of the algorithms are the same regardless. Machine learning approaches also have nearly identical results to more simple linear classifiers (Dressel and Farid, 2018). The main difference between the two is that machine learning based risk assessment produce more granularity in risk classes.

2013); the accuracy of different approaches does not vary much (Desmarais et al., 2018). For example, COMPAS' 137-feature assessment, which includes answers to a questionnaire, has a comparable accuracy rate – roughly 65% – to a risk assessment that uses only two variables, age and number of prior convictions (Dressel and Farid, 2018).

## 0.2 The Value-Ladenness of Algorithmic Risk Assessment

Since their inception, algorithmic risk assessments have been the target of much criticism (Feeley and Simon, 1994; Hannah-Moffat, 2013; Angwin et al., 2016). For one, the claim that risk assessment instruments are effective reforms in practice is largely speculative. The few existing empirical studies suggest that risk assessment tools have had little to no impact (Stevenson, 2018; Sloan et al., 2018; Garrett and Monahan, 2020; Stevenson and Doleac, 2021). As economist and criminal legal scholar Megan Stevenson starkly puts it, "Somehow, criminal justice risk assessment has gained the near-universal reputation of being an evidence-based practice despite the fact that there is virtually no research showing that it has been effective" (Stevenson, 2018, 306).

However, the criticism that has received by far the most attention concerns the tools' racial bias. A vocal chorus of critics has argued that risk assessments could exacerbate racial disparities in pretrial, sentencing, and parole decisions because they base predictions on (and reproduce) structurally racist patterns in the US criminal legal system. Legal scholar Bernard Harcourt argues that risk has become a proxy for race and thus that risk assessment instruments will "significantly aggravate the unacceptable racial disparities in our criminal justice system" (Harcourt, 2010, 2). Legal scholar Sonja Starr writes more broadly that basing sentencing decisions on risk assessment instruments "amounts to overt discrimination based on demographics and socioeconomic status" (Starr, 2014, 806). Likewise, Attorney General Eric Holder worries that basing sentencing decisions on demographic features "may exacerbate unwarranted and unjust disparities that are already far too common in our criminal justice system and in our society" (Horwitz, 2014).

Just as proponents of risk assessment tools emphasize their objectivity and superiority

to human judgment, this framing of the problems with risk assessment instruments centers the value-neutrality and fairness of the algorithms as an implicit goal. The premise that risk assessment instruments could in principle be made free of harmful social values – and should be used, were this neutrality to be the case – is often taken for granted in these debates. The conversation on risk assessment has largely neglected important empirical questions, such as whether judges and probation officers uncritically rely on predictive instruments – which are advisory – in their decision-making. Studies of risk assessment instruments have also sidelined the expertise of the communities most affected by, and most knowledgeable about, the ongoing effects of their implementation – communities impacted by incarceration. This dissertation faces these issues head-on.

I start by discussing the issue that has put risk assessment instruments in the spotlight: value-ladenness in the sense of algorithmic bias, that is, the systematic deviation of an algorithm's predictions from a normative standard. In chapter 1, I present an analysis of formal measures of algorithmic fairness, which underpin much of the scholarship dedicated to measuring and eliminating algorithmic bias. Building on work from the philosophical literature on mechanical objectivity, I illustrate the shortcomings of audits of risk assessment instruments that depend on these measures, with a focus on Carnegie Mellon University's audit of a recently-implemented recidivism risk assessment instrument, Pennsylvania's Sentence Risk Assessment Instrument. The work presented in the subsequent chapters invites readers to consider how social values can be expressed and reinforced by risk assessment instruments in ways that go beyond algorithmic bias.

In chapter 2, I use insights from jurisprudence to show that the use of risk assessment instruments in sentencing requires a version of legal formalism, which is widely rejected by legal scholars, as well as a consequentialist position on sentencing, two implicit normative commitments that can worm their way into how we reason about justice and prioritize a narrow set of risk-oriented interventions.[6] I use this case study to introduce my concept of *domain distortion*, the phenomenon in which scientific methods are both impacted and reinforce certain social values, thereby distorting how we reason about a domain of application.

[6]A version of this chapter is published in *Philosophy of Science* and won the Philosophy of Science Association's Mary B. Hesse graduate student essay prize.

In chapter 3, I examine the hype around computational crime prediction strategies in a markedly different geographical area and time period: the 1960s in the Soviet Union, when Soviet criminologists began adopting methods from cybernetics in attempts to predict crime, automate legal processes, and lend scientific credibility to the field. Using archival material I accessed and translated at the Moscow State Library in 2018, I illustrate another instance of domain distortion: I show how Soviet political values about crime and punishment became embedded in and gained scientific authority through formal modeling choices. Cybernetic models of crime – which excluded economic causal factors – helped revive the authority of Soviet criminology and were used to support political crime-reduction campaigns focused on 'moral rehabilitation', such as anti-alcohol campaigns. It is easy to recognize the political aims of formal models in sociotechnical systems we are external to; our own systems require the same scrutiny.

Finally, chapter 4 shows how social values become expressed through the interactions between individuals, organizational influences, and algorithmic tools. To study the impacts of the Sentence Risk Assessment Instrument, I interviewed judges, Pennsylvania Sentencing Commission members, and probation officers statewide about the tool. I designed my study in consultation with formerly incarcerated individuals from a justice reform organization to ensure I did not omit issues of critical importance to communities impacted by incarceration. I found that despite the promises of evidence-based sentencing and the perils of algorithmic bias, the Sentence Risk Assessment Instrument's effects are minimal because it is overwhelmingly ignored by judges, an instance of resistance to algorithms. I argue, however, that this algorithm aversion cannot be accounted for by individuals' distrust of the tools or automation anxieties, per the explanations given by existing scholarship. Rather, the instrument's non-use is the result of an interplay between three organizational factors: county-level norms about pre-sentence investigation reports; alterations made to the instrument by the Pennsylvania Sentencing Commission in response to years of public and internal resistance; and problems with how information is disseminated to judges. These findings shed new light on the important role of organizational influences on professional resistance to algorithms, which helps explain why algorithm-centric reforms can fail to have their desired effect. This study also supports an empirically-informed argument against the use of risk assessment instruments:

they are resource-intensive and have not demonstrated positive on-the-ground impacts.[7]

[7]A preliminary version of this paper is published in the *Data & Society Points* blog and a full version is forthcoming in the proceedings of the ACM conference on Fairness, Accountability, and Transparency (FAccT).

## 1.0 Meta-Mechanical Objectivity and the Limits of Algorithmic Fairness Audits

Amid the chaos of the early months of the pandemic, criminal courts in Pennsylvania were instructed to begin consulting the Sentence Risk Assessment Instrument when sentencing crimes. The actuarial tool uses demographic factors like age and number of prior convictions to estimate the risk that an individual will "reoffend and be a threat to society" – that is, be reconvicted within three years of release from prison (Pennsylvania Sentencing Commission, 2020). It was developed by the Pennsylvania Sentencing Commission, a legislative agency that advances "fairer and more uniform decisions at sentencing, resentencing, and parole" (Pennsylvania Commission on Sentencing, 2022).

Recently, however, recidivism risk assessment instruments like Pennsylvania's tool have become notorious for being racially biased. From 2017–2019, the Commission received over 100 overwhelmingly negative public testimonies about the Sentence Risk Assessment Instrument from sources including AI Now, the ACLU, high-profile academics, and local community organizations. Critics argued that the "racist tool" (ACLU of Pennsylvania, 2019) could "perpetuate the racial biases and stigmas inherent in our criminal legal system" (Coalition to Abolish Death by Incarceration, 2019). Much of the concern focused on the potentially biased predictions that would result from the tool's use of racially-correlated demographic variables and data.

In response to these allegations, the Pennsylvania Sentencing Commission solicited an external audit by researchers at Carnegie Mellon University (CMU). The audit evaluated the accuracy and algorithmic fairness of the instrument and recommended several technical changes, including numerical adjustments to the tool's risk category cutoffs (Becerril et al., 2019). It also estimated the tool's projected impact statewide.

In this chapter, I set the stage for the rest of the dissertation by sketching the bounds on what technical audits like these are able to demonstrate about the bias and impacts of algorithmic systems. I focus on the formal fairness definitions used in the field of fair machine learning, also known as algorithmic fairness. I argue that the methodology of algorithmic fairness reproduces the shortcomings of mechanical objectivity – the minimization of human

bias via strict rule-based protocols – but on a meta-level. Much like mechanical objectivity is intended to remove individual or idiosyncratic (human) bias through the use of a mechanical procedure (such as an algorithm), what I call *meta-mechanical objectivity* is intended to remove algorithmic bias through conformity to mechanical fairness rules. In Theodore Porter (1995)'s historical analysis, he argues that quantification in service of mechanical objectivity is adopted by (or imposed on) weak bureaucratic elites who lack public trust or authority. Building on this finding, I argue that the need for *meta*-mechanical objectivity arises from a corresponding lack of trust in or authority of algorithmic decisions. I show that the range of criticisms of algorithmic fairness approaches can be helpfully understood on this analogy, illustrating the limits of technical audits that use this methodology.

I begin by introducing the relationship between algorithms and mechanical objectivity. Next, I discuss the phenomenon of algorithmic bias and formal fairness definitions that have been proposed in fair machine learning to remedy the issue. Drawing on the analogy of meta-mechanical objectivity, I explain the shortcomings of standard statistical and causal approaches, which result from the partial nature of each measurement technique and require mediation through interpretation and value-laden choices. Using CMU's audit as an illustration of these issues, I discuss shortcomings in their fairness assessment and projected impacts, which I argue are unfounded based on the analysis done but serve to lend the instrument legitimacy.[1]

### 1.1 Algorithms and Objectivity

In a 2014 TED talk, the former Attorney General of New Jersey Anne Milgram describes assembling a team of statisticians at the Arnold Foundation[2] to figure out a way to put "dangerous people" in jail while releasing those who pose no threat to society (Milgram, 2014). Her data science team developed the Public Safety Assessment (PSA) tool, "a research-based,

[1] In chapter 4, I show empirically that the instrument's actual impacts are different from what either the audit or the critics of the instrument anticipated: the tool has no impact at all.

[2] This formerly non-profit philanthropic organization recently became an LLC and re-branded itself as Arnold Ventures.

data-driven pretrial risk assessment tool that provides judges with objective information about the likelihood that a defendant will commit a new crime or will fail to return to court" (Arnold Ventures, 2017). The PSA uses seven factors, mostly pertaining to age and past criminal history, to estimate the risk of new crime (that is, new arrest while on pretrial release). These factors are assigned weights and summed to a score ranging from 1 to 16 (Advancing Pretrial Policy and Research, 2023).

In describing her motivation for developing the PSA, Milgram appealed to the limitations of human decision-makers:

> Judges have the best intentions when they make these decisions about risk, but they're making them subjectively. They're like the baseball scouts twenty years ago who were using their instinct and their experience to try to decide what risk someone poses. They're being subjective, and we know what happens with subjective decision making, which is that we are often wrong.[3]

The blanket association between objectivity and algorithms continues to get a lot of popular traction. But algorithms, much like humans, are poor at predicting complex social outcomes. Arvind Narayanan (2019) coined the apt term "AI snake oil" to refer to algorithms that are falsely claimed as doing so. Typical recidivism risk assessment instruments – including the PSA, privately-developed tools like COMPAS, and publicly-developed tools like the Sentence Risk Assessment Instrument – have an area-under-curve (AUC)[4] in the mid 0.6 range, meaning they perform just better than a coin flip in distinguishing low- and high-risk individuals (Desmarais et al., 2018). This is a modest improvement over untrained human ability to predict recidivism (Goel et al., 2021), though trained human judgment and even simple classifiers with only two features achieve the same accuracy as more complex algorithms (Dressel and Farid, 2018). As I will discuss at length in this chapter, there is also widespread concern about the racial bias of risk assessment instruments. But despite abundant examples of discriminatory and inaccurate algorithms (O'Neil, 2016; Angwin et al., 2016; Noble, 2018; Eubanks, 2018), in practice, the perception that algorithms are objective

[3]This quote appears in Galison (2019).

[4]AUC is a standard statistical measure to assess the ability of algorithms to differentiate between higher and lower risk individuals. It represents the probability that a randomly selected data point with the outcome of interest (e.g., a person who went on to be rearrested) would have received a higher risk rating than a randomly selected data point without the outcome of interest (a person who was not rearrested) (Desmarais et al., 2018).

remains a key justification for their adoption in criminal courts and policing (Brayne and Christin, 2020).

Objectivity suggests a rigor and neutrality that Thomas Nagel famously described as the "view from nowhere" (Nagel, 1989). However, the concept of objectivity is multivalent and contextual – Heather Douglas (2009) identifies eight versions of objectivity just with respect to processes of scientific inquiry. This ambiguity in term has provoked a disparaging attitude from philosophers like Ian Hacking (2015), who in his provocatively titled "Let's Not Talk About Objectivity" writes that objectivity is an "elevator word" – a statement about a statement – and that we should just stick to talking about "ground-level" questions.

While I share Hacking's exasperation, I believe it is important to talk about objectivity. The objectivity associated with science is what gives it a privileged status as a way of generating knowledge about the world. Being labeled objective – or not – confers epistemic authority (and, in the case of crime prediction instruments, state authority). Insofar as false claims to objectivity or a misunderstanding of the limits of objectivity might lead to misplaced authority of technology or the adoption of faulty algorithmic decision-making methods, it is worth dwelling, if briefly, on the concept's most important senses and how they have evolved over time.

### 1.1.1 Mechanical Objectivity

Daston and Galison (2007) use the history of scientific images to show that the relationship between objectivity and science has gone through phases, in keeping with cultural and technological developments and the corresponding "epistemic virtues" of different historical periods. These changing norms of objectivity provide a helpful foil to understanding the contemporary "risk assessment era" in the criminal legal system (Starr, 2015).

Prior to the advent of photography, the $18^{\text{th}}$ century epistemic virtue "truth-to-nature" favored representing nature through scientific drawings – "reasoned images" – that captured the essential qualities of natural phenomena, perfecting and abstracting away nature's variability in the process. This produced "ideal" or "typical" representations of natural phenomena conducive to classification and standardization (as well as aesthetic pleasure).

Over time, the downsides of these idealizations became evident: scientists depicted nature in many different and inconsistent ways. The 1830s saw the emergence of a new epistemic virtue, "mechanical objectivity" – the repression of any conscious interventions by the creators of scientific images. The central tenet of mechanical objectivity was to minimize the influences of human contribution to the scientific process, including theoretical and idiosyncratic judgments, replacing these with the rigor and consistency of strict protocols and procedures that are "free from the inner temptation to theorize, anthropomorphize, beautify, or interpret nature" (139). With the spread of daguerrotypes and cameras in the 19$^{\text{th}}$ century, nature could be represented with its variability intact. Nature, it was said, could "speak for itself," (120) in all its particularities – in a manner that seemed "pure" and "uncontaminated by interpretation" (139). Of course, these techniques could not entirely rid scientific image production from error or interpretation, but objectivity served as a regulative ideal, which scientists pursued through the "self-surveillance" of their own discretion.[5]

Machines were also seen as embodying admirable qualities that humans, particularly human workers, sometimes lacked. As Charles Babbage, considered the father of computing, candidly put it: "One great advantage which we may derive from machinery is from the check which it affords against the inattention, the idleness, or the dishonesty of human agents" (Babbage, 1833, 54). Machines were considered especially advantageous for repetitive, delicate, physically strenuous tasks.

By the 1930s, however, scientists began confronting the limitations of mechanical procedures. Cameras were only able to capture certain dimensions of the natural world and could even add artifacts of their own to representations of scientific phenomena. Printouts created by technical procedures such as X-ray machines could be misleading or confusing because of their excessive (or insufficient) information. Mechanical objectivity, in other words, was "not sufficient" (314) and was even "costly – in different contexts, it demanded sacrifices in pedagogical efficacity, color, depth of field, and even diagnostic utility" (179). Moreover, trained observers were outperforming mechanical procedures at complex tasks, such as distinguishing different kinds of seizure readings from an electroencephalogram.

[5]Daston and Galison note, however, that photography was not coextensive with suspicion toward human intervention, and mechanical objectivity never replaced truth-to-nature; early photography was often used in service of truth-to-nature, and photos were commonly manipulated.

Throughout the 20th century, scientific "experts" abandoned mechanical objectivity in favor of the epistemic virtue of "trained judgment," in which scientists supplemented and altered the products of mechanical procedures with their trained intuitions, expertise and artistry. As scientists gained confidence through professional training and status, the scientific self began to be seen as something to be cultivated, rather than removed.

As Porter argues in his historical analysis of quantification in bureaucratic settings, and as I discuss at more length in chapter 3, increased interest in mechanical objectivity can emerge as a response to declining institutional authority. Although the earliest photographic techniques were already widely available in the 1830s, it was not until the 1880s, when the subjective contributions of scientists began to be seen as an "epistemological danger" (Daston and Galison, 2007, 198) to science, that mechanical image production techniques began to be widely adopted in service of mechanical objectivity. Porter writes that "Strategies of impersonality must be understood partly as defenses against such suspicions" – mechanical procedures are responses to declines in expert authority because they entail that decisions do not depend too much on any one individual (Porter, 1995, 229). Porter, Daston, and Galison's analyses are readily applicable in the context of algorithmic decision-making and to risk assessment instruments in particular: algorithms are used to promote the epistemic virtue of mechanical objectivity, in part as a response to declining trust in human judgment (Christin, 2016; Galison, 2019).

Consider the shift in the authority of expert decision-makers that took place in the US criminal legal system in the 1980s. The post-war period through the 1970s saw a rehabilitative era of criminal sentencing, characterized by discretionary decision-making by expert decision-makers – judges – who tailored punishment and 'treatment' decisions according to individual needs, with the aim of rehabilitating inmates (Phelps, 2011). But with evidence for the positive effects of this sentencing model lacking and political attacks on racially-biased and indeterminate sentencing decisions on the rise, in the 1970s, judicial discretion began to be scrutinized (Martinson, 1974; Garland, 2002). Since the 1980s, there has been a tendency in the criminal legal system to view the role of experts like judges with suspicion and to blame them for the system's unfairness or inefficacy. Reagan-era tough-on-crime policies, including federal sentencing guidelines and mandatory minimum sentences, limited judicial discretion

and increased sentence lengths – particularly for drug possession – which is widely thought to have led to the crisis of mass incarceration and the disproportionate incarceration of Black Americans (Alexander, 2012).

In addition to these discretion-limiting policy changes, the 1980s saw a sharp rise in the use of actuarial risk assessment tools to reduce the subjectivity and idiosyncrasy of penal decision-making (Feeley and Simon, 1992, 1994). Today, the "risk assessment era" is in full swing in the criminal legal system (Starr, 2015). Pretrial risk assessment instruments, in particular, are used in some capacity in preliminary arraignment decisions in most states in the US, though there is wide variability in their prevalence county-to-county (Figure 1). The contemporary reform movement of 'evidence-based sentencing', which promotes consistency and 'fairness' in sentencing decisions through the use of algorithmic risk assessment instruments, is one instance of a broader trend of compensating for the decline of expert decision-makers through the adoption of mechanical objectivity – by supplanting or limiting human judgment with mechanical procedures like risk assessment instruments.

As Daston and Galison showed, mechanical objectivity is one epistemic virtue among others, and it ebbs and flows over time. Much like scientists of the 1930s lamented the technical shortcomings of cameras and X-rays, concern about the shortcomings of algorithms as mechanical decision-making tools has now become widespread. Much of this concern been centered around algorithmic bias, to which I now turn.

### 1.2 What is Algorithmic Bias?

"There's software used across the country to predict future criminals. And it's biased against blacks." This is the headline of ProPublica's blockbuster audit of COMPAS, which shows that the commonly-used recidivism risk assessment instrument makes different types of classification errors for Black defendants and white defendants: Black defendants are more likely to be falsely classified by COMPAS as 'high-risk' for recidivism, while white defendants

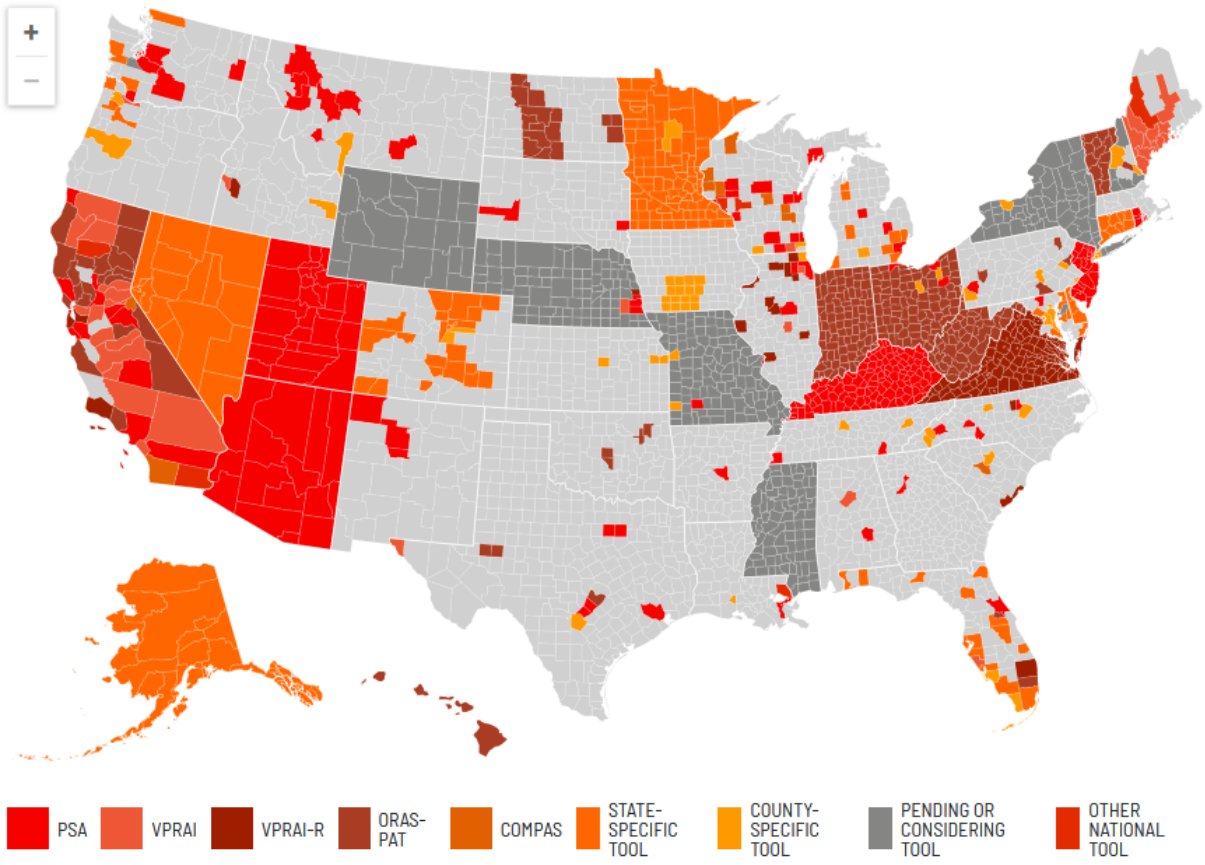


Figure 1: A map of pretrial risk assessment tools in the US, based on the Movement Alliance Project's database of over 300 jurisdictions (Movement Alliance Project, 2023).

are more likely to be falsely classified as 'low-risk' (Angwin et al., 2016).[6]

That COMPAS is a biased algorithm seems intuitive. It appears to violate at least two desirable standards, in this case an epistemic standard (the algorithm captures things as they really are in the world) and a moral standard (the algorithm is fair and impartial). At its most general, that is what bias is: a systematic deviation from some normative standard.

Philosophers of science (Friedman and Nissenbaum, 1996; Danks and London, 2017;

[6]COMPAS is one of the commonly used recidivism risk assessment instruments in the US. By comparing 137 factors, like answers to a questionnaire and defendant demographics (excluding information about race), to those of previous defendants, COMPAS calculates a recidivism risk score between 1 and 10 (NorthPointe, 2015). This score is included in a defendant's presentence investigation report, which is presented to a judge at the time of sentencing or preliminary arraignment (Forward, 2017).

Fazelpour and Danks, 2021; Johnson, 2020a) have given accounts of algorithmic bias that show the varied ways in which algorithms can be biased. Algorithms may have bias deriving from different standards, as well as different sources – training data, problem specification, technological constraints, and so on. Philosophical accounts of algorithmic bias are largely agnostic about there being one 'correct' way to measure bias – different measurements cater to different standards or virtues, which pull in different directions.[7]

Gabbrielle Johnson (2020a) argues that algorithmic bias is functionally analogous to implicit cognitive biases. She characterizes bias as a natural kind that plays a functional role in enabling induction in situations of underdetermination, which illuminates one important source of bias: the reproduction of existing patterns in training data, learned by cognitive and algorithmic systems alike as a heuristic for navigating an uncertain world. In Johnson's words, biases "bridge the otherwise limitless inductive gap that exists between evidence and theory" (Johnson, 2020b, 3) and operate anywhere induction does. For instance, a machine learning system built to navigate an obstacle course will face an underdetermination problem because the machine's sensors only have two-dimensional light stimuli available to them. To compensate, the algorithm will learn generalizations about the environment, such as the imperfect but often good-enough assumption that light tends to come from above (Johnson, 2020a, 13). Analogously, because the social world is shaped by social biases – racism, classism, ageism, and so on – an algorithm built to make inferences in this environment "necessarily adopts and utilizes assumptions that mimic patterns presently existing in the data on which it is trained" (14).

No doubt, the species of algorithmic bias that seem most alarming are those that most closely mimic familiar social stereotypes and implicit biases, which are often reflected in training data. Algorithmic bias can have sources that do not have a ready analogue in

[7]Allegations of algorithmic bias are sometimes driven by disagreements with social values underlying the construction or use of an algorithm in the first place, such as the prioritization of some performance goal over another or the kinds of interventions an algorithm is used for (Fazelpour and Danks, 2021). One might, for example, criticize the use of sentencing algorithms on the grounds that the instruments' objectives are flawed. For the purposes of this chapter, I will treat issues like these as distinct from the problem of algorithmic bias in the sense of biased predictions, though in practice these issues often intersect, insofar as what an algorithm optimizes influences the kind of biases its predictions have. Broader questions about the purposes and societal consequences of predictive technologies will be addressed at length in each of the subsequent chapters. To motivate the centrality of these substantive questions, my aim in this chapter is to first illustrate the limitations of the strictly technical approach to algorithmic fairness.

cognitive bias, however, which is in tension with Johnson's deflationary attitude that "issues surrounding algorithmic bias are not unique to algorithmic decision-making" (21). Consider Google Flu Trends (GFT), an algorithm that used Google search keywords pertaining to flu symptoms to predict flu season trends in the US. GFT was notoriously biased: it systematically overestimated the incidence of the disease. In the 2011–2012 flu season, for instance, GFT overestimated the prevalence of flu for 100 out of 108 weeks, with errors not randomly distributed – a previous week's errors were predictive of the current week's errors. Per Johnson, one source of this bias had to do with overfitting patterns in the training data; as one critic put it, this early version of GFT was "part flu detector, part winter detector" (Lazer et al., 2014, 1203). But the dynamics of the search algorithm itself also played a crucial role in the algorithm's biased predictions. Google search results for physical symptoms like 'fever' could lead to search recommendations about flu treatment, compromising further searches made in response to such suggestions but that were nevertheless represented in the training data.

Johnson's argument is also unable to account for the normativity sometimes, but not always, associated with algorithmic bias. Treating all generalizations equally – as strategies to overcome underdetermination – fails to demarcate technically incorrect yet morally acceptable generalizations (such as the generalization that all light comes from above) from generalizations that are technically accurate yet morally objectionable (such as the generalization that older people are less computer literate).

By contrast, Batya Friedman and Helen Nissenbaum (1996) give an explicitly normative definition of algorithmic bias: the phenomenon in which computer systems "systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others." This account implicitly takes deviation from the standard of neutrality to define algorithmic bias, which matches the common usage of the term in the contemporary literature on algorithmic fairness. But Friedman and Nissenbaum gloss over the trade-offs between neutrality at the decision-making stage and neutrality at the impact stage, which, as I will show in the section on statistical measures of fairness, are often in tension. They write, for instance, that "systematic discrimination does not establish bias unless it is joined with an unfair outcome," which not only equates bias and fairness but also groups what is known

in discimination law as disparate treatment with disparate impact (see section 1.3). In an improvement over Johnson's account, however, they identify multiple sources of bias: pre-existing bias in social institutions and practices; technical constraints such as hardware limitations and the formalization of qualitative categories; and the context of application, such as mismatches in expertise between users and system design expectations.

Sina Fazelpour and David Danks (2021) give the broadest and most comprehensive account of bias. While Friedman and Nissenbaum give a normative account of bias according to one standard, Fazelpour and Danks (see also Danks and London, 2017) use bias to refer to a deviation from any standard (e.g., statistical, ethical, legal); the normativity of bias in situated in the normativity of the relevant standard. For instance, a recidivism risk assessment instrument whose predictions systematically deviate from training data could be statistically biased, or morally biased if its predictions depend on an individual's race or gender. An algorithm can thus be biased in many ways, some problematic and some desirable; one form of bias can even be used to compensate for another type of bias (such as using smoothing to reduce the risk of overfitting noisy data).

Fazelpour and Danks also explain that the source of algorithmic bias can be helpfully understood as a problem of value-ladenness, where non-epistemic value judgments at each stage of the algorithmic development process become 'baked in' and expressed in the algorithms' predictions. These stages are similar to, but more detailed than, those given by Friedman and Nissenbaum, and include problem specification, data, modeling & validation, and deployment (see Figure 2); each of the many decision points within each stage requires value judgments. Notably, their account thus uses algorithmic bias broadly to refer both to biased algorithmic predictions and bias as a broader quality of a sociotechnical system (e.g., biased objectives and biased deployment); formal measures of algorithmic bias tend to focus on the former, though understanding the latter as a deviation from a standard is more challenging. Before turning how to measure bias quantitatively, I will briefly illustrate each of these sources of bias with reference to the running example of recidivism risk assessment instruments.

The problem specification stage determines the aims of the algorithm, including which target variables are to be predicted. In the case of recidivism risk assessment instruments, the outcome variable of interest is future crime. As I discuss in chapter 3, crime is a

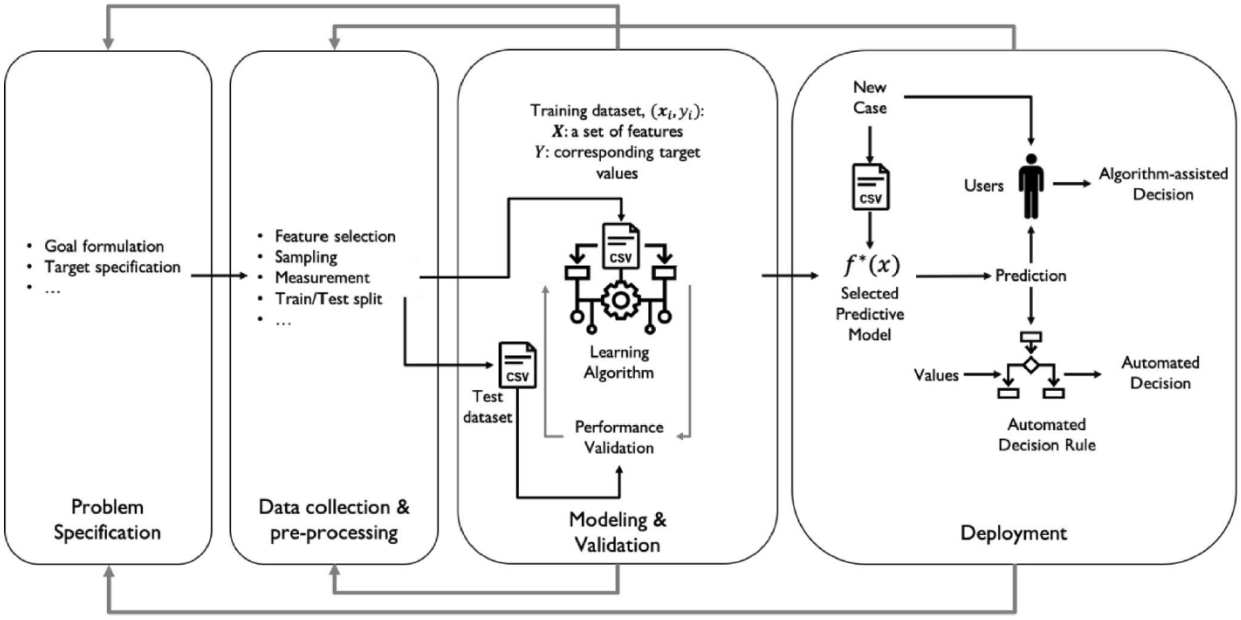


Figure 2: "Algorithm-driven and -assisted decision-making pipeline" (Fazelpour and Danks, 2021, 5).

legally, geographically, and culturally contingent category that evolves over time; in the US, it currently includes behaviors as diverse as drunkenness, welfare fraud, purse-snatching, prostitution, and bribery (Federal Bureau of Investigation, 2011). At the level of individual recidivism, crime is often operationalized as arrest or conviction over some time period, which simplifies the collection of training data. These specifications of crime are value-laden decision points that can result in biased predictions and are often contested. The Sentence Risk Assessment Instrument, for example, initially used re-arrest data as its target variable, but public testimonies argued that arrest is not only a poor predictor of actual crime – a technical probation violation like consuming alcohol can lead to rearrest – but also racially correlated due to racial profiling by police (Sassaman, 2018). In response, the Pennsylvania Sentencing Commission changed the target variable to re-conviction (Pennsylvania Commission on Sentencing, 2019c). The aims of risk assessment instruments, in turn, range from consistency in decision-making; public safety; prison population reduction; and, more rarely, allocation

of scarce resources for rehabilitation (within prisons or upon release). Any of these goals could be translated to a target variable for prediction other than crime, or could be attained without the use of predictive instruments. For instance, an algorithm that predicts which judges are most likely to give unduly long sentences could be an alternate strategy to increase consistency in sentencing, and ending cash bail could be a non-prediction-based strategy to reduce jail populations. How these aims are translated to prediction tasks is an important decision point that requires significant deliberation and value judgments.[8]

Biases in data, as Johnson's account showed, are inherited by algorithms. Fazelpour and Danks explain that data-level biases can be reflections of system-level biases as well as a result of measurement methods, such as under-sampling certain groups. Re-arrest and re-conviction data, for instance, is widely thought to inherit racial bias from structurally racist decision-making patterns within the criminal legal system and in US society more broadly – depending on the city, Black people can be up to four times more likely to be stopped by police than white people (Pierson et al., 2020). But racial biases in recidivism data also likely arise due to measurement problems. White collar crime is routinely underreported, and Black people are more likely than white people to be charged with drug crimes, even though drug usage rates are comparable between both groups (Rosenberg et al., 2017). Finally, how each risk category cutoff is set requires weighing the relative costs of false positives and false negatives (see the discussion of inductive risk in chapter 2).

Modeling and validation biases emerge when predictive models are optimized based on training data and later tested (validated) on a separate set of data. Developers may want to maximize the model's accuracy or minimize the disparate distribution of errors between groups, depending on their goals. This is also the stage at which the fairness of the algorithm's predictions is measured and the model is adjusted accordingly, potentially introducing new forms of bias, as I will discuss in the following section.

The final source of bias Fazelpour and Danks describe is deployment bias. Users' values can differ from the algorithm's values, perhaps without the users' awareness. Judges may not understand the purpose of a risk assessment instrument, what exactly it predicts or how

[8]In chapter 2, I explore the value-laden positions on punishment and intervention presupposed by the use of recidivism risk assessment instruments.

its predictions were generated, and the algorithm's outputs may be communicated poorly. For example, judges may use the algorithmic predictions selectively depending on factors like trust (Dietvorst et al., 2015) or fear of managerial surveillance (Brayne and Christin, 2020), and in a manner that can amplify or otherwise interact with existing human and institutional bias (Stevenson and Doleac, 2021).[9] In Kentucky, for instance, a risk assessment tool increased racial disparities in pretrial releases and ultimately did not increase the number of releases overall because judges ignored leniency recommendations for Black defendants more often than for similar white defendants (Albright, 2019). Deployment bias may also arise if an algorithm is deployed in a context sufficiently different from the one it was trained in.[10] The Sentence Risk Assessment Instrument was trained on recidivism data collected from 2004 to 2006, so demographic and reform changes over the last two decades could contribute to the instrument's biased predictions (Becerril et al., 2019). Algorithms can also interact in feedback loops, which shape future data; a choice to incarcerate someone longer on the basis of a high recidivism risk score could affect that individual's actual recidivism risk because past incarceration is predictive of future conviction – indeed, criminal history is one of the predictive factors the algorithm uses.

Given this understanding of the sources and kinds of algorithmic bias, I now discuss efforts to quantify and remove bias in algorithmic predictions, a sub-discipline of fair machine learning commonly referred to as algorithmic fairness.

### 1.3 Meta-Mechanical Objectivity: Quantifying Algorithmic Bias

We saw in subsection 1.1.1 that one possible response to the limitations of mechanical procedures is to pursue the epistemic virtue of trained judgment, that is, to develop human expertise and artistry in the use and interpretation of the mechanical protocol. But another possible response to the limitations of mechanical procedures is to compensate through additional mechanical procedures. The latter approach is implicitly adopted in the

[9] I discuss this human-algorithm interaction at length in chapter 4.

[10] This is a species of an external validity problem, well familiar in the replication crisis.

methodology of algorithmic fairness: relying on a quantitative procedure (a formal fairness metric) to compensate for the limitations of another quantitative procedure (an algorithm). To put this in the language of epistemic virtues, we can say that algorithmic fairness is a response to algorithmic bias that adds an additional mechanical objectivity constraint on top of the mechanical objectivity of the algorithm. We can helpfully think of this epistemic virtue as *meta*-mechanical objectivity. Meta-mechanical objectivity diverges from mechanical objectivity in an important respect, namely, that it is intended to minimize *algorithmic* bias rather than *human* bias, which, as we saw in the previous section, are not co-extensive. Formal measures of algorithmic bias concern the deviation from a particular kind of moral and epistemic standard, namely, fairness (as compared between two groups).[11] Much like mechanical procedures aimed at increasing consistency in human decision-making, formal fairness definitions aim at increasing how consistently the algorithm treats different groups of people.

Much like mechanical objectivity by way of algorithmic risk assessment emerged as a response to a crisis of confidence in human judges, the meta-mechanical objectivity of algorithmic fairness was a response to a crisis of confidence in technology. ProPublica's audit called attention to the racial bias of risk assessment instruments (Angwin et al., 2016), but this suspicion was directed not only at algorithms but also toward the companies that produce them. Anthropologist Rodrigo Ochigame (2019) argues that the discipline of algorithmic fairness was funded and promoted by big tech companies in the 2010s as a strategic response to public outcry over scandalous revelations in Silicon Valley, such as Facebook's breach of private data on 50 million users to the Trump-hired political marketing firm Cambridge Analytica, and Google's partnership with the Pentagon to analyze military drone surveillance (Project Maven). Private companies' interest in demonstrating that they could self-regulate – without the need for external, legal regulation – led to funding and support for the quantitative study of algorithmic fairness, an intentionally narrow and technical perspective on 'ethical AI'.

[11] There are also individual definitions of fairness, which are based on the idea that "similar individuals should be treated similarly," where similarity is determined on a case-by-case basis (Chouldechova and Roth, 2018). These require significant assumptions, such as an agreed-upon metric of similarity, "whose definition would itself seemingly require solving a non-trivial problem in fairness" (4; see also Fleisher, 2021). Because this metric is case-based, it also does not generalize easily – or have the mechanical quality of the other mechanical objectivity-promoting fairness measures, so I have chosen not to include it here.

The meta-mechanical objectivity of formal fairness definitions was thus a strategic move – an attempt to satisfy public demands by demonstrating that algorithms conform to industry-wide fairness standards, while skirting more substantive ethical scrutiny and regulation.

Like the algorithms adopted for mechanical objectivity, the algorithmic fairness metrics adopted for meta-mechanical objectivity have limitations. Each algorithmic fairness metric quantifies a shortcoming in an algorithm's predictions that must be remedied in order for it to be considered 'fair', depending on how the standard of fairness is interpreted. But relying on metric one rather than another requires value-laden choices, which do substantive normative work in the process of identifying and 'fixing' biased algorithms, which in turn introduces bias on other dimensions. Statistical definitions of algorithmic fairness, in particular, require equality in classification error rates (treatment parity) or actual outcomes (outcome parity) between members of different protected groups. These definitions each capture an important sense of systematic bias, but they have known limitations and can contradict each other. Others have proposed that causal fairness definitions could remedy the shortcomings of these approaches by requiring the absence of problematic causal structures involving protected variables, but causal approaches have their own shortcomings. In this section, I discuss statistical measures of algorithmic bias and causal measures of fairness, showing that each provides a partial and value-laden perspective on algorithmic fairness.

### 1.3.1 Statistical Measures of Fairness

The standard behind statistical fairness definitions is captured by two commonly cited legal concepts from anti-discrimination law. The first, known as disparate treatment, occurs when members of protected groups are treated unfairly during a decision-making process. Title VII of the US Civil Rights Act, for instance, says that employment decisions cannot discriminate against individuals based on their protected characteristics. This is intended to prevent intentional discrimination, whether based directly on protected features or based on proxies of protected features. Disparate impact, on the other hand, addresses practices that might not appear discriminatory at the decision-making stage, but nevertheless have adverse outcomes for members of a protected group.

Formal fairness definitions are essentially attempts to translate these qualitative legal doctrines into formal definitions. Consider the demographic features one might measure about an individual: age, gender, race, credit history, zip code, favorite 70s band, and so on. This list of features can be partitioned into its protected (e.g., race, gender) and unprotected (e.g., favorite 70s band) features. One might think that an easy way to satisfy treatment parity would simply be to exclude the protected features from an algorithm's consideration altogether. This attitude – that people may not be classified based overtly on their protected group membership – is known in discrimination law as 'anti-classification' and in algorithmic fairness as 'fairness-through-unawareness'. According to this definition, an algorithm is fair if it makes the same decision for two individuals with the same unprotected features (and protected features were not used in the decision).

The problem with anti-classification is that even if protected features are not directly considered, an algorithm can still make predictions that correlate with protected features by making predictions based on proxies of protected features (Johnson, 2020a). (A famous example of proxy discrimination is the use of literacy tests as a means of race-based disenfranchisement.) Moreover, making decisions blind to protected features can in fact penalize minority groups. For instance, after controlling for factors like criminal history and age, women tend to reoffend less often than men; this means that gender-neutral risk assessments might overstate the recidivism risk of women and, if used in sentencing decisions, could result in higher incarceration rates overall (Skeem and Lowenkamp, 2016; Corbett-Davies and Goel, 2018).

A more promising measure of treatment parity is 'classification parity', which says that some given classification error (false positive, false negative, AUC, etc.) must be equal across groups defined by protected attributes. For instance, ProPublica's audit of COMPAS used parity of false positive rates (the proportion of individuals wrongly classified as high risk) as its fairness metric, which requires that false positive rates are the same for a protected group as they are for the total population.

The problem with classification parity is that when different groups have different base rates, their calculated risk distributions will necessarily have different means and variances and, correspondingly, different error rates, regardless of which features are used in the calculation.

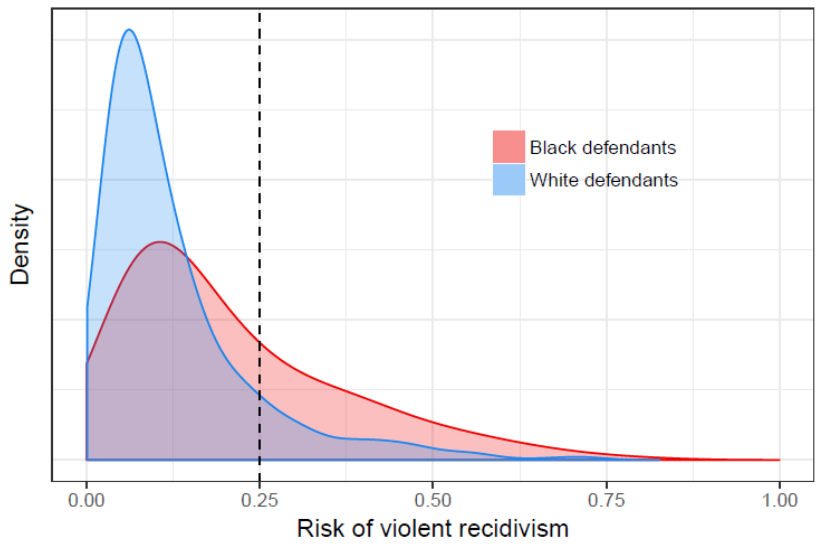


Figure 3: The risk distributions of recidivism between Black defendants and white defendants; the dotted line represents the decision threshold, showing that error rates in classification would be different for the two populations. Figure appears in Corbett-Davies and Goel (2018).

Error rate disparities between groups can be indicators that the shape of each group's risk distribution is different in the training data (see Figure 3). Enforcing classification parity in such cases would result in less accurate predictions, which is costly both for majority and minority groups (Berk et al., 2018). This is part of a broader phenomenon that Brent Mittelstadt et al. (2023) call "levelling down," where fairness is achieved by making the performance worse for every group, or by decreasing the performance of some groups to the level of the worst off.

By contrast, 'calibration' defines algorithmic fairness through disparate impact. It requires that actual outcomes are independent of protected attributes, conditional on a predicted risk score. In other words, calibration is satisfied when a risk score has the same meaning no matter which group an individual belongs to. For instance, a risk assessment instrument is calibrated if, for some recidivism risk score, the proportion of people who would go on to reoffend if released is the same across all protected groups.

Calibration is important if a risk score is to be treated as a meaningful quantification of

risk, as it is the only measure concerned with impact parity (which, recall, means that there is no disparate adverse impact on protected groups) (Barocas and Selbst, 2016). However, much like anti-classification, calibration cannot rule out proxy discrimination. Consider a classic case of redlining. Suppose that Black people and white people have the same rates of loan default within a zip code, but Black applicants tend to live in zip codes with high default rates. A bank's algorithm could discriminate against Black people by rejecting loans from zip codes with high default rates and ignoring all other information. Such a score would be calibrated because white and Black applicants with the same score would default equally often, but the bank could still deny loans to most Black applicants (Corbett-Davies and Goel, 2018). On the other hand, disparate impact can result from differences on the basis of which discrimination is considered perfectly appropriate.[12]

Complicating matters further, impact parity actually requires disparate treatment when base rates between groups differ – even if a system is calibrated correctly, it will still result in unequal treatment for different groups unless those groups are homogeneously represented in the data in the relevant respects. This means that in most cases, an algorithm can satisfy either classification parity or calibration, but not both.[13]

This tension is illustrated well by the response to ProPublica's audit of COMPAS given by Equivant (formerly Northpointe), the company that makes the instrument. Recall that ProPublica's algorithmic bias claim appeals to classification disparity (that Black defendants are more likely to be falsely classified as future criminals). Equivant responded that because COMPAS is calibrated (white defendants and Black defendants with the same score reoffend at similar rates) it is therefore *not* racially biased (Dieterich et al., 2016). ProPublica, in turn, rebutted this rebuttal, arguing that from the perspective of someone who is part of

[12]In the context of recidivism risk, a more general problem with calibration is that risk estimates – whether informally made by judges or algorithmically-derived – influence *actual* risk. For example, if a high risk prediction influences the likelihood that someone is incarcerated, this will influence a person's opportunity and incentive to commit crime – "it is impossible without additional strong assumptions to distinguish the 'true' behavior of individual defendants from the behavior that results from their non-random treatment within the existing system" (Bushway and Smith, 2007). The influences of these prior risk assessments will be present in the data, and if judges have historically estimated risk differently for Black defendants than for white defendants, then conditional accuracy measures for different groups are biased and a flawed metric of fairness (Stevenson, 2018).

[13]See Chouldechova (2017) and Kleinberg et al. (2016) for formal proofs and discussion of this impossibility result.

the group more likely to be wrongly classified, simply sorting Black and white defendants correctly at the same rate is not enough – as the statistician Andrew Gelman puts it, "from the perspective of the sentencer it might be unbiased, but from the perspective of a criminal defendant it could be biased" (Angwin and Larson, 2016).

COMPAS, in short, is biased according to classification parity, but it is less clear whether this means this renders it unacceptable for use. As Corbett-Davies, Goel, and González-Bailón maintain in an op-ed arguing for the use of risk assessment instruments, "It is not biased algorithms but broader societal inequalities that drive the troubling racial differences we see ... throughout the country. It is misleading and counterproductive to blame the algorithm for uncovering real statistical patterns" (Corbett-Davies et al., 2017). Others argue that classification disparities are indicators that "something is likely amiss" (Hellman, 2019, 8) and that this unfairness will likely be compounded by the algorithm. Bernard Harcourt argues that prior criminal history has essentially become a proxy for race, meaning that risk assessment instruments are likely to produce a "ratchet effect" that will exacerbate racial disparities in the criminal legal system, with downstream effects on disparities in social outcomes such as employment and education (Harcourt, 2015, 2007).

Deborah Hellman points out that calibration and classification parity are "geared to different tasks" (Hellman, 2019, 10). Calibration informs beliefs about the meaning of an algorithm's classifications, while classification parity informs actions on the basis of those classifications. Hellman argues that inductive risk – the relative cost of different kinds of errors – is what guides our actions. Insofar as we might care more about incarcerating a person who would not have gone on to reoffend than about releasing a person who would have gone on to reoffend, differences in prediction errors between different racial groups turn out to be especially important for whether and how information from the algorithms is used in decision-making.

In short, much like other mechanical procedures, each of these measurements responds only to certain inputs from the environment and thus provides a partial perspectives on fairness (Giere, 2006; Mitchell, 2009). In response to these limitations, other researchers have countered that the absence of a problematic causal mechanism might be necessary to ensure fairness. I turn briefly to this other class of fairness measures.

### 1.3.2 Causal Measures of Fairness

Building on Pearl (2009)'s counterfactual causal framework, causal measures of fairness rest on the intuition that protected attributes should not affect predictions unless they come from 'acceptable' causal pathways (Kusner et al., 2017).[14] Appealing to David Lewis, Kusner et al. (2017) suggest that a decision is fair toward an individual if it is no different between our actual world and a counterfactual world in which the individual belongs to a different protected group. This means that changing a protected attribute, such as race, while holding anything not causally dependent on that attribute fixed should not change the outcome, regardless of the protected attribute's predictive power. This is called 'counterfactual fairness'.

Counterfactual fairness is intended to capture the sentiment expressed in the legal definition of employment discrimination:

> The central question in any employment-discrimination case is whether the employer would have taken the same action had the employee been of a different race (age, sex, religion, national origin etc.) and everything else had been the same (7th Circuit Court, 1996; Pearl, 2009).

Causal fairness measures like this one are more epistemically demanding than statistical fairness definitions. There are at least three types of discrimination that statistical fairness definitions are blind to but that can be captured using counterfactual reasoning: direct discrimination ($X \rightarrow Y$), indirect discrimination ($X \rightarrow W \rightarrow Y$), and spurious discrimination (e.g., via a common cause, $X \leftarrow Z \rightarrow Y$) (Zhang and Bareinboim, 2018).

As an illustration of what a counterfactual fairness assessment might look like, consider a scenario in which a car insurance company assigns insurance prices based on an individual's accident rate (Figure 4). Some unobserved factor $U$ (like aggression) causes drivers to be more likely to have an accident and also causes them to prefer driving red cars $X$. Suppose that individuals of some race $A$ are more likely to drive red cars but are no more likely to get into accidents than other individuals. Using car color to predict the accident rate $Y$ is unfair because it charges individuals of a certain race higher prices, even though race does not cause driving behavior. This will not be captured by statistical measures of fairness.

[14] Several other authors have taken similar approaches, notably Kilbertus et al. (2017), Nabi and Shpitser (2018), and Zhang and Bareinboim (2018).

Counterfactual fairness does capture this intuition because it shows that intervening on race would affect whether red cars are preferred ($X$) but not accident rates ($Y$).


Figure 4: $A$ represents race, $X$ represents driving a red car, $U$ represents aggressive driving, and $Y$ is the accident rate. In the right graph, $A$ is intervened on, and $Y$ is unchanged (adapted from Kusner et al., 2017).

Counterfactual fairness can also be used to identify cases in which causal pathways do exist between protected variables and outcomes, but that we may wish to remove from our decision-making procedure. In cases where a protected feature and an outcome are associated due to "a world that punishes individuals in a way that is out of their control," (5) Kusner et al. echo the finding that treatment parity and impact parity cannot always be reconciled. Counterfactual fairness suggests a reason for this: "this is the result of $A$ [race] being a cause of $Y$ [an outcome such as recidivism risk]" in the algorithm (Kusner et al., 2017, 6). Thus predictive instruments should strive not to use $Y$ (recidivism risk) as the basis for decision making, but rather some $\hat{Y}$ that estimates another predictor that is "closest" (6) to $Y$ but independent of the protected variable $A$ (race).

But causal fairness approaches, too, are partial and require value-laden choices. In particular, causal models make assumptions that limit them to representing only certain cases of discrimination and require choices about variable inclusion and 'disallowed' causal paths that presuppose normative judgments about discrimination. For brevity, and because the audit that is the focus of the following section does not use causal fairness definitions, I will not elaborate on these problems here. Interested readers may find more detail about the value-ladenness of counterfactual fairness in Appendix A. I will only note here that enforcing causal notions of fairness may introduce new and undesirable algorithmic biases, such as hiring algorithms that treat all job candidates equally regardless of their qualifications

(Nilforoshan et al., 2022).

### 1.3.3 Calls for Trained Judgment Over Meta-Mechanical Objectivity

In imposing meta-mechanical objectivity on algorithms, fairness definitions inherit the limitations of mechanical objectivity that prompted the metrics' use in the first place. Formal fairness definitions capture only a partial sense of fairness and require value-laden choices. Altering algorithms to conform with these metrics, whether by pre-processing training data or adding fairness metrics as an additional optimization constraint, addresses existing biases by introducing new ones (Danks and London, 2017). Trade-offs between fairness metrics mean that these mitigation strategies can result in social costs, such as decreases in predictive accuracy for some groups (Mittelstadt et al., 2023).

More broadly, the last several years have seen mounting criticism of algorithmic fairness, from arguments that it neglects legal subtleties and interpretation in anti-discrimination definitions (Wachter et al., 2021) to calls for a more substantive notion of fairness that includes its objectives and broader sociotechnical context (Barabas, 2019; Barabas et al., 2020; Green and Viljoen, 2020; Green, 2022). Resolving algorithmic bias, in other words, may require resolving structural social bias (Johnson, 2020a; Hellman, 2019).[15] In practice, audits that rely on formal fairness definitions overlook these issues. Worse, the apparent value-neutrality and universality of fairness definitions – their perceived meta-level objectivity – can trickle down to the perceptions of objectivity of algorithmic methods that satisfy them. This can serve as a stamp of legitimacy for algorithms that have serious shortcomings in practice; Aïvodji et al. (2019) call this phenomenon 'fairwashing'. To illustrate these issues, I turn now to a recent example of these fairness definitions in action.

[15]This reflects a classic tension in anti-discrimination doctrine between anti-classification and anti-subordination; the latter holds that guarantees of fairness "cannot be realized under conditions of pervasive social stratification and argue that law should reform institutions and practices that enforce the secondary social status of historically oppressed groups" (Balkin and Siegel, 2003).

## 1.4 Case Study: CMU's Technical Audit

Prior to its statewide deployment in 2020, the Pennsylvania Sentence Risk Assessment Instrument (SRAI) was evaluated by a team of master's students from CMU's Heinz College, by request from the Pennsylvania Sentencing Commission. Like other advocates of risk assessment instruments, the audit situates the motivation for the use of the SRAI in terms of mechanical objectivity: "Since someone with the same circumstances and offenses will be treated the same by an algorithm, it removes the problems of human decision making and bias" (Becerril et al., 2019, 9). The review has two main components: replication and assessment of the mechanical procedure. Based on this analysis, the audit makes several recommendations, most notably a recommendation to increase the high-risk category cutoff by two points.

The team started by replicating the logistic regression used for variable selection and re-creating both the general SRAI and the Crime Against a Person SRAI (the latter estimates risk of violent recidivism for crimes such as murder, rape, and assault).[16] The selected factors in the final version of the tool are age, gender, number of prior convictions, prior conviction offense type, current conviction offense type, multiple current convictions, and prior juvenile adjudication (Pennsylvania Commission on Sentencing, 2020). Variables are given weights, depending on their degree of association with the outcome variable. Young age, for example, is strongly associated with recidivism; individuals under 21 receive 5 points and those over 49 receive 0 points (Figure 5). These points are summed to make the risk score, which ranges from 0 to 18 and, at the time of CMU's analysis, were binned into low (0–4 points), typical (5–9 points), and high (10–18 points), categories that corresponded to one standard deviation above and below the mean risk score.

The audit notes that the training dataset is imbalanced in several senses. Most individuals do not go on to reoffend; roughly 8,000 individuals recidivate, while roughly 15,000 do not. Because of this, the overall error rate for the high risk category (the proportion of incorrect high risk predictions) is much higher than for the error rate for the low risk category. The

[16]The Pennsylvania Sentencing Commission selected the factors that were to be used by the SRAI through the unweighted Burgess method, a bivariate analysis between each factor and recidivism.

| Risk Factors | | Risk Score |
|---|---|---|
| Gender | Male | 1 |
| | Female | 0 |
| Age | <21 | 5 |
| | 21-25 | 4 |
| | 26-29 | 3 |
| | 30-39 | 2 |
| | 40-49 | 1 |
| | >49 | 0 |
| Current Conviction Offense Type | Murder | 1 |
| | Person-Felony | 1 |
| | Person-Misd. | 1 |
| | Sex-Felony | 0 |
| | Sex-Misd. | 0 |
| | Burglary | 2 |
| | Property-Felony | 2 |
| | Property-Misd. | 2 |
| | Drug-Felony | 1 |
| | Drug-Misd. | 1 |
| | Public Admin. | 1 |
| | Public Order | 1 |
| | Firearms | 2 |
| | Other Weapons | 2 |
| | Other | 1 |
| Number of Prior Convictions | None | 0 |
| | 1 | 1 |
| | 2-3 | 2 |
| | 4-5 | 3 |
| | >5 | 4 |
| Prior Conviction Offense Type | Person/Sex | 0 |
| | Property | 1 |
| | Drug | 1 |
| | Public Order | 1 |
| | Public Admin. | 1 |
| | DUI | 0 |
| | Firearm/Weapon | -1 |
| Multiple Current Convictions | Yes | 1 |
| | No | 0 |
| Prior Juvenile Adjudication | Yes | 1 |
| | No | 0 |
| Scale | | 0 to 18 |

Figure 5: "Recidivism Risk Scales" (Pennsylvania Commission on Sentencing, 2020, 13).

defendant dataset is also mostly white (60%), young (median age of 28), and male (80%).

The team assessed both the performance and the fairness of the instrument. Along the AUC metric, which shows how good a model is at distinguishing between classes, the SRAI performed "moderately well" (an AUC of 0.66, meaning that the SRAI could distinguish high-risk and low-risk defendants 66% of the time).[17] The Crime Against a Person SRAI performed "very poorly" on most of the metrics used because of the very low likelihood of violent recidivism, and the team recommended not deploying the it at all, a recommendation the Commission followed.

In evaluating the fairness of the general SRAI, the CMU team used five fairness metrics. Two are basic measures: 'overall accuracy' and 'demographic parity', or the overall proportion of positive or negative predictions. The other three measures are statistical measures of algorithmic fairness:

- Two versions of classification parity, including 'conditional procedure accuracy' (true positive and true negative rates, i.e., sensitivity and specificity) and 'treatment equality' (ratio of false positives to false negatives);
- Calibration, which they call 'conditional use accuracy' (the proportion of positive predictions that are true positives and the proportion of true negative predictions that are true negatives).

These fairness metrics were calculated for race and gender groups. "If there is no difference between the fairness metrics of a subpopulation, by gender or race for example," the review states, "then total fairness has been achieved" (10).

To start, the team calculated the recidivism rate for each group and found that the Black population had the highest recidivism rate (36.4%), followed by the white population (31.5%) and the Hispanic population (29.0%). Men also had a higher recidivism rate (34.6%) than women (26.3%). Recall from subsection 1.3.1 that classification parity is typically not possible when base rates differ between groups, as they do in this case. In keeping with this, the audit found disparities for both race and gender on both basic measures (overall accuracy,

[17]This qualitative label comes from the performance of the SRAI relative to similar risk assessment instruments, which have AUCs that range from 0.61 to 0.67. However, the rule of thumb given in a widely cited textbook on applied logistic regression places the SRAI's AUC of 0.66 within the "poor discrimination" category (an AUC of 0.5 is as good as a coin flip, or "no discrimination") (Hosmer Jr. et al., 2013).

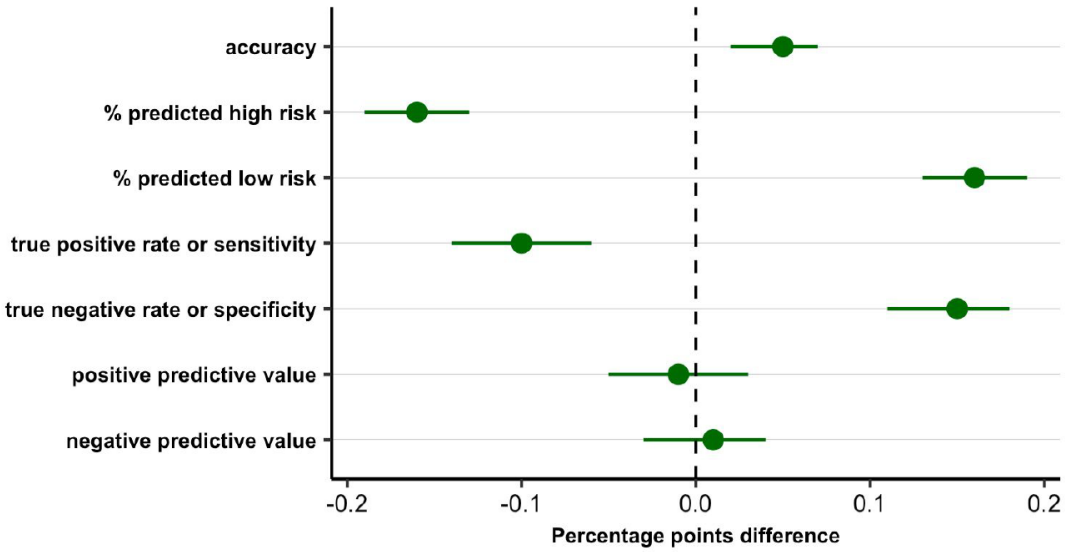


Figure 6: "Fairness Metrics Differences Between White and Black Offenders." Shows each metric's value for the white population minus the metric's value for the Black population, with 95% confidence intervals based on a bootstrapping method the team used on the validation dataset. Note that the chart does not include the difference in treatment equality between white and Black defendants, which is −1.89 (Becerril et al., 2019, 40).

proportion of positive and negative predictions) and both markers of classification parity (sensitivity and specificity, ratio of false positives to false negatives), but found that the tool was calibrated because the positive/negative predictive values were the same between groups. The differences in the values of each metric between white and Black defendants are shown in Figure 6.

The team found that accuracy was 5 percentage points higher for white defendants relative to Black defendants and 12 percentage points higher for females relative to males. The SRAI identifies high-risk Black defendants more accurately than high-risk white defendants, and identifies low-risk white defendants more accurately than low-risk Black defendants. Conversely, the SRAI identifies high-risk males more accurately than high-risk females and identifies low-risk females more accurately than low-risk males. "It seems that the general SRAI is biased as more white offenders who do not recidivate are classified as low risk in comparison to the Black offenders," the audit states, but adds that "this result is a direct

consequence of having different recidivism rates across groups and an instrument that does not perfectly separate both classes (recidivate vs not recidivate)" (36). They also found that, "in relative terms, there are more false positives than false negatives for Black offenders compared to whites" (36) and "more false positives than false negatives for male offenders than for female offenders" (38). The positive and negative predictive values, however, are equal between white and Black defendants and between male and female defendants; the probability that any high risk defendant, regardless of group membership, will go on to reoffend is 53%, and the probability that a low risk defendant, regardless of group membership, will not go on to reoffend is 83%. In other words, the tool is calibrated but has various race and gender classification disparities.

However, the graphic in Figure 6 is misleading for two reasons. First, it excludes the ratio of false positives to false negatives ('treatment equality') because it is so much higher than the other values that including it would have made the scale on the graph impractically large.[18] The only thing the authors mention about the false positive:false negative ratio is that it is higher for Black defendants, without emphasizing the magnitude of the difference or its significance. A clearer way of highlighting the issue is to compare the false positive rate and false negative rate for white relative to Black defendants and male relative to female defendants. I found that Black defendants are 1.5 times as likely as white defendants to be falsely classified as recidivating, and white defendants are 1.6 times as likely as Black defendants to be falsely classified as non-recidivating, a finding similar to ProPublica's audit (Angwin et al., 2016). Males are also over 3 times as likely as females to be falsely classified as recidivating, and females are 3 times as likely as males to be falsely classified as non-recidivating.[19] Putting these error disparities in relative terms is more informative and better illustrates the scale of the classification disparity.

Second, during my own replication of the audit's figures, I encountered some troubling inconsistencies. When I attempted to reproduce each fairness metric based on the confusion

[18]The difference in treatment equality between white and Black defendants is −1.89 (2.44 for white and 4.33 for Black). This figure is not even mentioned in the main text; I was able to calculate it based on charts in the appendix of the paper.

[19]I calculated the false positive rate FP/(FP+TN) for white defendants (0.32) and Black defendants (0.46); the ratio Black:white is 1.46. Next I calculated the false negative rate FN/(FN+TP) for white defendants (0.27) and Black defendants (0.17); the ratio white:Black is 1.61. I used the FP and FN that the audit appears to have used, despite the labeling inconsistencies (see Figure 7).

matrices provided, my figures differed substantially from what was reported. However, I found that I was able to reproduce their numbers by swapping the false positive and false negative rates in each confusion matrix (for both race and gender). This means that either the 'Prediction' and 'Truth' labels in each confusion matrix are incorrectly labeled, or the team mistakenly used the false positive rate instead of the false negative rate (and vice versa) in their analysis, which would invalidate their findings. Fortunately, I believe it is the former. The confusion matrices are consistently analyzed in this manner, and the total recidivism incidence reported in the text approximately matches the sum of the 'Prediction' 'Yes' column but not the 'Truth' 'Yes' row, which suggests that 'Prediction' and 'Truth' are incorrectly labeled (Figure 8).

| White and Others | | |
|---|---|---|
| | Prediction | |
| Truth | No | Yes |
| No | 6,775 | 1,290 |
| Yes | 3,154 | 3,463 |

| Black | | |
|---|---|---|
| | Prediction | |
| Truth | No | Yes |
| No | 2,258 | 451 |
| Yes | 1,954 | 2,219 |

| | | White and Others | | | Black | | |
|---|---|---|---|---|---|---|---|
| | | My analysis | | Original | My analysis | | Original |
| Metrics | | Standard | FP <-> FN | | Standard | FP <-> FN | |
| Accuracy | (TP+TN)/(TP+TN+FN+FP) | 0.7 | 0.7 | 0.7 | 0.65 | 0.65 | 0.65 |
| Sensitivity | TP/(TP+FN) | 0.52 | 0.73 | 0.73 | 0.53 | 0.83 | 0.83 |
| Specificity | TN/(TN+FP) | 0.84 | 0.68 | 0.68 | 0.83 | 0.54 | 0.54 |
| Positive predictive value | TP/(TP+FP) | 0.73 | 0.52 | 0.52 | 0.83 | 0.53 | 0.53 |
| Negative predictive value | TN/(FN+TN) | 0.68 | 0.84 | 0.84 | 0.54 | 0.83 | 0.83 |
| Treatment equality | FP/FN | 0.41 | 2.44 | 2.44 | 0.23 | 4.33 | 4.33 |

Figure 7: Based on the confusion matrices by race (top two boxes; Becerril et al., 2019, 35), I calculated each fairness metric with the false positive (FP) and false negative (FN) (values in green), and then calculated the same metrics with the FP and FN swapped (FP<->FN) (values in red). The audit's original analysis (also in red) matches the calculations with the swapped FP and FN.

| General | | |
|---|---|---|
| | Prediction | |
| Truth | No | Yes |
| No | 9,660 | 1,842 |
| Yes | 5,501 | 6,020 |

Figure 8: "Confusion Matrices of Recidivism" (Becerril et al., 2019, 40). Based on the text, approximately 15,000 people did not reoffend and 8,000 did (27). The 'Truth' rows 'No' and 'Yes' do not sum to these values, but the 'Prediction' columns 'No' and 'Yes' do, which suggests that 'Prediction' and 'Truth' are incorrectly labeled.

### 1.4.1 Recommendations and Projections

Based on predictive performance, the SRAI "falls within industry standards" (52). Based on the fairness assessment, the audit makes several recommendations to improve the tool.

The first is to not use the high risk category in the SRAI, due to the high error rate (48%) in classifying high-risk defendants, as well as race and gender disparities. "Due to the different base rates in recidivism between each subpopulation," they write, "it is practically unattainable to achieve both a high level of accuracy and fairness" (53). Instead, they recommend only using the SRAI to identify low-risk defendants. The Commission chose not to follow this recommendation, responding that it is more important to consider the accuracy of the high risk category within the actual outcome: out of people that did recidivate, the instrument correctly predicts 77% as high risk. However, the Commission did agree to follow the audit's second recommendation not to use the Crime Against a Person SRAI, which I did not discuss in my analysis above.

The third recommendation concerns the cutoffs for the risk categories. The audit found that accuracy improved when the high-risk cutoff was increased from 10 points to 12 points: fewer defendants were included in the high risk category (5%, down from 16%), and those labeled high risk were more likely to reoffend. Increasing the low risk cutoff, which would increase the number of defendants labeled low risk, reduced the classification disparity between

Black and white defendants; the audit does not comment on how this change would affect relative false positive and false negative rates. The Commission agreed to this change, though they noted their preference for avoiding the "political/unsystematic process of picking and choosing cut points that consider the ratio of false positives to false negatives" (Pennsylvania Commission on Sentencing, 2019b).

The audit also assesses what the instrument's performance would be if it were to stop using gender as a predictive factor, "in case of legislative action that mandated its [gender variable] removal" (Becerril et al., 2019, 49). They found that removing gender increased the accuracy of the instrument, improved sensitivity, and decreased specificity. Fairness metrics between male and female defendants, as well as between white and Black defendants, stayed the same or improved, and male defendant scores decreased by a point while female defendant scores stayed the same. They thus recommended removing gender as a predictive variable, but the Commission did not follow this recommendation on the basis that excluding gender increased how many females are classified as high risk.

Finally, the audit makes a projection about the impacts of the tool on judges' rates of order presentence investigation report (PSI) in the state. A PSI contains additional information about the defendant, including an interview with the defendant conducted by a probation officer and, in some counties, an additional 3$^{rd}$ generation risk and needs responsivity assessment. The SRAI recommends that judges order a PSI for low- and high-risk defendants, with the idea that atypically high- or low-risk defendants could be candidates for alternative sentences, such as community supervision. PSI-ordering rates in Pennsylvania vary substantially county-to-county, as do the contents of the reports; one of the audit's projected outcomes of the tool's adoption was thus the minimization of county-level disparities in how often, and for which kinds of defendants, judges choose to order a PSI (Figure 9). I discuss this projection in more empirical detail in chapter 4, showing why it is unfounded.

### 1.4.2 The Limits of Algorithmic Fairness

As the authors of the audit acknowledge, their analysis inherits the shortcomings of the algorithmic fairness definitions it uses, as well as the mathematical impossibility of satisfying

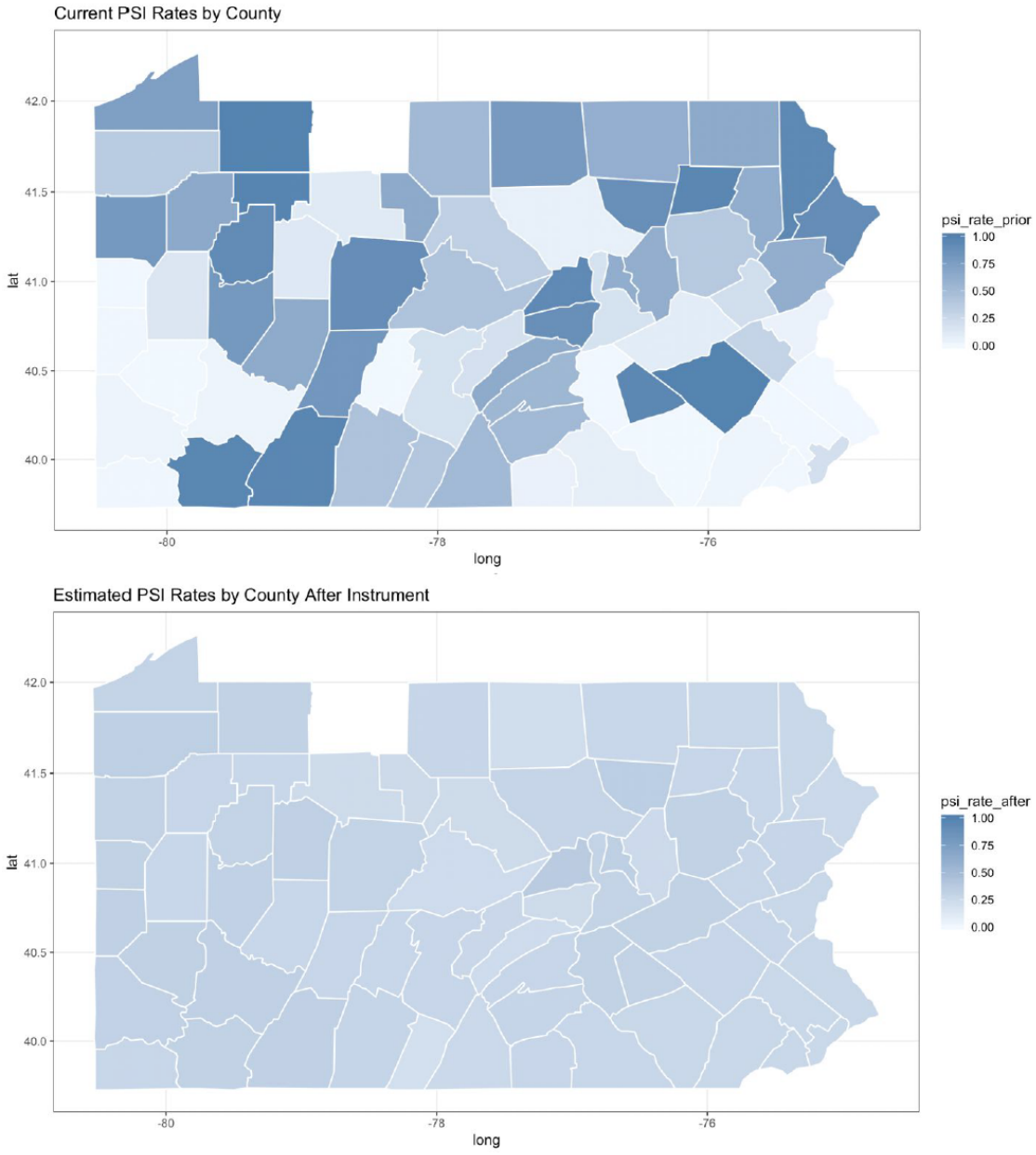


Figure 9: "Comparison of PSI Rates Before and After the Instrument," which states that "if PSIs were to completed [sic] following the rate at which the instrument identifies high- or low-risk offenders, the PSI rates across counties will be more consistent" (Becerril et al., 2019, 24).

classification parity and calibration when base rates differ. The authors recognize the challenge of their task, writing that "Evaluating risk assessment instruments can be difficult because there is no standard for 'acceptable' values of accuracy, fairness, and performance" (52). The audit instead resorts to a kind of balancing act, in which different changes are tested in their impacts on each fairness and accuracy metric. Like the artifacts scientists noted in their use of mechanical image production techniques in the 19$^{\text{th}}$ century, each of the changes the auditors test introduces artifacts of its own: it improves algorithmic bias on one standard while exacerbating algorithmic bias on other standards.

The meta-mechanical objectivity of formal fairness metrics, intended to minimize algorithmic bias, thus proceeds through the introduction of additional value-laden decisions and biases – not so mechanical, perhaps. Following mechanical rules is complex and requires discretion and interpretation. Moreover, the audit's attempts to make whatever improvements they had in their power to make – modest improvements to one fairness metric or another – were blocked in practice by the Commission's reticence to implement most of the changes, and, as I will show at length in chapter 4, the reticence of judges to use the SRAI in the ways the auditors suppose in their charmingly optimistic Figure 9. The scientists in Daston and Galison's historical narrative ultimately overcame the shortcomings of mechanical objectivity through trained judgment, a positive exemplar of how humans can use mechanical tools to improve their decision-making. In the case of risk assessment instruments, judges apply their own trained judgment, though it is not trained to use or interpret the instruments, per se.[20] The situation on the ground is thus in some respects an unhappy marriage between truth-to-nature and mechanical objectivity: the losing combination of biased human decisions and biased mechanical procedures (recall the Kentucky pretrial case, in which race disparities were exacerbated because of inconsistent adherence to the risk assessment tool (Albright, 2019)).

Regardless of the researchers' aims, their audit played a strategic role for the tool's

[20]Megan Stevenson (2018) writes, "risk assessment tools may not be used as designed: they may be ignored or used off-label to accomplish something other than what was intended. Judges may not understand exactly what the risk score is measuring, or what level of statistical risk is associated with each risk category. The tool may be good at predicting misconduct, but the interventions taken to ameliorate risk may actually exacerbate it. The pressures of re-election or re-appointment may impact how and when the risk tool is used" (306).

ultimate adoption. Once it could be said to have been found to conform to "industry standards," including alarming error disparities between race and gender groups and coin-flip-level accuracy for high risk predictions, the instrument could move forward to the approval stage. As Porter argues, the mechanical objectivity derived from numbers confers epistemic authority "even when nobody defends their validity with real conviction" (8). Much like the suspicion of the judge was to be assuaged through an algorithmic bureaucrat's mechanical objectivity, suspicion of the algorithmic bureaucrat was to be assuaged through formal fairness metrics' meta-mechanical objectivity. Following the external review's completion in 2019, the Commission voted to adopt the tool, and the Sentence Risk Assessment Instrument was formally rolled out in July 2020.

## 1.5 A Broader Picture of Values in Algorithms

This chapter started with the relationship between algorithms and mechanical objectivity and provided a survey of algorithmic bias and the project of measuring it. I argued that formal fairness rules are a means of achieving meta-mechanical objectivity, where the bias to be removed is algorithmic rather than human. I illustrated the shortcomings and value-ladenness of formal fairness definitions and demonstrated these issues in action through an analysis of CMU's fairness audit of the Sentence Risk Assessment Instrument. I showed that the mechanical objectivity of risk assessment instruments and meta-mechanical objectivity of formal fairness definitions emerge as a response to suspicion toward human and algorithmic decision-makers, respectively, and that both forms of objectivity serve as a source of authority even when they introduce their own value-laden decisions and shortcomings.

The limits of the project of quantitatively measuring and removing algorithmic bias become even more apparent once algorithmic bias is conceived of more broadly than biased prediction, as Fazelpour and Danks do, and once algorithmic fairness is considered more substantively, as many others have proposed (Barabas, 2019; Fazelpour and Lipton, 2020; Green and Viljoen, 2020; Green, 2022; Mittelstadt et al., 2023). In the subsequent chapters, I use the case study of crime prediction to illustrate how algorithms can presuppose and

influence values in ways that go beyond biased algorithmic predictions: the normative jurisprudential and moral positions required by and reinforced by criminal risk assessment, which de-prioritize structural interventions (chapter 2); the value-laden formalization of crime evident through variable choices in Soviet cybernetic models of crime (chapter 3); and the values introduced by judge-AI-interaction in the context of the Sentence Risk Assessment Instrument (chapter 4).

## 2.0 Domain Distortion: How Predictive Algorithms Warp the Law

In the discourse on evidence-based sentencing, a movement that advocates grounding sentencing decisions in scientific and empirical methods, recidivism risk assessment instruments have taken on central importance (Monahan and Skeem, 2016). Risk assessment instruments, which estimate an individual's risk of rearrest or reconviction for a future crime, are often presented as a 'progressive' reform – a way to reduce mass incarceration, reduce bias in judgment and sentencing, reform cash bail, and make sentencing "objective," "smart," and "evidence-based" (Hannah-Moffat, 2013; Estelle and Phillips, 2018; Starr, 2014).

The objectivity associated with algorithms is subject to familiar critiques of the value-free ideal in science, the idea that scientific reasoning should strive to be free of non-epistemic values (Douglas, 2009). Much like other scientific methods, algorithmic decision-making contends with non-epistemic values introduced by dealing with inductive and epistemic risk.[1] In chapter 1, I also discussed the limits of the mechanical objectivity derived from the use of algorithmic risk assessment. There is overwhelming evidence that algorithms can perpetuate and exacerbate the biases that plague human judgment – harmful social values can get 'baked in' (Danks and London, 2017; Fazelpour and Danks, 2021). To date, much of the concern about the value-ladenness of risk assessment instruments has centered around their algorithmic bias and which fairness metrics the algorithms need to satisfy.

This focus tacitly assumes the following conditional: if risk assessment instruments can be made free from harmful social values, they should be adopted in criminal sentencing. In other words, as long as algorithms come as close as possible to satisfying the value-free ideal, their use is preferable to biased judgment. Among other problems, this perspective neglects three problematic value commitments of risk assessment instruments, which illustrate another avenue by which algorithms can be value-laden: by influencing (and being influenced by) the concepts, assumptions, and normative aims that are taken for granted in its context of application. This relationship between dominant social values and the choice to approach

[1] I follow Biddle and Kukla (2017) in using the term 'epistemic risk' to refer to the risk of error at any stage of knowledge production, which includes inductive risk.

questions through particular methodologies is bidirectional and mutually reinforcing, subtly normalizing the social values that led to the adoption of the methodology in the first place. I call this phenomenon *domain distortion* because the social values in an algorithm's domain of application become distorted through its use.

First, insofar as risk assessment instruments are intended to remove judge discretion and produce consistent sentencing results, their application presupposes a formalist interpretation of legal principles – namely, that laws have one correct, mechanically discoverable meaning. Formalism, sometimes disparagingly referred to as 'mechanical jurisprudence', sustained heavy criticism from 20$^{\text{th}}$ century legal realists; it is rejected by many contemporary legal scholars for failing to capture, descriptively, what judges actually do and, normatively, what judges ought to do. It is, in essence, the value-free ideal of the legal world. Risk assessment instruments distort the domain of criminal sentencing by reifying a widely disparaged jurisprudential presupposition and neglecting the essential interpretive component of judging. In practice, risk assessments are selectively considered by judges to *augment* judgment, sometimes amplifying existing racial biases in human judgment.

Second, the use of risk assessment instruments blurs the line between the domain of liability assessment (choosing a verdict) and the domain of sentencing (given a verdict, choosing a punishment). Jurisprudence – the philosophy of law – has traditionally been concerned with the former domain, while the latter is up to the personal discretion of judges. Risk assessment instruments explicitly take future liability assessments into consideration when deciding sentences for current liability assessments, which I argue effectively dissolves the separation between these domains. One consequence of this blurring of domains concerns the implicit purpose of criminal sentences: deciding criminal sentences based on predictive features that have nothing to do with prior criminal conduct, such as demographic and socioeconomic information, presupposes that the purpose of punishment is consequentialist (crime control) rather than deontological (retribution).[2] My aim here is not to advocate for either of these positions, but rather to point out that, in blurring the domains of liability assessment and sentencing, the use of risk assessment algorithms in sentencing means an implicit normative commitment to a consequentialist view of sentencing.

[2]Monahan and Skeem (2016) have also pointed out this issue.

Third, risk assessment as a project is compatible with a narrow set of interventions. In particular, instruments that predict recidivism risk are compatible with interventions that differentially target individuals at different parts of a probability distribution of recidivism, rather than structural interventions intended to shift the mean downward or target the causes of the distribution. This is characteristic of the dominant penal school of thought since the 1980s – the "new penology" (Feeley and Simon, 1992) – which emphasizes the criminal legal system's role of controlling and surveilling risky populations. Recidivism risk assessment instruments support the social values of efficiency and control of groups and are at odds with abolitionist social values.

I begin by providing context on debates in values in science; I discuss existing work on values in risk assessment instruments and highlight work that recognizes the role of values as causes and effects of scientific methodologies, including work by Hugh Lacey, Langdon Winner, and Elizabeth Anderson. Given this background, I introduce the concept of domain distortion, in which a scientific methodology not only presupposes certain value-laden assumptions or aims but in turn reifies them. Using the example of risk assessment instruments, I defend, in turn, the claims that their use in sentencing (1) presupposes formalist reasoning, (2) blurs the line between liability assessment and sentencing, and (3) privileges individual-level over structural penal interventions. These are routes by which algorithmic decision-making distorts how we reason about their domain of application, introducing value in a deeper sense than mere epistemic risk.

## 2.1 The Battle Over the Value-Free Ideal

In chapter 1, I discussed the evolution of scientific objectivity given by Daston and Galison (2007). Here, I address a view of scientific objectivity conceptually related to mechanical objectivity, one that has been the focus of much debate in philosophy of science: the position that scientists should favor and confirm hypotheses only on the basis of scientific evidence and facts, without the influence of moral or political values.

Broadly, values are things considered worthy of pursuit (Elliott, 2017, 11). An epistemic

value is one that is considered to be conducive to knowledge production. Accuracy, scope, and consistency with existing knowledge of the world are examples of epistemic (also known as cognitive) values (Kuhn, 1977). By contrast, non-epistemic values are not necessarily knowledge conducive; these include political and ethical values, which can be held dogmatically, such as religious faith, or can be responsive to empirical evidence (Anderson, 2004).

Proponents of value-free science argue that non-epistemic values should be irrelevant to evaluating the goodness of a scientific claim. This position is known as the value-free ideal: science can or should strive to be free from non-epistemic value judgments. Scientific theories should instead be assessed only according to epistemic values, such as empirical adequacy, simplicity, and so on (Lacey, 1999; Betz, 2013).

It is widely accepted that values may play a legitimate role in the "context of discovery" – that is, in selecting what should be investigated and which hypotheses should be tested. It is also acceptable for non-epistemic values to be used to constrain scientific procedures, such as ethical treatment of human subjects, but this is strictly for the sake of promoting an ethical value and is not knowledge-conducive. Finally, non-epistemic values can play an uncontroversial role in determining how much certainty is needed for a scientific claim prior to using it as a basis for action (Anderson, 2004).

But proponents of the value-free ideal argue that values must be excluded from the "context of justification" – that is, in the evaluation of hypotheses. As Lacey (1999) puts it, "science and values only touch; they do not interpenetrate" (1). It is agreed by proponents and opponents of the value-free ideal alike that scientists should not engage in wishful thinking – that is, they should not overlook evidence just because it does not conform with their values or use values to achieve a predetermined conclusion (Anderson, 2004; Douglas, 2009; Elliott, 2017). One of the archetypal examples of values playing this kind of illegitimate role in science is Joseph Stalin's promotion of Lysenkoism, the theory that environmental alterations made to agricultural crops can be passed down to their offspring. Despite its poor empirical adequacy, Lysenkoism fit well with the party's Marxist philosophy, which emphasizes the importance of acquired traits over inherited ones; geneticists, on the other hand, were condemned and imprisoned for engaging in Western pseudoscience (Elliott, 2017). The question, then, is whether non-epistemic value judgments can legitimately play a more

intimate role in scientific inquiry without leading us down the Lysenkoism path.

There are two standard arguments against the value-free ideal. The first is the underdetermination argument. Given some scientific evidence, we can inductively infer many different conclusions – so theory is underdetermined by evidence. Whether some premises support a given conclusion depends on additional auxiliary assumptions about the world. The Duhem-Quine thesis holds that no hypothesis can be tested in isolation from these assumptions; any time a piece of evidence appears to refute a hypothesis, it could be because one of the assumptions is incorrect, because the hypothesis is wrong, or because the evidence is wrong (Brown, 2013). Thus there is what Helen Longino (1990) calls a 'gap' between theory and evidence, which she and others have argued can and should be bridged with moral or political values, such as preferring the hypothesis that conforms with feminist values or will do the least harm (Brown, 2013; Anderson, 2004; Elliott, 2011; Longino, 2019).

The argument from inductive risk is the second serious challenge to the value-free ideal. Inductive risk refers to the potential consequences of false positives (accepting a false claim) and false negatives (rejecting a true claim). Scientists must make non-epistemic judgments to manage this risk – if the consequences of false positives are considered more important, then a higher degree of confidence should be necessary to accept a hypothesis. In other words, how much evidence is *enough* evidence depends on the value-laden assessments of importance we make for different kinds of errors (Rudner, 1953; Hempel, 1965). Richard Rudner (1953) famously illustrates this point by noting out that our standards for demonstrating that a drug is nontoxic are much higher than our standards for showing that belt buckles are not defective; this difference is due to the "grave" consequences of making a mistake in the former case. Heather Douglas (2000, 2009) later extended this argument to value-laden choices of statistical significance and inductive risk at the 'internal' stages of science: methodological choices, data gathering, and data interpretation; she also argues that inductive risk is particularly pressing for scientific judgments made by experts in policy-relevant contexts. A broader characterization of the argument from inductive risk involves not just the potential harms posed by inductive risk but also potential harms posed by risks of epistemic errors at *any* stage of knowledge production; this is called "epistemic risk" (Biddle and Kukla, 2017).

Opponents of the argument from inductive risk respond that epistemic values alone can be

used to trade off the risks of different errors, or that non-epistemic values need not *necessarily* enter the process (Levi, 1960; Mitchell, 2004). Advocates typically respond that epistemic values alone are insufficient because they do not uniquely determine evidential thresholds and can pull in different directions; weighing epistemic values in a non-arbitrary way thus requires non-epistemic judgments (Winsberg et al., 2014; Douglas, 2017).[3]

Some philosophers have argued that social values are not just a "necessary evil" in science but rather are essential to help science achieve legitimate social and policy goals, so long as the values are stated explicitly and chosen deliberately to represent social and ethical priorities. Kevin Elliott, for instance, argues that policy-relevant scientific research should be guided by values held by relevant community stakeholders. One way to achieve this is through methodologies such as community-based participatory research (CBPR), in which community members play a central role in designing and carrying out research in ways that center community needs and values. "CBPR is a natural outgrowth of the realization that research on chemical pollution and other policy- relevant issues incorporates a host of value judgments. The assumptions that scientists make, the specific questions that they ask, the methods that they employ, the standards of evidence that they demand, and the terms and concepts that they use for communicating their findings can all be influenced by implicit values. Citizens can help bring these values to light and suggest ways of steering research in directions that best fit their own concerns" (Elliott, 2017, 16).

### 2.1.1 Epistemic Risk and Underdetermination in Risk Assessment Instruments

Arguments against the value-free ideal have recently been discussed in the context of recidivism risk assessment instruments by philosophers Gabbrielle Johnson and Justin Biddle. My interest in the value-ladenness of risk assessment instruments in this chapter is substantively different from these arguments, but each rightly points out ways that managing

[3]Ward (2021) argues that opponents of the argument from inductive risk focus on whether values must be motivating reasons behind scientists' choices, whereas proponents of the argument focus on whether values must be justifying reasons – non-epistemic values need not motivate scientists choices here, but "decisions about the acceptance and rejection of hypotheses that run inductive risk cannot be justified without non-epistemic values" (7). Opponents would then need to provide an account on which epistemic values alone can justify choices with practical consequences, or argue that scientists are simply using a conventional threshold for accepting hypotheses.

epistemic risk and underdetermination of theory by evidence introduces value judgments to the creation and use of risk assessment instruments.[4]

Johnson (n.d.) applies the argument from underdetermination and the argument from inductive risk to illustrate the ways in which non-epistemic values are present in recidivism risk assessment instruments. Building on Longino (1995)'s argument that the decision to adopt one set of values over another requires justification that takes into account historical injustices, Johnson argues that assumptions used to bridge the evidence-theory gap by machine learning researchers are not neutral; that is, "If the warrant for our induction is grounded in the uniformity of a pattern in the world, and if the uniformity of that pattern is predicated on oppressive mechanisms of social reproduction, then the warrant for induction is founded on oppression" (10). Johnson also argues that determining how much and which kind of accuracy is enough, which is at the center of algorithmic fairness debates, depends on non-epistemic values in the ways illustrated by Rudner (1953). She adds that a full assessment of inductive risk must take into account not only that the harms of false positives are shared among Black and white defendants, but also the differential harms on Black defendants given that they already bear the brunt of injustices in the criminal legal system (15).

Biddle (2022) takes a broader, epistemic risk perspective on the different ways in which value-laden tradeoffs in human decisions render risk assessment instruments value-laden. He identifies the many decision points at which values enter the development process: the choice of predicting recidivism and the operationalization of the outcome variable of recidivism; the choice of which predictive factors to use and whether to use socioeconomic factors; the choice of risk categories and the relative tolerance for different errors; the choice of fairness criterion; and the tradeoff between transparency or explainability and accuracy. Biddle recommends three possible improvements in how to use recidivism risk assessment instruments, each an increasingly larger departure from the status quo: (1) having judges consider the outputs of multiple additional instruments developed by different entities; (2) making risk assessment instruments transparent and accessible to users and "foster[ing] engagement of relevant stakeholders and publics in the tool-design and implementation

[4]The points made in both of these papers can be helpfully situated as emerging from value-laden decision points in algorithm development described by Fazelpour and Danks (2021), which I illustrate in chapter 1.

process to ensure that instrument design and use reflects the values of affected communities" (17); or (3) to "prohibit the use of recidivism-prediction algorithms that can be shown to disadvantage groups that are already unjustly disadvantaged," in the sense of classification disparity (18). Each recommendation is contestable.

Biddle's first recommendation will only reap the benefits of model robustness if the models underlying different risk assessment instruments are independent in the relevant ways (Weisberg, 2006). In the US, recidivism risk assessment instruments are developed with comparable assumptions and accuracy rates (Desmarais et al., 2018). The main differences between them are how exactly how they operationalize recidivism (re-arrest vs. re-conviction; 2 years vs. 3 years), which variables they use (static or dynamic), and which geographical populations are represented in their training data. Notably, using instruments that are not developed on data in the jurisdiction in which they are applied could decrease, rather than increase, the tools' validity.

Biddle's second point raises an important issue about engaging relevant stakeholders. But legal advocacy organizations, justice reform organizations, and communities impacted by incarceration tend to be highly critical of the premise of recidivism prediction. In Pennsylvania, for instance, transparency was not at issue because the Sentence Risk Assessment Instrument was developed publicly and in a manner responsive to public criticism, but the public was still overwhelmingly opposed to the basic principle of the instrument (Sassaman, 2018; ACLU of Pennsylvania, 2019; Coalition to Abolish Death by Incarceration, 2019). In this case, ensuring that the risk assessment tool "reflects the values of affected communities" (Biddle, 2022, 17) would mean not using it at all and instead implementing a more substantive reform.

To the third point: as Biddle discusses, measures of fairness or group disadvantage are contested and value-laden. Critics of risk assessment instruments argue that the algorithms' classification disparities are disadvantageous to those already unjustly disadvantaged, but advocates say that the tools do not have disparate impact and so do not meet Biddle's criterion (Dieterich et al., 2016; Corbett-Davies et al., 2017).

Biddle briefly discusses the fact that different tools facilitate different punishment goals to different degrees and suggests that democratic deliberation about these goals and their intersections with prediction systems should be encouraged. I believe this is one of the core

Four ways in which values relate to choices.

| Values serve as reasons for making choices. | | Values stand in causal relations with choices. | |
|---|---|---|---|
| **Motivating Reasons** | **Justifying Reasons** | **Causal Effectors** | **Affected Goods** |
| Values motivate an agent to make a choice. | Values justify a choice. | Values causally impact a choice. | Values are promoted or undermined by a choice. |

Figure 10: "Four ways in which values relate to choices" (Ward, 2021, 5). These need not be mutually exclusive.

issues at stake in the use of risk assessment instruments and warrants further attention, as I aim to show in this chapter.

### 2.1.2 Beyond Epistemic Risk: Causal Effectors and Affected Goods

The sense of value-ladenness at issue in algorithmic fairness is typically value-ladenness in the sense of algorithmic bias (chapter 1). As we saw in the previous two sections, philosophers arguing for the presence of values in science, and the value-ladenness of risk assessment instruments specifically, have tended to focus on the values needed to manage inductive risk or epistemic risk more broadly. To illustrate another sense in which risk assessment instruments are value-laden – what I call domain distortion – it is helpful to briefly introduce a taxonomy of values.

Zina Ward (2021) helpfully identifies two categories of ways that values can possibly bear on scientific choices. First, values can serve as *reasons* for making choices, including 'motivating reasons' and 'justifying reasons' (Figure 10). Per the arguments from underdetermination and inductive risk, if one scientific decision is not obviously better than another according to epistemic values, then non-epistemic values can motivate that choice, even if they are not endorsed explicitly or consciously; values can also serve as explicit justifications, such as the preference for a methodological option that has faster computing time or greater social good.

Second, values can stand in *causal relations* with scientific choices. A value is a 'causal effector' if it makes a difference to the outcome. For example, value-laden choices early on in the design of a research study have a causal impact on what kind of evidence is gathered. Robyn Bluhm (2017) argues that methodological assumptions in the design of randomized control trials, such as the comparison of an intervention to an active control rather than a placebo, are relevant to whether and how evidence generated by the trial confirms a scientific hypothesis. Value-laden choices shape which evidence is available to confirm a hypothesis; a value can thus causally influence a choice without motivating it explicitly. How research conceives of an object of inquiry can also depend on the point of view taken on the object of inquiry, which may depend on researchers' personal or professional relations to it (Longino, 1990).

Conversely, values can be 'affected goods' – they can be causally impacted by scientific choices. As Ward rightly notes, "This conception of the relationship between values and choices is pervasive but often hidden, appearing mostly in the background of work on values in science" (4). The central point is that different research approaches and standards of evidence provide more support for some values and less support for others (Elliott, 2017, 41, 99). This may seem like a trivial claim about the consequences of science, but showing that certain methodological choices advance or are only compatible with certain societal values can be surprising and substantive. Hugh Lacey (1999) develops this idea in detail, arguing that choices about 'research strategies' can "interact in mutually reinforcing ways with particular (social and moral) values" (20) and can thus shift society in different directions. In particular, Lacey argues that many current research strategies dominate due to their mutually reinforcing interactions with "modern values of control," particularly the value of exercising control over natural objects (20).

Lacey illustrates this point through his discussion of agricultural research.[5] He argues that since the 20$^{\text{th}}$ century, scientific approaches to improving agriculture have focused on methods from biotechnology, especially the genetic modification of desirable traits, which has often been coupled with research on fertilizers and pesticides with the aim of maximizing crop yields. These approaches have promoted modern values of control and development

[5]Elliott (2017) provides an excellent summary of this case study (45–48).

(189) and fit the interests of agricultural biotechnology companies and research universities; however, this research strategy has tended to neglect its negative effects on small-scale farmers, impoverished rural communities, and the environment. Agroecology is an alternative research strategy that is attentive to local traditions and community needs, strains of seeds that are specific to a geography, and 'natural' pest controls. Lacey argues that agroecology promotes the values of "environmental sustainability, food sovereignty, social justice and democratic participation" (Lacey, 2021) because it aims to grow crops in a way that is ecologically friendly, requires few upfront costs, and benefits the health and economy of rural communities in the Global South.

In the phenomenon I term domain distortion, scientific methods are influenced by (values as causal effectors) and influence (values as affected goods) social values in a manner that reifies and normalizes those social values. Lacey's view and mine diverge in a substantive way: Lacey maintains that science merely interacts or "touches" values at certain points, including the adoption of a research strategy and the application of scientific knowledge, and that epistemic values are the only grounds on which theories and hypotheses are evaluated. In particular, he maintains in his neutrality thesis that scientific findings are (1) "consistent with all value judgments," (2) have "no (cognitive) consequences in the realm of values," and (3) are "evenhandedly applicable regardless of values held" (21).[6]

In contrast to Lacey's thesis, I argue in my own case study in this chapter that using recidivism prediction instruments in sentencing (1) is consistent only with consequentialist value judgments about sentencing, (2) reinforces those values in practice through its compatibility with only certain types of interventions, and (3) requires a value-laden position on legal interpretation. Lacey acknowledges that neutrality "may not be highly manifested in actual fact" because the current conditions of scientific knowledge "may be significantly applicable only in support of certain values," but he maintains that in principle this thesis may be fully met (Lacey, 1999, 18). I am less concerned about describing the state of a hypothetical, idealized version of science. I aim to show how the state of scientific inquiry actually is value-laden in both causal directions, which is particularly noteworthy when these values

[6]Anderson (2004) argues that this thesis is false once one treats 'values' as entities that need not be held dogmatically but are responsive to empirical evidence; scientific theories may also presuppose value judgments, such as "classifying data according to a preferred normative theory" (4) (Bluhm, 2017 makes a similar point).

are entrenched and treated as a given in some domain. My interest is thus more descriptive than normative. We may disagree with the values presupposed and reified by recidivism risk assessment instruments (indeed, I show that many legal scholars have) or argue that values influence risk assessment instruments in unjustified ways, but that is not my primary aim in this chapter. Indeed, as I show in chapter 4, there are other good arguments for opposing the use of risk assessment instruments in criminal legal decision-making.

Scholars in Science and Technology Studies have likewise argued for the bidirectional causal relationship between values and technology in ways that call Lacey's neutrality thesis into question. In his 1980 essay "Do Artifacts Have Politics?" Langdon Winner writes that technological artifacts can reinforce existing power structures because they are "political phenomena in their own right" (123). Winner describes two routes by which technological artifacts can have political characteristics: when they are a means of settling an issue in a community (e.g., Robert Moses' low overpass design blocking marginalized communities from accessing public parks in an affluent white neighborhood in Long Island), and when technology requires or is strongly compatible with certain kinds of political relationships (e.g., a ship's practical necessity for the undemocratic structure of captain and obedient crew). In both scenarios, choices in the design of a technology (or choices in whether a technology should be adopted) not only influence the social world in important and unexpected ways but also affect the physical instantiation of the technology, economic investment into the technology, and social behaviors surrounding the technology (127-128). Once made, these adoption and design choices endure rigidly and can have downstream effects.[7]

Moreover, these consequential decisions need not require specific malicious intentions on the part of developers of technology. As Winner puts it, the technological deck can be "stacked" in advance, ushering in decisions that favor particular (typically dominant) social interests. For instance, disabilities rights activists in the 70s brought to light how public structures like buses, buildings, and sidewalks served to systematically bar people with disabilities from participation in public life. The reason for these discriminatory designs was more plausibly general neglect of non-dominant interests in society – a stacked deck –

[7]For an example from science, see Lenhard and Winsberg (2010) on the entrenchment of design decisions in early weather simulations in contemporary climate models.

rather than intentional marginalizing aims on the part of sidewalk-designers, but the artifacts nevertheless have political characteristics and downstream effects, exacerbating these groups' marginalization.

Elizabeth Anderson (2004)'s case study of research on the impacts of divorce likewise provides an excellent illustration of both values as causal effectors and values as affected goods, and most closely parallels my own illustrations of domain distortion. She argues that feminist research on divorce challenged the dominant orientation toward "traditional family values" present in most scientific research on the impacts of divorce. On the traditional view, divorce separates parental roles from spousal roles and thus breaks down families, which harms children; the evidence for this position is the disparity in measures of well-being in individuals with and without divorce, particularly for negative outcomes like poverty and behavioral problems. But Anderson points out that comparing the well-being of family members with and without divorce is like comparing the effect of hospitalization on health – it requires accounting for pre-existing problems in well-being and health that divorce and hospitalization are a response to, respectively.

Feminist researchers approached the issue of divorce with more ambivalence: perhaps divorce reinforced women's disadvantages by making it easier for men to leave, or perhaps it allowed women to liberate themselves from oppressive marriages. These researchers were also open to seeing non-traditional families as families, rather than as 'broken down'. Instead of focusing on comparisons of quantitative measures of well-being, the feminist study Anderson discusses (Stewart et al., 1997) instead asked how individuals varied over time in the meanings they gave to divorce and its effects. This qualitative data showed that divorce resulted in some transformative and positive effects over time; for example, 70% of women judged that their personalities had improved since divorce, and while previous research had shown that divorce negatively impacts the financial condition of women, Stewart's study found that despite this, women were overall pleased to have more autonomy over the income they did have. This case is an example of domain distortion: negative values about divorce not only causally impacted the results of the original research but also empirically supported negative values about divorce, further reinforcing this normative position.

Similarly to Anderson, what I aim to illustrate in the rest of this chapter is an example

of values causally impacting and being impacted by a methodological choice, namely, the choice to predict recidivism. The values in question are not stated explicitly or easily visible (internally or externally), but I argue that social values of crime control and legal formalism inadvertently get reinforced and normalized through the widespread use and acceptance of risk assessment instruments. The widespread use of risk assessment instruments means that interventions compatible with these values have become the norm, and interventions promoting or presupposing different values have been marginalized. Scientific choices thus influence values, which in turn influence which scientific approaches and interventions are considered legitimate, which in turn entrenches values further. These values become treated as natural or 'correct', especially when they are held by the state and others in dominant social positions. Domain distortion characterizes the phenomenon in which choices of scientific or technological inquiry become distorted in this way: by being influenced, and lending support to, a social value.

In the rest of this chapter, I illustrate this point through the case study of risk assessment instruments. Recall that the working assumption in algorithmic fairness and the evidence-based sentencing movement is that, so long as risk assessment instruments are free from harmful values, they should be adopted in criminal courts to reduce judge bias. A closer look at two jurisprudential problems and one methodological consequence not only calls the value-free ideal of risk assessment instruments into doubt, but also shows that the value-ladenness of risk assessment instruments is deeper than mere biased predictions and epistemic risk. The methodological choice of predicting recidivism is compatible with and promotes certain social values, in ways that are surprising and rarely made explicit.

## 2.2 What Is It That Judges Do?

A longstanding debate within jurisprudence concerns what it is that judges do when they interpret laws or deliver judicial decisions. Legal formalism is the view that laws are rules derived from the linguistic meaning of legal texts, and as such have a determinate, discoverable meaning that is applicable to facts (Solum, 2005). With respect to judicial

reasoning, formalism holds that judges should (and do) decide cases based on this linguistic meaning of 'black letter law' and consistent with earlier precedent. As such, formalism implies that there is one correct way to decide cases. This adherence to rules thus restricts discretion in legal decision-making (Schauer, 1988).

Once a mainstream legal philosophy, formalism met heavy criticism from early 20$^{\text{th}}$ century scholars from a jurisprudential school of thought known as legal realism. In contrast to formalists, legal realists hold that jurisprudential reasoning does – and should – depend on factors outside of the strict textual meaning of a law.[8] Law, legal realists argue, is found not in the meaning of legal statute and precedent, but rather in the behavior of judges and legal actors – "law in action," rather than "law in the books" (Kruse, 2011; Pound, 1910). Legal realism is thus a negative claim about formalism: single, objective interpretations of legal rules are impossible, undesirable, or fail to capture what judges really do in practice.

The realist critique take many forms. One modest realist argument is that, even if legal formalist reasoning is in principle possible, it is nevertheless undesirable. For one, laws tend to outlive the worlds of their creators, and mechanically applying laws in our current context can have unanticipated harmful consequences contrary to the drafters' intentions. Hence, formalism is disparagingly referred to by its critics as "mechanical jurisprudence."[9]

Other realist critiques question the very coherence of formalism. Singer, for instance, argues that legal rules often lack the certainty demanded by formalism, and further that there are different (and sometimes contradictory) ways of reading legal precedents (Singer, 1988). Similarly, Llewelyn argues that there are always multiple "correct" ways to interpret cases. A case's interpretation depends in part on context and the "sense of the situation" of the court – in other words, an element of ineffable judicial expertise is a part of law itself (Llewellyn, 1950, 397). Other realists, like Cohen, go farther and question the coherence of legal concepts, like 'corporation' or 'person'. These concepts, Cohen writes, depend on the very questions they are used to ask, like 'is entity $x$ subject to suit'; they are thus viciously circular and empty, an illusion covering up the true social forces that drive judicial decisions (Cohen, 1944, 816).

[8]Note that legal realism, as the term is used in jurisprudence, has the opposite connotation of scientific realism.

[9]Pound, 1908 first coined this term.

Even proponents of legal realism, however, tend to agree that certain factors ought not influence judges' determination of guilt, such as a criminal defendant's race, socioeconomic background and the like. Nevertheless, jurisprudential decisions seem, in practice, to be influenced by such factors. Recent empirical studies on judges, though such studies are fairly rare, consistently lend support to legal realism as a descriptive thesis – judges' decisions are influenced not only by political leanings of judges and social climate, but also by factors like defendant characteristics (Rachlinski and Wistrich, 2017). In one such study, Spamann and Klöhn presented four fictitious scenarios to US federal judges; in each case, caselaw either strongly or weakly supported the defendant, and the defendant was described as having either favorable or disfavorable personal characteristics. These legally irrelevant defendant characteristics were stronger predictors of the judgment outcome than caselaw, even though the judges' written reasons appealed exclusively to legal principles for their decision (Spamann and Klöhn, 2016).

In sum, legal realists hold that jurisprudential reasoning necessarily depends on factors not contained in the text of the law, such as public good, popular sentiment, political climate, and the like – that there is an ineliminable human component to jurisprudence.

## 2.3 Mechanical Jurisprudence, Realized

The dialectic about the merits and value-ladenness of risk assessment instruments shares a structural similarity with debates about legal formalism and realism.[10] A standard formalist response to realist critiques of biased judges is that, even if judges are not formalists in practice – that is, they do not make decisions based strictly on legal rules – they still *should* be making decisions as formalists. Legal rules may not be unbiased, but following them to the letter, warts and all, is still more justified than idiosyncratic judgment. After all, if legal reasoning is not constrained in the formalist sense, then it is unclear what distinguishes it

[10] Green and Viljoen (2020) recently analogized algorithmic reasoning to legal formalism to critique of the former. They argue for "algorithmic realism," a call to recognize "the internal limits of algorithms and to the social concerns that fall beyond the bounds of algorithmic formalism" (1). By contrast, I am arguing that using algorithms for legal decision-making necessarily casts legal interpretation as a formalist enterprise.

from mere politics and opinion. Realist claims about the untenability of formalism does not justify its absence; at best, realism calls for greater transparency about the real nature of decisions, without providing grounds for their justification. Similarly, we might think that algorithmic decision-making in sentencing, even if it has its own sources of bias, is still preferable to idiosyncratic bias that pervades human decision-making.

Legal scholars like Ronald Dworkin have offered some middle-of-the-road responses to this issue from the perspective of jurisprudence. On Dworkin's account, legal principles do constrain judges, but not in the formalist sense – decisions cannot be mechanically derived from laws because there is an ineliminable interpretive component to jurisprudence. What judges do, on Dworkin's law-as-interpretation account, is a combination of finding and making law: much like literary interpreters, judges interpret the law to make it the best it can be while remaining consistent with what has come before (Dworkin, 1986). In particular, judges should interpret law in such a way as to maximize certain desirable features of a legal system, including justice, fairness, and due process, as well as the system's 'integrity' (in essence, its moral coherence). This, Dworkin argues, not only descriptively captures what judges claim to be doing, but also provides satisfactory *grounds* for law, i.e., justification for the use of force to enforce laws.

We need not agree with every aspect of Dworkin's story to derive a broader moral from it: the dichotomy between exclusively mechanical and idiosyncratic decisions is a false one. Law is a human enterprise and requires dynamic interpretation, but judgment is nevertheless undergirded by legal principles.

Risk assessment instruments, however, are not dynamic or interpretive in this way; they provide the same recommendation given the same demographic information, precluding the possibility to reinterpret legal rules as the world changes and a defendant's context shifts. The presumption that it is possible to generate correct mechanical recommendations from legal principles and the facts of a case is formalist, and must contend with the critical reasons realists have given against legal formalism. This means that the use of risk assessment instruments comes with a normative presumption about jurisprudence, even if the algorithms could be made value-free in a superficial sense.

The extent to which risk assessment instruments instantiate formalist reasoning in practice

depends on an empirical question, namely, how much the judge's ultimate decision is influenced by the risk score. This question – whether risk assessment instruments effectively automate judgment – was at the core of State v. Loomis, a 2016 Wisconsin supreme court dismissal of an appeal against the use of COMPAS in sentencing decisions. Loomis, a man who received a high risk score and a correspondingly harsh sentence, appealed on the basis that his due process was violated by the use of COMPAS, since the algorithm is proprietary and the details of its function are not up for dispute (State v. Loomis, 2016). The court ruled that because the output of such algorithms is merely supplementary information and is not the sole basis for a judge's decision, their use does not violate due process. The judge who sentenced Loomis even insisted that he "would have imposed the same sentence regardless of whether it considered the COMPAS risk scores" (Forward, 2017).

Here it is worth considering the prevalence of cognitive biases in human reasoning. Relevantly, automation bias refers the human tendency to assign higher levels of authority and trust to automated sources relative to non-automated sources, like other people (Park, 2019). Related is the issue of complacency, which refers to the tendency to rely uncritically on automated systems that require human oversight – people become complacent when an automated system appears to be performing its job well (Parasuraman and Manzey, 2010). Complacency is sometimes blamed for easily preventable accidents involving machines and human operators, such as recent deaths of drivers of semi-automatic Tesla cars (Boudette, 2016) or accidents involving airplane pilots relying uncritically on faulty data outputs from cockpit machinery (Parasuraman and Manzey, 2010). Considering that the US criminal legal system is overloaded and decision fatigue among judges appears to be a pervasive problem – for one, judges' decisions are influenced by how recently they have had a break (Danziger et al., 2011) – automation bias plausibly jeopardizes the legitimate use of sentencing algorithms assumed by the Wisconsin supreme court.[11] Empirical evidence is still limited, but studies on recidivism risk assessment instruments in Kentucky showed that judges are more likely to override a low risk assessment in favor of harsher bond conditions for Black defendants than for white defendants, suggesting that the real story is more complicated (and

[11]This fact is even offered as a key motivation for the development of COMPAS: "In overloaded and crowded criminal justice systems, brevity, efficiency, ease of administration and clear organization of key risk/needs data are critical. COMPAS was designed to optimize these practical factors" (NorthPointe, 2015).

more troubling) than simple automation bias (Stevenson, 2018; Albright, 2019). I discuss this issue at more length in chapter 4.

In short, the use of risk assessment instruments distorts the domain of criminal sentencing because it requires a problematic view of jurisprudence, which in turn normalizes the assumption and could shape judges' behavior. This demonstrates one striking way in which the use of algorithmic decision-making can introduce value to the legal process.

### 2.3.1 What's Special About This Case?

At this point, one might object that domain distortion, even if present in this case, is not specific to risk assessment instruments. Efforts to reduce bias and discretion in sentencing are not unique to the current move toward algorithmic decision-making – similar motivations underpinned the 1984 introduction of federal sentencing guidelines to limit "unwarranted disparity" of sentences for similar crimes, in part by establishing a system of mandatory sentencing guidelines (98th Congress, 1984). Among the changes introduced by the guidelines was a 258-box grid called the "Sentencing Table" (Figure 11), which through a complicated series of rules mechanically determines the severity of a sentence based on a defendant's criminal history (Stith and Cabranes, 1998, 3). The guidelines were introduced at a moment of draconian crackdown on crime in the heyday of the drug war in the US. Today, the federal sentencing guidelines are perhaps most notorious for requiring longer sentences for the possession of crack cocaine compared to powder (Murphy, 2002), a recognized race proxy that resulted in harsher sentences for Black people for the crime of drug possession. The guidelines also raised the percentage of crimes with a mandatory prison sentence from 50% to 85% and are one of the causes of the present crisis of mass incarceration (Starr, 2014).

At first, the domain distortion introduced by risk assessment instruments may seem different in degree, not in kind, from that of federal sentencing guidelines: both impose formalism in ways that shape sentencing practices. Wendy Espeland and Berit Vannebo, for instance, argue that sentencing guidelines have profoundly reshaped criminal sentencing by shifting the power of discretion from judges to prosecutors, who determine sentencing outcomes by deciding which and how many criminal charges to press. This has led to a surge

**SENTENCING TABLE**
**(in months of imprisonment)**

| | Offense Level | Criminal History Category (Criminal History Points) I (0 or 1) | II (2 or 3) | III (4, 5, 6) | IV (7, 8, 9) | V (10, 11, 12) | VI (13 or more) |
|---|---|---|---|---|---|---|---|
| Zone A | 1 | 0–6 | 0–6 | 0–6 | 0–6 | 0–6 | 0–6 |
| | 2 | 0–6 | 0–6 | 0–6 | 0–6 | 0–6 | 1–7 |
| | 3 | 0–6 | 0–6 | 0–6 | 0–6 | 2–8 | 3–9 |
| | 4 | 0–6 | 0–6 | 0–6 | 2–8 | 4–10 | 6–12 |
| | 5 | 0–6 | 0–6 | 1–7 | 4–10 | 6–12 | 9–15 |
| | 6 | 0–6 | 1–7 | 2–8 | 6–12 | 9–15 | 12–18 |
| | 7 | 0–6 | 2–8 | 4–10 | 8–14 | 12–18 | 15–21 |
| | 8 | 0–6 | 4–10 | 6–12 | 10–16 | 15–21 | 18–24 |
| Zone B | 9 | 4–10 | 6–12 | 8–14 | 12–18 | 18–24 | 21–27 |
| | 10 | 6–12 | 8–14 | 10–16 | 15–21 | 21–27 | 24–30 |
| | 11 | 8–14 | 10–16 | 12–18 | 18–24 | 24–30 | 27–33 |
| Zone C | 12 | 10–16 | 12–18 | 15–21 | 21–27 | 27–33 | 30–37 |
| | 13 | 12–18 | 15–21 | 18–24 | 24–30 | 30–37 | 33–41 |
| Zone D | 14 | 15–21 | 18–24 | 21–27 | 27–33 | 33–41 | 37–46 |
| | 15 | 18–24 | 21–27 | 24–30 | 30–37 | 37–46 | 41–51 |
| | 16 | 21–27 | 24–30 | 27–33 | 33–41 | 41–51 | 46–57 |
| | 17 | 24–30 | 27–33 | 30–37 | 37–46 | 46–57 | 51–63 |
| | 18 | 27–33 | 30–37 | 33–41 | 41–51 | 51–63 | 57–71 |
| | 19 | 30–37 | 33–41 | 37–46 | 46–57 | 57–71 | 63–78 |
| | 20 | 33–41 | 37–46 | 41–51 | 51–63 | 63–78 | 70–87 |
| | 21 | 37–46 | 41–51 | 46–57 | 57–71 | 70–87 | 77–96 |
| | 22 | 41–51 | 46–57 | 51–63 | 63–78 | 77–96 | 84–105 |
| | 23 | 46–57 | 51–63 | 57–71 | 70–87 | 84–105 | 92–115 |
| | 24 | 51–63 | 57–71 | 63–78 | 77–96 | 92–115 | 100–125 |
| | 25 | 57–71 | 63–78 | 70–87 | 84–105 | 100–125 | 110–137 |
| | 26 | 63–78 | 70–87 | 78–97 | 92–115 | 110–137 | 120–150 |
| | 27 | 70–87 | 78–97 | 87–108 | 100–125 | 120–150 | 130–162 |
| | 28 | 78–97 | 87–108 | 97–121 | 110–137 | 130–162 | 140–175 |
| | 29 | 87–108 | 97–121 | 108–135 | 121–151 | 140–175 | 151–188 |
| | 30 | 97–121 | 108–135 | 121–151 | 135–168 | 151–188 | 168–210 |
| | 31 | 108–135 | 121–151 | 135–168 | 151–188 | 168–210 | 188–235 |
| | 32 | 121–151 | 135–168 | 151–188 | 168–210 | 188–235 | 210–262 |
| | 33 | 135–168 | 151–188 | 168–210 | 188–235 | 210–262 | 235–293 |
| | 34 | 151–188 | 168–210 | 188–235 | 210–262 | 235–293 | 262–327 |
| | 35 | 168–210 | 188–235 | 210–262 | 235–293 | 262–327 | 292–365 |
| | 36 | 188–235 | 210–262 | 235–293 | 262–327 | 292–365 | 324–405 |
| | 37 | 210–262 | 235–293 | 262–327 | 292–365 | 324–405 | 360–life |
| | 38 | 235–293 | 262–327 | 292–365 | 324–405 | 360–life | 360–life |
| | 39 | 262–327 | 292–365 | 324–405 | 360–life | 360–life | 360–life |
| | 40 | 292–365 | 324–405 | 360–life | 360–life | 360–life | 360–life |
| | 41 | 324–405 | 360–life | 360–life | 360–life | 360–life | 360–life |
| | 42 | 360–life | 360–life | 360–life | 360–life | 360–life | 360–life |
| | 43 | life | life | life | life | life | life |

Figure 11: The Sentencing Table from the US Federal Sentencing Guidelines. "The Offense Level (1–43) forms the vertical axis of the Sentencing Table. The Criminal History Category (I–VI) forms the horizontal axis of the Table. The intersection of the Offense Level and Criminal History Category displays the Guideline Range in months of imprisonment. 'Life' means life imprisonment" (United States Sentencing Commission, 1987).

in plea bargains and a general pressure on defendants to plead guilty to minimize prison time (Espeland and Vannebo, 2007). Critics of federal sentencing guidelines also make reference to an issue similar to automation bias, pointing out that the system of rules in the federal sentencing guidelines "lends an appearance of having been constructed on the basis of science and technocratic expertise, giving it a threshold plausibility to a general public not familiar with its actual contours and operation" (Stith and Cabranes, 1998, xi).

To this I respond that, though risk assessment instruments and federal sentencing guidelines share a similar goal and exacerbate racial disparities in practice, sentencing guidelines do not shift how the domain of criminal sentencing is reasoned about in the same way. This is because sentencing guidelines do not fall into the purview of jurisprudence and thus are not subject to critiques of formalism, whereas risk assessment instruments do and are. To show why, it is necessary to introduce a second form of domain distortion due to risk assessment instruments, namely, the shift in how liability assessment and sentencing are treated in relation to each other.

## 2.4 Blurred Lines

Traditionally, jurisprudence has considered sentencing and liability assessment (i.e., determination of guilt) as distinct enterprises, except in unusual circumstances like capital punishment cases, which can be decided by juries. The separation of these domains is reflected in courtroom practices – juries are instructed not to consider the punishment when making liability assessments; facts are held to a different standard in sentencing than in liability; and even back when federal sentencing guidelines were mandatory, judges had far more discretion about sentencing than they do about liability assessment (Ross, 2002). I argue, however, that the line between these domains is blurred by the use of risk assessment in sentencing. This is because risk assessment instruments are *predictive* algorithms: they explicitly take future liability assessments into consideration when deciding sentences for current liability assessments. Federal sentencing guidelines, on the other hand, belong to the domain of sentencing; as such, they remain comfortably insulated from jurisprudential critiques, though

they can (and should) be criticized on other grounds.

Presuming that sentencing and liability assessment are separate domains (or not) carries important normative baggage. When the Federal Sentencing Commission set out to draft sentencing guidelines in 1984, it confronted what it referred to as the "philosophical problem" of determining "the purposes of criminal punishment": is the purpose of punishment to serve retribution proportional to a defendant's culpability for a crime ("just desert"), or is it to lessen the likelihood of future crime, either by deterring others or incapacitating the defendant ("crime control")? Rather than dealing with this difficult issue, the commission simply assumed that following the former will help with the latter (Monahan, 2006). Ultimately, it was decided that information about criminal history could be used in determining sentences, but that defendant characteristics like age or race, which have "little moral significance" (Moore, 1986, 317) cannot be used in sentencing, even if they are statistically predictive of recidivism (Monahan, 2006).

Conversely, many risk assessment instruments (including COMPAS, the PSA, and the Sentence Risk Assessment Instrument) do take 'morally insignificant' variables – including age, gender, education history, and familial relationships – into account. This, in effect, presupposes that the purpose of punishment is consequentialist (crime control) rather than deontological (retributive), and breaks down the separation between liability and sentencing. My purpose here is not to advocate for a particular position on sentencing, but to point out that the consequentialist values implicit in risk assessment instruments distort how the domain of criminal sentencing is reasoned about when using other methods, like sentencing guidelines.

There is, however, important nuance here. Notably, even before the advent of risk assessment instruments, judges were permitted to consider recidivism risk, historically based on clinical judgment, when deciding sentences. This suggests that the boundary between liability and sentencing may not have been particularly sharp to begin with. Risk assessment instruments make the role of future liability assessment in current liability assessment more explicit, but how much further they dissolve the separation between these domains in practice depends on how much judges considered recidivism in the first place, which is an empirical question. Indeed, views about the role of recidivism risk are tied to positions about the

purpose of the US criminal legal system more broadly, which have shifted substantially over the last several decades. These positions, respectively, come with distinctive sets of responses to crime. To illustrate why risk assessment instruments are compatible with only certain kinds of interventions, I briefly provide some context on these historical shifts.

## 2.5 The New Penology: Surveillance and Control

Early 20$^{th}$ century positivist criminology in the US focused on the incapacitation of dangerous criminal types through imprisonment and attempts to limit the spread of hereditary criminality via eugenics, especially forced or coerced sterilization. This idea originated with Italian criminal anthropologists like Cesare Lombroso, who argued that many criminals were destined for a life of crime by their biological inheritance – in short, that they were "born criminals" and thus unreformable (Simon, 2005, 2145). Since born criminals are unchangeable, the theory held, they should be given harsh treatment, such as permanent detention or execution. By contrast, 'occasional criminals' (a more rare type) could be changed through treatment and could thus be subject to softer penal practices.

Post-World War II and until the 1970s, the focus of US positivist criminology shifted to individual responsibility and therapy. Rehabilitation emerged as the broader purpose of the US penal system; the causes of criminality were to be diagnosed and 'treated', much like doctors treated illnesses, in order to reform inmates and return them to society (Phelps, 2011). But with evidence for significant effects of the rehabilitative ideal lacking and political attacks on discretionary and indeterminate sentencing models from critics on the left and right, rehabilitation became discredited. A major 1974 report concluded that "nothing works" in rehabilitating prison inmates and that sentences should be considered separately from rehabilitative goals (Martinson, 1974; Garland, 2002).

Scholars argue that a new paradigm of punishment, characterized by a focus on control and surveillance of populations (Foucault, 1975), has dominated in the US criminal legal system since the 1980s, leading to a sharp increase in incarceration (Feeley and Simon, 1994; Harcourt, 2007; Garland, 2012). The result has been the crisis of mass incarceration; since

the 1980s, the incarcerated population has risen by 500% to over 2 million people (Calabresi, 2014). The focus in the 1980s shifted toward the actuarial classification and management of 'risky' classes of individuals, with high-cost maximum security prisons for the highest risk groups and an explosion of low-cost electronic surveillance techniques for the lowest risk groups (Feeley and Simon, 1992; Garland, 2012).[12] As legal scholar Jonathan Simon argues:

> At the heart of this project is the conviction – which American penal policies continue to reflect – that crimes are committed by a distinguishable group of persons with a proclivity toward law-breaking and that crime control policies should seek to isolate and repress these dangerous classes.

Risk assessment instruments are one part of this broader strategy of control and management.[13] The instruments are in keeping with tenets of positivist criminology, including the belief that criminality is rooted in the measurable differences found between criminal and normal people, and the resulting science of crime control, particularly the link between criminal legal institutions and mechanisms of scientific data collection, analysis, and surveillance promoted by evidence-based sentencing (Simon, 2005).[14] This is reflected in practice through the use of risk assessments to inform interventions on risky individuals rather than structural interventions.[15]

Sociologists Seth Prins and Adam Reich (2017) illustrate this point incisively in their paper "Can we avoid reductionism in risk reduction?" They focus on the "risk-needs-responsivity" (RNR) framework, a social psychology theory of crime that is the basis for one of the commonly used[16] 3rd generation recidivism risk assessment instruments in use today, the "Level of Services Inventory" (LSI) (Andrews and Bonta, 2010). Based on risk factors

[12]This phenomenon generalizes more broadly to a neoliberal "risk society" in which actuarial predictions of risk govern numerous aspects of social life as they are increasingly embraced by institutions and organizations (Baker and Simon, 2002; Fourcade and Healy, 2013).

[13]This echoes Lacey (1999)'s observations about the modern values of control upheld through biotechnological agricultural research.

[14]Several critics have argued that Feeley and Simon (1992)'s account overstates the effects of actuarial risk thinking. Hannah-Moffat and O'Malley, for instance, argue that in practice, considerations of risk have evolved and are hybridized with rehabilitation, for instance with the increasing uptake of 3rd generation risk needs assessments, which are supposed to identify dynamic loci for rehabilitation (Hannah-Moffat, 2005, 2019; O'Malley, 2010). If criminal sentencing is informed by risk assessment, however, even the inclusion of dynamic factors seems at odds with any rehabilitative aims of incapacitation.

[15]Similar assumptions underpin place-based crime prediction, such as predictive policing instruments like PredPol.

[16]The LSI is used in over 900 correctional institutions in the US and Canada (Lowenkamp et al., 2009).

that are most associated with recidivism (re-arrest) in a population sample of individuals in community corrections supervision, individuals are classified into 'risk classes', which inform everything from pre-trial detention to resource allocation. The four risk factors most predictive of recidivism are all psychological in nature: a history of antisocial behavior; antisocial personality pattern; antisocial cognition; and antisocial associates (Andrews and Bonta, 2010, 131; Prins and Reich, 2017). The focus on dynamic risk factors that are potential targets for intervention means the LSI is in principle concerned with not only with risk assessment, but also on risk reduction (Andrews and Bonta, 2010, 132–133), though this aspect of risk assessment is typically ignored in practice (Latessa and Lovins, 2014).

Prins and Reich point out several theoretical problems with this approach.[17] The first is simply that the LSI is an actuarial tool – it identifies risk factors that are "likelihoods based on group averages" (Prins and Reich, 2017, 5) – but those very factors are used to target individuals for intervention, which conflates predictive variables with causal ones. Consider the risk factors 'antisocial behavior' and 'antisocial peers'. If antisocial behavior causes antisocial peers and recidivism, then intervening on antisocial peers will not affect recidivism. Conversely, if having antisocial peers is a cause of antisocial behavior and recidivism, then intervening on antisocial behavior will not affect recidivism (Figure 12).

The second critique, which illustrates one way in which risk assessment instruments promote values of control, is that the RNR theory focuses on individual-level psychological predictors of crime and downplays population-level causes of crime, such as class and poverty. This is because the studies on which the RNR framework is based measure inter-individual variation within a selected population, namely, people who have had some involvement with criminal legal institutions. But as Prins and Reich note, distal factors like low socioeconomic status are virtually ubiquitous in the population of criminal defendants, so the focus on predicting inter-individual differences in recidivism could effectively mask the contribution of those population-level causes. In fact, the original authors of the framework explicitly discount population-level causes of crime widely discussed in sociology, such as class, arguing that it is a "myth" that the "roots of crime are buried deep in structured inequality" (Andrews

[17] I focus here on the critique by Prins and Reich, but Monahan and Skeem (2016) and others make similar arguments.

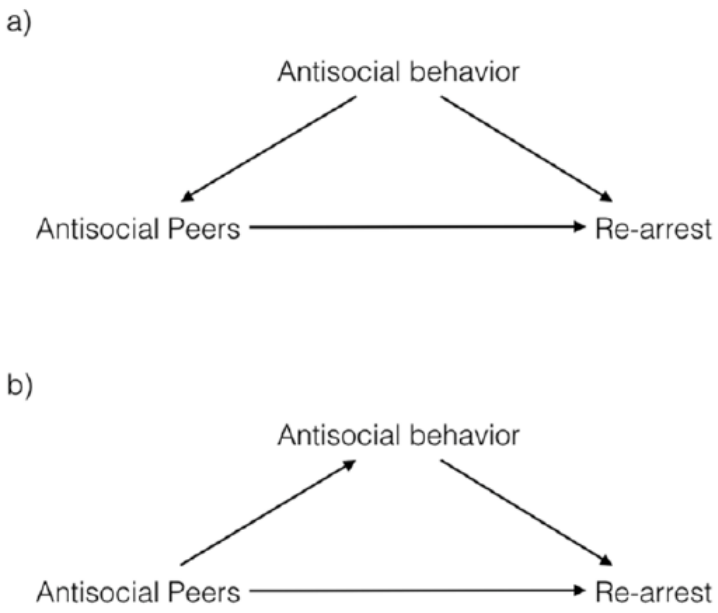


Figure 12: Whether intervening on antisocial peers and antisocial behavior would be effective for reducing recidivism depends on which DAG explains the observed actuarial associations (Prins and Reich, 2017, 6).

and Bonta, 2010; 79, 93). As epidemiologists like Geoffrey Rose (2001) have long pointed out in critiques of case-centered epidemiology, to understand differences in a distribution, it is necessary to study features of the populations, not of individuals – "the causes of a distribution are rarely the same as the causes of an individual's place within a distribution" (Prins and Reich, 2017, 8).

The RNR framework also ignores second-order risks, distal causes that put people at risk of proximate causes of crime. Intervening on a dynamic risk factor for recidivism will have little effect if the causes of the risk factor are not addressed. For instance, rates of obesity and diabetes are high among impoverished communities, but instructing those communities to avoid processed foods without addressing their limited access to such food is not likely to be impactful (9).

Finally, factors that may seem static (not a target for intervention) at the individual level might be dynamic (changeable) through interventions on a population-level. Race,

for instance, is a static variable that means not just skin color but also a package of lived experiences that someone in a racialized social position experiences. What this social position looks like – for instance, the way it is perceived and treated by institutions – *is* changeable. If the focus is on individual interventions, then some potentially dynamic loci for intervention will be missed (9-10).

Together, these value-laden parts of the RNR theory, characteristic of actuarial risk assessment approaches more broadly, are predisposed to favoring interventions for individuals at the highest risk part of the risk distribution over population-level interventions to shift the mean down, including interventions like improving housing, reducing economic inequality, investing in communities, and redressing racial discrimination. These structural interventions are essential for substantive progressive reforms or abolitionist values in the criminal legal system; the risk assessment tool is compatible only with the control and management of classes of people perceived to be risky. Following O'Malley (2010), Prins and Reich conclude that attention should be directed toward systemic harm reduction rather than simply punishment and individual-level interventions.

The social values that influence and are influenced by risk assessment instruments are thus a case of domain distortion. In addition to presupposing values of formalism and control, recidivism risk assessment thus reinforces social values of control in practice through its compatibility with criminal justice interventions focused on individual-level risks, rather than the structural inequities that "put people at risk of individual-level risks" (10).

## 2.6 Summary

The value-ladenness of algorithmic methods is typically discussed in the context of epistemic risk and algorithmic bias. In this paper, I examined a deeper sense of value introduced by algorithmic methods: *domain distortion*, the influence and reification of social values, which distorts how a domain of application is reasoned about. I illustrated how domain distortion can occur through an analysis of the use of risk assessment instruments in criminal sentencing. Using insights from jurisprudence, I argued that risk assessment instruments

presuppose legal formalism and blur the line between liability and sentencing, which presumes that the purpose of punishment is consequentialist and reinforces social values of control that are currently dominant in the criminal legal system. Finally, I showed how risk assessment promote values of control in practice through their compatibility with interventions on certain parts of a risk distribution, rather than structural interventions. Domain distortion provides a distinctive avenue for values to become entrenched in the domain that algorithms are applied to, a value entry-point that is neglected by a focus on epistemic risk.

In the following chapter, I examine the domain distortion of crime prediction methods in another historical and political setting – legal cybernetics in the post-Stalin Soviet Union – to illustrate the consequences and rhetorical utility quantitative methods can have for the scientific authority of criminology.

## 3.0 Mathematizing Crime and Punishment: Legal Cybernetics in the Post-Stalin Soviet Union

Information theory has, in the last few years, become something of a scientific bandwagon. ... Applications are being made to biology, psychology, linguistics, fundamental physics, economics, the theory of organization, and many others. In short, information theory is currently partaking of a somewhat heady draught of general popularity.

Although this wave of popularity is certainly pleasant and exciting for those of us working in the field, it carries at the same time an element of danger. It will be all too easy for our somewhat artificial prosperity to collapse overnight when it is realized that the use of a few exciting words like information, entropy, redundancy, do not solve all our problems.

–Claude E. Shannon, "The Bandwagon," 1956.

"Science only achieves perfection when it becomes mathematical." This adage, attributed to Karl Marx and often cited in preludes to criminology articles published in the Soviet Union,[1] reflects a traditional and pernicious view of the natural world: as sciences mature, they become grounded in mathematics, and thus increasingly objective. Despite much criticism by philosophers of science,[2] the privileging of reductionist 'hard' science is deeply ingrained and continues to persist in many disciplines.

Science in the Soviet Union was no stranger to this theme. The preference for quantitative sciences post-Stalin, in particular, had two additional motivations that made it particularly widespread: conforming with the prevailing Marxist-Leninist view that science reaches perfection as it becomes mathematical, and purging the influence of decades of Stalinist ideology on science. In this vein, Soviet science in the 1960s saw a promulgation of methods from 'cybernetics', the study self-regulating systems, a field closely associated with early work in computer science, information theory, mathematical modeling, and artificial intelligence. In the two decades after Stalin's death in 1953, the language and methodology of cybernetics was adopted by many previously non-mathematical fields of study, including economics, genetics, and linguistics, often with the explicit motivation of making them scientific and objective.[3] The bandwagon Claude Shannon warned about in 1956 was in full swing in the Soviet Union in the 1960s and 70s.

Criminology, the study of crime, features on this bandwagon. The appeal that cybernetics methods held for post-Stalin Soviet criminologists is clear: theorizing about the causes of crime was not only deeply laden with ideological values – in the 1930s crime had been officially considered to be a vestigial feature of capitalism and its study had been correspondingly outlawed – but also severely needed – a massive spike in crime during the Khrushchev political thaw in the 1960s and 70s made reviving criminology a key part of the Soviet political agenda.[4] Much like in other historical episodes in which quantification was a solution for declining institutional authority,[5] the mathematical apparatus of cybernetics was used strategically by

[1]For instance, this quote appears in Poshkiavichius, 1974, 7, Polevoi and Shliakhov, 1977, 3, and Pankratov, 1967, 134. Because of the high quantity of references in this chapter, they will be delegated to footnotes.
[2]Fodor, 1974; Mitchell, 2009.
[3]Gerovitch, 2002.
[4]Solomon, 1974.
[5]Porter, 1995.

Soviet criminologists to give their field authority by making it appear more scientific. Like the use of risk assessment instruments in the US, legal cybernetics was also part of as a broader set of expansionist projects that institutionalized the observation and control of populations, with the aim of eliminating crime and moral vices such as alcohol consumption, in keeping with the 20th century rise in sociological surveillance and data collection in the West.[6]

Although the application of cybernetics to criminology is not inherently problematic, the application of quantitative methods on its own of course cannot lend scientific status to a discipline and in fact may produce the illusion of objectivity. Indeed, as cybernetics gained popularity in the Soviet Union in the 1960s, it rapidly began to lose its intellectual content; appealing to cybernetics increasingly served as a rhetorical strategy, sometimes to promote claims with poor theoretical grounding.[7] Using archival material I accessed and translated at the Moscow State Library in 2018, I argue that the genesis of illusory objectivity captures much of the role that the adoption of cybernetics methods in the study of law and crime – known as *legal cybernetics* – played in the rise of the field of Soviet criminology in the 1960s and 1970s. While legal cybernetics did coincide with the rising scientific authority of Soviet criminology, I argue that it also inherited, obscured, and promoted existing ideological values in the field – an instance of domain distortion.

Although Soviet cybernetics and criminology have both been examined by historians at some length,[8] no historical writing exists on the intersection of these two fields, nor the relationship between legal cybernetics and the scientific authority of Soviet criminology. With the prevalence of quantitative recidivism risk assessments in contemporary criminal legal systems, often adopted with the motivation of making penal decision-making more 'objective', it is more important than ever to understand the historical context of crime prediction and its relationship with objectivity. This chapter not only remedies an important absence in the literature on Soviet science, but also provides a valuable route to this understanding.

I begin with a novel synthesis of the rich and fascinating parallel histories of cybernetics and criminology in the rapidly shifting political landscape of the post-Stalinist period. Next,

[6]Bratich, 2018.

[7]See the discussion of attempts to optimize the Soviet economy using cybernetics in Gerovitch, 2002.

[8]See Gerovitch, 2002 for an excellent history of Soviet cybernetics, and Solomon, 1974 and Shelley, 1979b for discussions of Soviet criminology.

I discuss the relationship between quantification and epistemic authority and how it relates to the notions of objectivity employed in legal cybernetics. I argue that the mechanical objectivity that derived from the use of quantitative methods served to raise the epistemic authority of criminology, frequently coupled with rhetorical claims about legal cybernetics' absolute objectivity. As an illustration, I focus specifically on the work of Vladimir Nikolaevich Kudriavtsev (1923–2007), a prominent Russian criminologist and law professor who was a major player in the revival of criminology and the promotion of legal cybernetics, and his applications of cybernetics to study the causes of crime. I identify one important value-laden methodological assumption in this body of work: the exclusion of economic causes of crime in causal variable choice in his cybernetic models of crime, which served to reinforce long-standing dogmatic values in Soviet criminology.

## 3.1 Stalin's Dark Legacy

From the time Joseph Stalin took charge of the Soviet Union in the 1930s to his death in 1953, Stalinist ideology permeated every part of Soviet life, including science. Indeed, if there is any one factor that distinguishes the history of science in the Stalinist period, it is the stock-in-trade saturation of political and social values in the lives of scientists, as well as in scientific theory itself.[9] Soviet attitudes toward 'bourgeois' science, in turn, typically fell into one of two perspectives: the "criticize and destroy" perspective, which decried certain scientific developments and forbade their study (the paradigm example of this is the Lysenko affair in genetics), and the "overtake and surpass" perspective, which emphasized the superiority of Soviet science and competition with the West (e.g., the 'space race').[10] Attitudes toward criminology and cybernetics both underwent dramatic shifts in the post-Stalinist period. After a long period of suppression and abuse, the field of criminology re-emerged in the late 1950s, aiming to re-establish its credibility. Cybernetics, once derided as a Western pseudoscience, became prominent and was applied in many disciplines – including criminology

[9]See Graham, 1987 for a thorough discussion of science in the Soviet Union.
[10]This push and pull is a recurring theme in Gerovitch's history of Soviet cybernetics.

– that sought to purge themselves of Stalinist ideology; ironically, cybernetics itself soon lost intellectual content.

### 3.1.1 The Renaissance of Soviet Criminology

Under Stalin, the study of criminology was effectively forbidden. As the repression of the Stalinist period subsided in the late 1950s, Soviet criminology went through a renaissance: it became a major *scientific* area of study, complete with numerous governmental institutes devoted to its activities; it became included in legal education; and it exerted influence on political measures. Nevertheless, criminology continued to be imbued with political values. Understanding the rebirth of Soviet criminology as science sheds light on why the objectivity associated with legal cybernetics held such appeal.

Russian criminology has gone through several distinct phases. In the early 1900s, prior to the revolution, a sociological school of criminology prevailed, emphasizing the socioeconomic factors of crime causation, rather than "moral defects," as had been vogue in earlier schools of Russian criminology; legal scholars of this period emphasized the importance of preventing crime rather than punishing it, and opposed measures like the death penalty.[11] After the Bolshevik revolution in 1917, the new Marxist-Leninist party held that crime was primarily a result of the social and economic conditions of capitalism. Crime "arose only on that stage of development of society when private property, classes, and the state appeared."[12] Any remaining crime in socialist society was due to the "outdated values of capitalism instilled in old-fashioned individuals," personal defects which could be changed by re-education and rehabilitation.[13] Correspondingly, criminological studies focusing on criminal psychology and medical anthropology flourished during this time.[14]

Early Soviet criminology's vigorous period of activity was short-lived. By the late 1920s, party leaders gradually began dismantling laws and legal institutions that impeded their goals

[11]Semukhina, 2017, 422; Gilinskiy, 2017, 114.
[12]Large Soviet Encyclopedia, 1940.
[13]Gernet, 1922, quoted in Semukhina, 2017, 423.
[14]Solomon, 1974, 123–124; in 1918, the department of "moral statistics" was created, which collected data about alcoholism, crime, and suicide (Gilinskiy, 2017, 114); in 1925, the State Institute for the Study of Crime and the Criminal was established, which conducted empirical studies on crime, including a survey of 125,000 prisoners and labor camp inmates (Semukhina, 2017, 423).

and introduced a politicized criminal law code;[15] the party became hostile to criminology, a field intimately tied up with criminal law and legal institutions.[16] Compounding this growing suspicion of criminology, in the 1930s psychiatric and medical studies of criminals were accused of neo-Lombrosianism and "pseudo-science serving the bourgeois interest by confusing the minds of proletariat."[17] The study of the transient phenomenon of crime was deemed to serve no purpose, and criminological studies effectively disappeared.[18] What remained of criminology during the Stalinist period existed primarily to justify the repression and murder of the state's political enemies, including wealthy peasants (Kulaks) and ethnic and religious minorities.[19] Criminal convictions as a form of political repression occurred on a mass scale under Stalin, and political opponents known as 'enemies of the people' ('vragi naroda') or 'enemies of the proletariat' ('vragi proletariata') were convicted of outlandish crimes in high-profile 'purge trials'; millions of people were also sent to Gulags or executed without trial – over one million people were killed just in the years 1937–1938.[20]

Soon after Stalin's death in 1953 and the start of Khrushchev's political thaw, criminology was reinvigorated. Khrushchev publicly denounced Stalin, released millions of Gulag prisoners, and reformed criminal law; a brief 'soft-line' approach to crime began.[21] By the 1960s, with crime rates on the rise and showing no signs of abating, understanding the causes of crime became a top priority of Khrushchev's government.[22]

The revival of criminology came with some constraints. First, criminology ultimately needed to serve the Soviet government's aims of crime prevention;[23] because it had been suppressed in part for its lack of utility, the practical role that criminology could play for the state was frequently emphasized by criminologists in their writing.[24] At the same time – often

[15]Maggs, 2017; Shelley, 1979a, 394.
[16]Solomon, 1974, 124.
[17]Bulatov, 1929; Solomon, 1974, 125; Semukhina, 2017, 424.
[18]Maggs, 2017; Gilinskiy, 2017, 114. In 1933 the State Institute for the Study of Crime and the Criminal became the Institute of Criminal and Correctional-Labor Politics; in 1937 it was closed down completely (Semukhina, 2017, 424).
[19]Shelley, 1979b; Solomon, 1974; Shelley, 1979a, 395; Kotljarchuk and Sundström, 2017.
[20]Maggs, 2017, Ellman, 2002.
[21]Dobson, 2009, 5; Solomon, 1974, 131.
[22]Dowling, 2013, 1.
[23]Solomon, 1974, 135–136.
[24]I observed this firsthand in primary sources. For instance, Andreev and Kerimov discuss a 1959 American conference presentation on information theory and law, but ultimately dismiss the paper's significance on the basis that it is "insufficiently tied to the practical needs of jurisprudence" (Andreev and Kerimov, 1961;

in tension with the first constraint – criminological theories needed to be compatible with Marxist-Leninist philosophy, which condemns the innateness of human traits. The 'bourgeois' study of biological and economic causes of crime continued to be condemned, while research on the 'personality of the criminal' and social causes of crime was emphasized.[25]

In part for these reasons, Soviet criminology in the 1960s and 70s continued to be plagued by corruption and ideological bias. Criminological texts published well into the 60s continued to claim that the crime rate was going down, even as it was steadily on the rise.[26] To suppress the official crime rate, the state diagnosed political dissidents with mental illnesses like "sluggish schizophrenia," whose symptoms included "reform delusions," "struggle for the truth," and "perseverance"; by the 1970s, an estimated one third of political prisoners in the Soviet Union were locked up in mental institutions.[27] Possible directions of criminological research were significantly cordoned by assumptions in the Marxist-Leninist framework and the practical necessity criminology was to provide to the state.

Nevertheless, from the 1960s through the 1980s, criminology steadily grew in stature and re-established itself as a prominent academic discipline.[28] Demonstrating the utility and reliability of criminology came along with demonstrating that its findings were not determined by the political desires of the state – that they were objective. One prominent strategy to achieve this goal was to show that criminology was grounded in mathematics and formalism, which cybernetics provided.

### 3.1.2 Cybernetics: From Western Pseudoscience to Paragon of Objectivity

Historian Slava Gerovitch writes that no other field of science in the Soviet Union was the target of such dramatic changes in attitude as cybernetics.[29] Decried under Stalin as a "reactionary pseudoscience" and "an ideological weapon of imperialist reaction," cybernetics emerged in the late 1950s as a prominent science.[30] By the 1970s, as numerous disciplines

Allen, 1959).

[25] Solomon, 1974, 135; Semukhina, 2017, 424.

[26] For instance, Kudriavtsev, 1967 writes that the number of convicts in 1964 was 42.7% lower than in 1958, despite the population having grown (10).

[27] van Voren, 2010, 33.

[28] Semukhina, 2017, 424.

[29] Gerovitch, 2002.

[30] *Short Philosophical Dictionary*, (Rosental' and Iudin, 1954). I first saw these quotes in Gerovitch, 2002, 4.

began adopting cybernetics, it began to lose its intellectual content and its popularity eventually dwindled.

The origins of cybernetics can be found in Norbert Wiener's 1948 book "Cybernetics: or Control and Communication in the Animal and the Machine," in which he discusses information processes, control, communication, and entropy in living and artificial systems. Other prominent contributors to the field were Claude Shannon (cited in the quote at the start of this chapter), who created information theory, and George Boole, who studied mathematical logic.[31] In Soviet writing, cybernetics is often used interchangeably with its methods: information theory, logic, and computation. The line between computation and cybernetics, in particular, is often blurred in Soviet writing on cybernetics; one scholar calls computers the "children" of cybernetics.[32]

During the Stalinist period, the emerging Western science of cybernetics and was publicly treated from the "criticize and destroy" perspective. Soviet journalists, seeking to fulfill anti-Western propaganda quotas, published numerous articles in the early 50s decrying a strawman cybernetics, with titles like "Cybernetics – An American Pseudo-Science" and "The Science of Modern Slaveholders."[33] In classified military work of the same period, however, scientists readily applied cybernetics methods and created computer technology as part of the Soviet strategy to "overtake and surpass" the Western military.[34]

After Stalin's death, cybernetics emerged as a prominent science, whose mathematical methodology was seen as holding promise for reforming other ideology-laden branches of Soviet science.[35] Two years post-Stalin, three military scientists – who had for years been studying cybernetics in classified military research – published an influential article in which they praised cybernetics and discussed its value for Soviet science.[36] Soon, realizing their mistaken condemnation of cybernetics, catching up to Western advances in computing became a top priority for Soviet officials. In 1961, cybernetics was included in the *Program of the Communist Party* as a science important for communism, and a series of published volumes

[31] A textbook on legal cybernetics, *Foundations of Legal Cybernetics*, lists Shannon and Boole as the prominent contributors to the field. Polevoi and Shliakhov, 1977, 11–12.
[32] Shliakhov, 1967, 7.
[33] Gerovitch, 2002, 119.
[34] Gerovitch, 2002, 131.
[35] Gerovitch, 2002, 4.
[36] Sobolev et al., 1955; also discussed in Gerovitch, 2002.

called *Cybernetics in Service of Communism* began the same year.[37] The technical and formal nature of cybernetics held appeal to scientists and Soviet intellectuals, who hoped to use it as a tonic for ideology-laden academic discourse of the Stalin era.[38] The perceived objectivity of cybernetics led to its adoption in many areas of science, social science, and the humanities; genetics was studied using 'biological cybernetics', structural linguistics using 'cybernetic linguistics', and so on.[39]

The popularity of cybernetics eventually led to its downfall. Gerovitch argues that, as its concepts became broad and universally applied, cybernetics became polysemous – and, in effect, contentless. By the 1970s, former proponents of cybernetics joked: "They told us before that cybernetics was a reactionary pseudo-science. Now we are firmly convinced that it is just the opposite: cybernetics is not reactionary, not pseudo-, and not a science."[40]

The publication *Cybernetics in Service of Communism* ended in 1981.[41] Before its descent into insignificance, however, cybernetics was taken up by criminology, forming a subfield known as legal cybernetics.

## 3.2 The Origins of Legal Cybernetics

One of the notable features of criminological research in the post-Stalinist period was its fervor for mathematical and quantitative methods. Publications from this period frequently began by quoting Marx,[42] Lenin,[43] or sometimes even Kant[44] regarding the importance of mathematics in science, and emphasize the objectivity of these approaches as significant for the burgeoning science of criminology. A later textbook on legal cybernetics is quite explicit about this: "the use of tools and methods of mathematics contributes to increasing the objectivity and accuracy of the research and the results obtained, on the basis of which legally

[37] Gerovitch, 2002, 256; Nauchnyi sovet po kibernetike [Scientific Council on Cybernetics], 1961.
[38] Gerovitch, 2002, 154–155.
[39] Ibid.
[40] Molchanov, 1998, 402. Quote and reference taken from Gerovitch, 2002, 4.
[41] According to the database I was able to access at the Russian State Library.
[42] E.g., Kudriavtsev, 1967, 6.
[43] E.g., Ratinov, 1967, 180.
[44] E.g., Polevoi and Shliakhov, 1977, 7.

important decisions are made."[45][46] The desirability of objective criminology is clear, given its recent historical context. Indeed, Gerovitch argues that Soviet scholars in this period became fixated on making their fields more 'objective'.[47] In the next section, I revisit some notions of scientific objectivity and relate them to epistemic authority and discussions of objectivity in the legal cybernetics literature.

### 3.2.1 Objectivity, Quantification, and Authority

As I discussed in chapter 1, scientific objectivity has taken on different meanings and importance in different contexts and historical periods.[48] In general, objectivity connotes ideas of impartiality and unbiasedness, both as an epistemic ideal and sometimes as a moral value. It is a key idea both in law and in science: the impersonal treatment of individuals according to "objective standards" is a central idea in law and combines both moral and epistemic standards,[49] and objectivity is intimately tied up with scientific authority – the epistemic authority of science derives from the presumed objectivity of scientific reasoning.[50]

Objectivity is frequently mentioned in Soviet writing on legal cybernetics. It is typically described, if at all, as what it is not: devoid of a factual basis, or based on subjective judgments.[51] This is how one legal cybernetics publication describes the most epistemically demanding version of objectivity:

> Objective truth is a characteristic of knowledge ... that is determined by the very nature of the displayed object, and does not depend on man or humanity. The objective truth of a judgment or position is in opposition to its falsity.[52]

[45] Polevoi and Shliakhov, 1977, 7: "использование средств и методов математики способствует повышению объективности и точности проводимых исследований и получаемых при этом результатов, на основе которых принимаются юридически значимые решения."

[46] A note on translations: every quote for which I include original text was translated by me. Each primary source that was accessed on my trip to Moscow in 2018 is labeled as such in the bibliography.

[47] One rough proxy of this is that a Google Books N-grams search on the word "объективность" (objectivity) in the Russian corpus shows a steady rise up through around 1917 (the year of the Bolshevik revolution), after which it steadily declines and plateaus until Stalin's death in 1953, when it skyrockets. Thanks to John Norton for pointing me toward this trend.

[48] Lorraine Daston and Peter Galison's *Objectivity* discusses the epistemic values of truth-to-nature, mechanical objectivity, and trained judgment through the evolution of scientific image production.

[49] See Porter, 1995's discussion of Kent Greenawalt's *Law and Objectivity*; 5.

[50] Reiss and Sprenger, 2017.

[51] E.g., Kudriavtsev and Eisman, 1964: "We must free ourselves of the misconception that the social sciences are based on maybes, the sciences of subjective judgments devoid of an objective basis."

[52] Trusov, 1967, 30: Объективная истинность – это характеристика знания ... которое определяется

When discussing the benefits of cybernetics, Soviet criminology scholarship of this period often slips between this kind of absolute objectivity – in the sense of grasping the true facts 'out there' in the world – and mechanical objectivity, or freedom from individual contribution.[53] For instance, another legal cybernetics publication, which discusses the creation of causal models based on tabulated statistics on crime, discusses objectivity in three different ways within a single paragraph. These correspond roughly to objectivity [1] naturally emerging from large quantities of data (absolute), [2] free from individual views and bias (mechanical), and [3] agreed upon by multiple methods (absolute/mechanical):

> In our experience, objectivity of research is largely achieved due to [1] the *mass character* of the analyzed facts. In the mass study of a trend, causality blows a hole even in the wall of subjectivism. The questionnaire for each criminal case is filled out based on the basis of evidence, and [2] *does not merely reflect the views and assessments* of the persons giving answers and the investigators filling out the questionnaires. ... In addition, the findings of the research questionnaire are checked by other methods of sociological research, such as a social experiment, statistical data, reporting documents, etc., which are also used to generalize the results of the crime prevention work in the preliminary investigation stage. [3] *Superior synthesis of research methods* is the most important criterion of objectivity.[54]

Rhetorically, the term 'objectivity' in Soviet writing of this period is frequently associated with mathematics and quantification.[55] The association between cybernetics and objectivity, in turn, came from its association with mathematics – computer programming in the Soviet Union originated as a branch of mathematics, and computer algorithms were "mathematical machines," with the corresponding association of rigor, universality, and incorruptibility.[56] Two Russian historians put it this way: "when it turned out that words lied, formulas looked

самой природой отображаемого объекта, не зависит ни от человека, ни от человечества. Объективная истинность суждения или положения противостоит его ложности.

[53]Mechanical objectivity is discussed at length by Daston and Galison, who define it as the generation of knowledge with no trace of the person who generated it (Daston and Galison, 2007).

[54]Chugunov and Gorskii, 1967, 155. "В нашем опыте объективность исследования во многом достигается за счёт массовости анализируемых фактов. В массовом исследовании тенденция, причинность пробьют брешь даже в стене субъективсизма. Анкета по каждому конкретному уголовному делу заполняется по материалам доказывания, а не отражает в себе только взгляды и оценки лиц, дающих ответ на вопросы её, и следователей, заполняющих анкету. ...Кроме того, выводы анкетного исследования перепроверяются другими методами конкретно-социологического исследования, такими, как социальный эксперимент, данные статистики отчётные документы и т.п., которые также используются при обобощении результатов таботы по борьбе с преступностью в стадии предварительного расследования. Подовный синтез методов исследования является важнейшим критерием объективности."

[55]Gerovitch, 2002, 161.

[56]Gerovitch, 2002, 161.

more trustworthy[;] ... exact knowledge seemed an equivalent of moral truth; an equals sign was put between honesty and mathematics."[57]

Historian Theodore Porter analyzes the objectivity associated with quantitative measures and its relationship to institutional politics and epistemic authority. He argues that numbers are a technology of distance, creating distance from local and personal contributions and overcoming distance through their consistency and standardization. This mechanical objectivity – knowledge based on a consistent set of rules and thus free of individual or human contribution – often emerges as an "adaptation to the suspicions of powerful outsiders" and provides institutions an alternative to placing trust in individuals.[58] In the context of mechanical objectivity, epistemic authority derives from compliance with rules and quantitative procedures, rather than tacit expert judgment. Even though quantitative methods "provide no panacea,"[59] they are often considered trustworthy even when nobody vouches for their validity.[60] The push for replacing interpersonal trust with quantification and mathematization, Porter argues, is thus a symptom of weakness, vulnerability, and distrust of institutions.[61]

Porter shows that the spread of mechanical objectivity in institutional settings is often forced on the elites whose authority it threatens, rather than embraced. For example, Porter describes how accountants in the US were forced to adopt quantitative rules to standardize their practice after the Depression sparked intense suspicion and political scrutiny of accountants' discretion by the newly formed Securities and Exchange Commission. Legal cybernetics departs from the case studies in his book in that Soviet criminologists were themselves eager to seek out the mechanical objectivity that results from the consistent application of mathematical methods (in this case, cybernetics) as a strategy for excising the subjectivity and bias that had rendered their field's expertise illegitimate and untrustworthy during the Stalinist period. Mechanical objectivity served not only as a rhetorical proxy for absolute objectivity, but also a strategy to bolster the scientific status and authority of Soviet criminology.

[57] Vail' and Genis, 1996, 100; quoted in Gerovitch, 2002.
[58] Porter, 1995, 89.
[59] Porter, 1995, 5.
[60] Porter, 1995, 8.
[61] Porter, 1995, ix–xi.

### 3.2.2 Criminology: An Art or a Science?

> It is our aspiration that justice, humanity, inevitability, truth and all other legal concepts will be grounded in indisputable data and be therefore as exact as the fields of mathematics, physics, and chemistry.[62]
> –Vladimir Kudriavtsev (Soviet law professor; prominent figure in legal cybernetics), 1965

Legal cybernetics emerged as an embodiment of this aspiration. A 1961 article in the periodical volume *Cybernetics in Service of Communism*[63] made a claim that would be echoed in numerous legal cybernetics publications for decades to come: that the successful application of cybernetics to other areas of science left no doubt that it could similarly be applied to solve any number of legal and criminological problems.[64] The extent to which criminology and law were conducive to mathematization was intimately tied up with the scientific status of the field. On Soviet interpretations of Marxist theory, the more mathematics criminology incorporated, the more scientific it was, though the applicability of mathematical methods to criminology was frequently debated. In spite of its uncertain scientific status, the popularity of legal cybernetics rapidly grew. In 1971, just a decade later, legal cybernetics would become a required course for law students at Moscow State University, the largest and most prestigious university in the Soviet Union.[65] Legal cybernetics, with its mechanical objectivity and aura of absolute objectivity, promoted the credibility and status of criminology.

Cybernetics is a notoriously slippery term, and legal cybernetics inherited these ambiguities in definition to a fault. A 1962 publication on legal cybernetics follows Wiener's definition in describing cybernetics as the branch of science studying control processes in machines, living organisms and society.[66] It is clear, however, from the papers that were ultimately published under the subject 'legal cybernetics' that the term was actually used to refer to the application of any methods even remotely associated with cybernetics, including the use of computers, formal logic, and information theory to broadly legal matters. This suggests

[62] Kudriavtsev, 1965. Original quote: "Мы хотим, чтобы справедливость, гуманность, неотвратимость, истина и все прочие юридические категории стали столь же точными, основывались бы на таких же бесспорных данных, как это имеет место в категориях математики, физики, химии."

[63] This was the oldest publication with the keywords "cybernetics" and "legal," "law," or "criminology" in the two largest library systems in Russia.

[64] Andreev and Kerimov, 1961, 234; Kerimov, 1962 says something almost identical.

[65] Polevoi and Shliakhov, 1977, 6; Semukhina, 2017 argues that the scientific status of criminology increased in this period.

[66] Kerimov, 1962, 99.

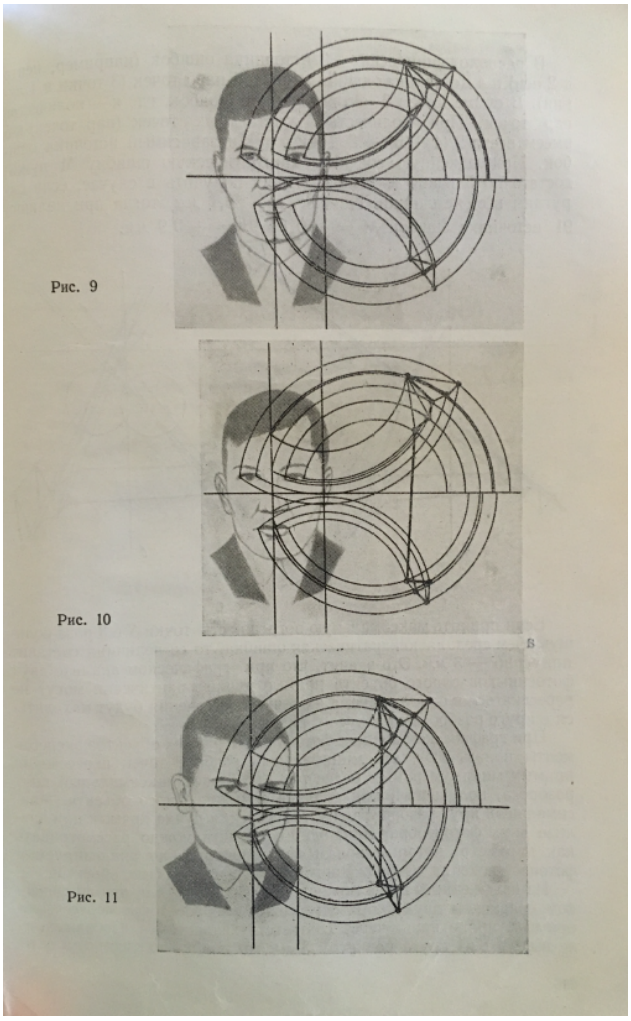


Figure 13: An image from a legal cybernetics publication on analyzing images of faces in forensic evidence (Polevoi, 1970). Photographed at the Russian State Library in Moscow.

that 'legal cybernetics' was less a distinctive discipline than a buzzword.

For instance, the introduction to a 1970 collection of research papers titled *Legal Cybernetics*[67] describes legal cybernetics as the "widespread use of computer technology, information theory and mathematical methods ... to facilitate the activities of legal institutions and, above all, the courts, the prosecution authorities and protectors of public order, forensic examination, and law enforcement and law-making activities of public authorities and admin-

[67]In Russian, *Pravovaia Kibernetika*.

istration."[68] To say this is a broad definition is an understatement. Indeed, the potential applications of cybernetics discussed in these early publications were seemingly boundless. Some authors argued that the organization of large amounts of information in the legal system could be aided by cybernetics.[69] Others pointed to the potential value of logical modeling in simplifying laws or finding logical inconsistencies in them. Still others claimed out that cybernetics could be valuable in assessing or creating logical models of forensic evidence in court proceedings,[70] or even in analyzing similarities between signatures[71] or images of faces[72] in forensic investigations (Figure 13).

Perhaps the most important application of cybernetics to legal sciences, however, was to study and theorize about the causes of crime. In a volume on legal cybernetics, multiple authors pointed out that mathematical and logical modeling of crime as a sociological phenomenon was important in criminology because of the impossibility of experimental studies.[73] In this vein, several institutes and laboratories for mathematical and empirical studies of crime were established in the 1960s. The All-Union Institute for Study of Causes of Crime and its Prevention,[74] established in 1963 and headed by Kudriavtsev, collected empirical data on the causes of crime,[75] and in 1966, the Central Research Institute of Forensic Expertise[76] created a laboratory for adopting cybernetic methods in criminology, criminal statistics, forensics, and in the organization of legal information.[77]

Nevertheless, the extent to which criminology and the legal sciences were conducive to mechanical methods was a topic of vigorous debate in the 1960s. At the start of their short, propagandic book, *Cybernetics in the Fight Against Crime* (1964) (Figure 14), Kudriavtsev

[68] Shliakhov, 1970, 6:"Правовая кибернетика основывается на широком использования электронно-вычислительной техники, теории информации и математических методов. Её задча - облегчить деятельность юридических учреждений и прежде всего судов, органов прокуратуры и охраны общественного порядка, судебной экспертизы, правоприменительную и правотворческую деятельность органов государственной власти и управления."

[69] E.g., Andreev and Kerimov, 1961, 236.

[70] E.g., Shliakhov, 1967, 18–19, and Trusov, 1967.

[71] E.g., Zhuravel' et al., 1970.

[72] E.g., Polevoi, 1970.

[73] Bluvshtein, 1970, 105; Gavrilov and Kolemaev, 1970; Chugunov and Gorskii, 1967.

[74] Всесоюзный институт по изучению причин и разработке мер предупреждения преступности.

[75] Universitet Prokuratury Rossiiskoi Federatsii, 2018; Unspecified, 2007.

[76] Центральный научно-исследовательский институт судебных экспертиз.

[77] Akademiia Nauk SSSR: Nauchnyi sovet po kibernetike [USSR Academy of Science: Scientific Council on Cybernetics], 1967, 6.

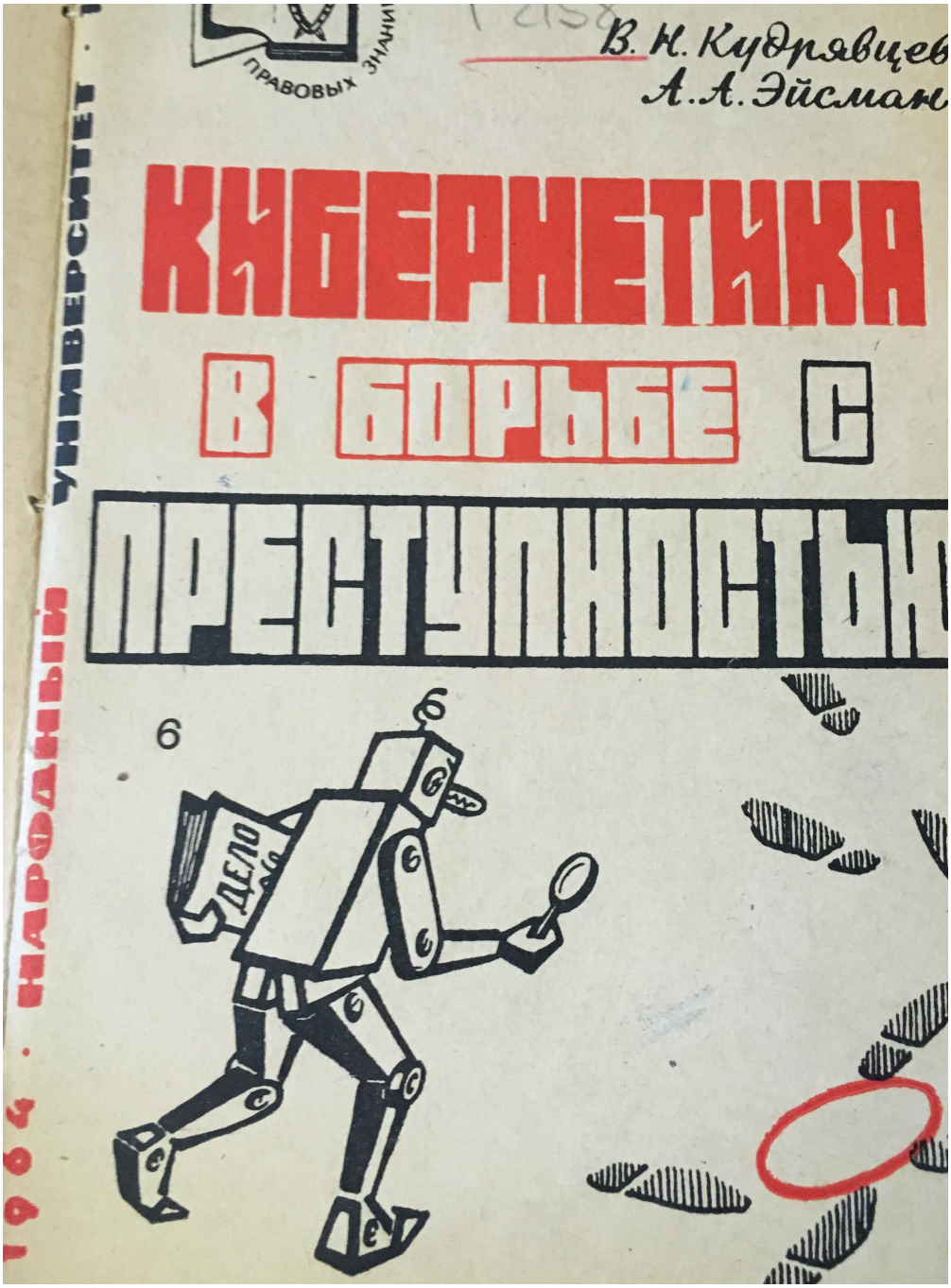


Figure 14: The cover of Kudriavtsev and Eisman's 1964 book "Cybernetics in the Fight Against Crime." Photographed at the Russian State Library in Moscow.

and his colleague Aleksei Eisman recount a conversation they had with a fellow lawyer named Olga Ivanovna, who was skeptical about the prospect of applying cybernetics to study crime:

> [Olga Ivanovna:] "Cybernetics and the fight against crime? What could you have possibly found in common between these things? ... Intuition and cybernetics are incompatible. How do you expect to apply your mathematical formulas in a field where everything is based on human impressions, experiences, evaluations, different opinions and judgments? ... The very nature of legal work precludes the possibility of mathematical precision in criminal decisions."
> [Kudriavtsev and Eisman:] "Do you mean to say, Olga Ivanovna, that jurisprudence is not a science at all? After all, if it does not lend itself to precise methods, then isn't that the only possible conclusion? ... We must free ourselves from the misconception that the social sciences are a matter of subjective judgments, devoid of an objective basis. Legal science and practice should be based not on impressions and subjective opinions, but on facts, on precise, well-founded reasoning, on the knowledge of objective laws of nature. ... Cybernetics can improve the precision of our work and thus help us in this fight [against crime]."[78]

Kudriavtsev's concerned colleague was not alone in questioning the applicability of cybernetics to criminology. In a 1965 book, *On the Possibility of Using Cybernetics Methods in Law,* one Soviet lawyer argues that socialist legal proceedings cannot be predetermined in advance by a finite number of logical and mathematical formulas, and that it is impossible for "cybernetic machines" to simulate the cognitive process of evaluating evidence because they in principle do not have access to one of the main elements of evidence assessment: faith or disbelief in the veracity of evidence.[79] Still another author points out that computers use formal logic and abstract categories, which must be constructed out of complex events containing numerous connections and complicated dimensions, information that would be lost if converted to a formal language.[80]

[78] Kudriavtsev and Eisman, 1964, 3–9: [Ольга:] Кибернетика и борьба с преступностью? Что вы нашли между ними общего? Интуиция и кибернетика – несовместимые вещи. ... Как вы мыслите применить ваши математические формулы в такой области, где всё основано на человеческих впечатлениях, переживаниях, оценках, на разных взглядах и суждениях? ... Сама природа юридической работы исключает возможность математической точности в решениях. [Кудрявцев и Эйсман:] Не хотите ли вы этим сказать, Ольга Ивановна, что юриспруденция это вообще не наука? Ведь если она не поддаётся точным методам исследования, то вывод должен быть именно таков? ... Мы должны освободиться, если они есть ещё у кого-либо, от неправильных представлений, будто общественные науки суть науки гипотез, науки субъективных суждений, лишенных объективной основы. Юридическая наука и практика должны быть основаны не на впечатлениях и субъективных мнениях, а на фактах, на точных, обоснованных рассуждениях, на познании объективных закономерностей. ... Киберентика может повысить точность нашей работы и тем самым помочь нам в этой борьбе."
[79] Knapp, 1965, 138
[80] Pekelis, 1986.

The majority position, however, denied that cybernetics could not *in principle* formalize and, eventually, automate many criminal and legal matters, even if it was not currently feasible. A 1967 publication, "The use of cybernetics and computers in sociological studies into the causes of crime and the personality of the criminal," argues that cybernetics is not only applicable, but would also improve the objectivity and quality of such studies.[81] The editors of a 1970 volume on mathematical models of crime argue that "in the objective world there is no forbidden zone where the quantitative and structural methods of modern mathematics would not be applicable,"[82] and a publication on the application of cybernetics to analyze forensic evidence argues that "if human thinking proceeds in a natural, not supernatural way, then, like any natural process, it is knowable and, therefore, ultimately formalizable. ... The task of scientific knowledge is precisely to transform informal things into formal ones."[83]

These quotes may give the impression that quantitative methods were applied to criminology indiscriminately or carelessly. This is not the case – on the contrary, even the papers arguing for the mathematization of criminology at least claim to be aware of the dangers of using of quantitative methods under false assumptions. Indeed, Western criminology was often harshly criticized by Soviet criminologists on precisely these grounds.[84] Eisman, a frequent contributor to applications of logical models to criminology, complains that "bourgeois theorists suggest a purely mechanical evaluation of evidence using various mathematical tools, without taking into account the specifics of this field. Such a crude approach to this extremely complex problem has not been eradicated in our time."[85] It is ironic that prominent criminologists like Eisman's co-author, Kudriavtsev, were lauded and rewarded for their adoption and promotion of legal cybernetics, even while these mathematical methods served

[81] Chugunov and Gorskii, 1967, 150–151.

[82] Gavrilov and Kolemaev, 1970. Original quote: "Происходящий на наших глазах процесс математизации знаий свидетельствует о том, что в объективном мире нет той запретной зоны, того "островка," где были бы неприменимы количественные и стрктурные методы современной математики."

[83] V. M. Glushkov, "Cybernetics and cognitive work," 1965; quoted in Trusov, 1967, 32. "Если мышление человека происходит естественным, а не сверхестественным путём, то, как и всякий естественный процесс, оно является познаваемым и, следовательно, в конечном счёте формализуемым. ... Задача научного познания как раз и состоит в превращении неформальных вещей в формальные ."

[84] Selivanov et al., 1978, a textbook on criminology, writes that one of the tasks of Soviet criminology was to expose these false scientific concepts in 'bourgeois' criminology.

[85] Eisman, 1967, 164: в ряде работ буржуазных теоретиков предлагалось чисто механическое применение в оценке субедбых доказательств некоторых математических аппаратов, без учёта специфики указанной области. Вульгарный подход к этой чрезвычайно сложной проблеме не изжит и в наше время.

to obscure and promote the value-laden aspects of their work.

## 3.3 The Objective Side of Crime

Kudriavtsev wrote several foundational texts on the causes of crime, including the *Objective Side of Crime* and *Causation in Criminology*.[86] His framework for thinking about the causes of crime, which focuses on criminal personality and crime as a social phenomenon, was influential within criminology and was often appealed to in legal cybernetics papers on causal models of crime. This framework is one of the most prominent examples of early Soviet attempts to formalize criminology using causal models and is illustrates the limits of mechanical objectivity. After outlining the main features of this framework, I discuss one of its most clearly value-laden assumptions: the exclusion of economic causes of crime in causal variable choice, which Kudriavtsev's framework inherits from its broader Marxist-Leninist framework, and which legal cybernetics models that use it inherit in turn. I argue that the mechanical objectivity that resulted from the adoption of a formal framework rhetorically promoted the methods' absolute objectivity and reinforced Marxist-Leninist values about crime.

### 3.3.1 Crime and its Causes

In his foundational 1960 book *The Objective Side of Crime*, Kudriavtsev emphasizes that crime is a social phenomenon. He writes that crime has two dimensions, 'subjective' and 'objective': like any form of human behavior, he writes, crime not only has mental (subjective) content, but also is expressed in external (objective) types of behavior and action/inaction, and causes changes in the external world.[87] These interact in feedback loops, a common concept in cybernetics.

The subjective side of crime consists in the motives, goals, and "personality of the criminal."

[86] In Russian, "Ob'ektivnaia Storona Prestupnosti" Kudriavtsev, 1960 and "Prichinnost' v Kriminologii" Kudriavtsev, 1968.
[87] Kudriavtsev, 1960, 8.

The objective side of crime has three components. The first is the socially dangerous *action* (or inaction), which is partially subjective (since it arises partly due to personality) and partially objective (since it arises partly due to the method, place, time, and circumstances the crime was committed in).[88] The second component is the *causal connection* between the act and the criminal result, which can depend both on the personality of the criminal and the circumstances of the act.[89] The third is the *criminal aftermath*, or the "socially harmful changes" that are a result of the crime.[90] "It would be wrong," Kudriavtsev writes, "to limit ourselves to studying the psychological aspects of crime, forgetting that it is the objective side of crime that is the real embodiment and expression of the subject's goals and intentions, and that it ultimately shows the main social characteristic of the crime – its social danger."[91]

The most important task of Soviet criminology, Kudriavtsev writes in his 1968 book *Causation in Criminology*, is to understand the causes of crime in socialist society, for this is the only way that crime can be prevented and eradicated.[92] In Marxist-Leninist philosophy, the general concept of causality is the same for all areas of knowledge: a causal relationship is a relationship between phenomena in which one or more interacting phenomena (cause) generates another phenomenon (effect), and these causal connections exist in the external world, independently of human consciousness.[93]

Criminality in society consists in statistical patterns that manifest due to causal laws affecting individuals.[94] A person's life situation on its own, however, does not result in a criminal act – it arises in combination with their personality, expressed through their subjective interests, perspectives, habits, psychological characteristics, and other individual traits.[95] A person's personality, in turn, is not something they are born with, but something that forms as they mature, through interaction between their external social environment

[88]Ibid, 10.
[89]Ibid, 11.
[90]Ibid, 11: "вредных изменний в объекте посягательства."
[91]Kudriavtsev, 1960, 9: "Неправильно было бы ограничиваться психологическим аспектом, забывая отом, что именно объективаная сторона преступления является реальным воплощением и выражением во вне целей и намерений субъекта и что именно в ней в конечном счёте проявляется основное социальное свойтсво преступления – его общественная опасность."
[92]Kudriavtsev, 1968, 3.
[93]Kudriavtsev, 1960, 185–186.
[94]Kudriavtsev, 1968, 7.
[95]Kudriavtsev, 1968, 15.

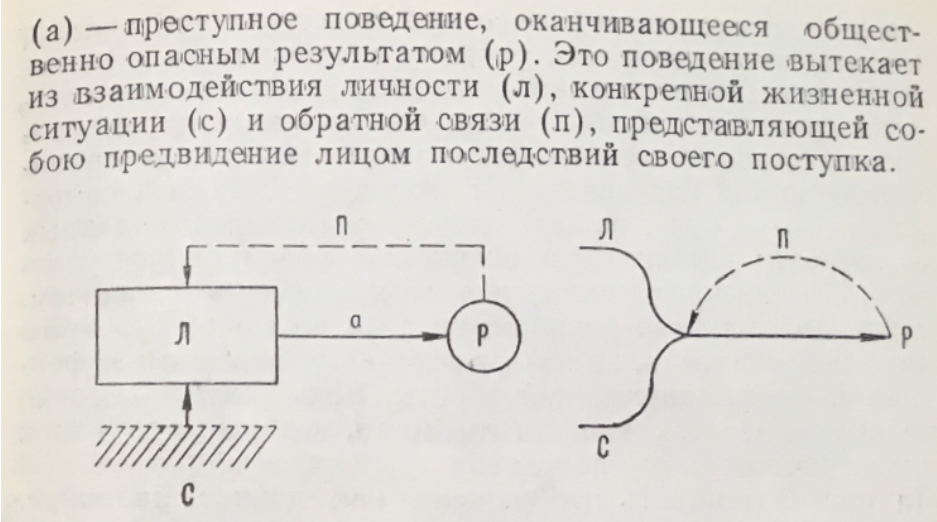


Figure 15: "(а) – criminal behavior, resulting in a socially dangerous result (р). This behavior arises from the interaction of their personality (л), specific life situation (с) and backward causation (п), which represents the person's foreseeing of the consequences of their action." Kudriavtsev (1968), 17. Photographed at the Russian State Library in Moscow.

and "individual psycho-physiological data."[96] In sum, the causes of crime stem from the "personality of the criminal," the particular life situation in which they find themselves, and the interactions between these things.[97] These causal connections are not necessarily unidirectional or acyclic – in the process of committing the crime, the objective side of crime affects the subjective side in a feedback loop (see Figure 15). For instance, committing a crime might cause the perpetrator to self-reflect and change for the better, or it may cause them to commit more crimes.[98] Man, Kudriavtsev writes, is a complex system and "at the highest level self-regulating, self-sustaining, self-recovering, self-correcting and even self-perfecting."[99]

Kudriavtsev writes that the commission of a crime is the result of a threshold being crossed by the sum of these different factors in the objective and subjective sides of crime; he frequently discusses the contributions of these factors to the likelihood of the commission of a

[96]Ibid, 19, 53.
[97]Ibid, 16.
[98]Kudriavtsev, 1960, 19.
[99]Kudriavtsev, 1968, 16: "Человек – это сложнейшая система, 'в высочайшей степени саморегулирующая, сама себя поддерживающая, восстановливающая, поправляющая и даже совершенстовующая'." (Kudriavtsev is quoting Pavlov, whose work is often referred to in Soviet cybernetics.)

crime in terms of information theory.[100] Because Marxist theory emphasizes the importance of free will and is opposed to determinism of human action, however, Kudriavtsev stresses that committing a crime is, at bottom, a conscious act of a person and is always guided by the mental properties of the subject – by their will.[101][102]

Nevertheless, the greatest contributing causes of the majority of offenses, Kudriavtsev writes, are "defects of upbringing, deficiencies in the *domain of moral personality formation*."[103] He advocates for the collection of extensive "moral statistics" about the "antisocial events" that are closely tied to crime, including drunkenness and alcoholism, addiction, breaking of social order, child neglect, and so on.[104] He reports that 80% of cases of crimes by minors are connected to familial neglect, and drunkenness is involved in many crimes.[105] Consequently, he recommends prophylactic measures that involve educating and "raising the culture" of people: the establishment of social clubs, giving people time off work in special vacation homes ('dom otdykha'), sports groups, playgrounds for children, and so on.[106] Similar rationale – namely, the causal connection between alcohol, "moral personality," and crime – led in part to Gorbachev's 1985 anti-alcohol campaign two decades later, which raised the price of alcohol, starkly decreased its production, and instituted social measures to discourage drinking (Figure 16).[107]

### 3.3.2 Hidden Values, Reinforced

One might expect, given the emphasis on individual social factors and absence of economic factors in this framework, that wealth inequality and poverty were at a minimum under socialism in the Soviet Union. This, of course, was far from the case – lack of availability

[100]Kudriavtsev, 1968, 130–132.
[101]Ibid, 12.
[102]The challenge of reconciling Marxist views on free will and determinism in dialectical-materialism is a complicated issue in Marxist theory. Kudriavtsev emphasizes that these circumstances represent only the *possibility* of a future crime, and a person "can be responsible only for those intentions and desires that were actually realized in criminal behavior," or when they could not have acted otherwise.
[103]Kudriavtsev, 1967, 19; emphasis his. "Непосредственной причиной большинства правонарушений являются дефекты воспитания, недостатки в области нравственного формирования личности."
[104]Kudriavtsev, 1967, 5,7; "создание так называемой 'моральной статистики'."
[105]Ibid, 17.
[106]Kudriavtsev, 1967, 19.
[107]Bhattacharya et al., 2013.


Figure 16: A 1986 Soviet poster discouraging alcohol consumption. "A shot, a glass, a bottle, here comes the hooligan / Indulging in drinking is dangerous: there is but one step from drunkenness to crime." Artist unknown.

of food and clothing, long lines for consumer goods, and inadequate housing were all major challenges in the Soviet Union at this time; even by the standards of 1967 Soviet economists, over half the population in 1965 lived in poverty, with a substantial wealth distribution gap.[108] It is plausible that an economic factor such as poverty could be the common cause of, for instance, high rates of alcohol consumption and criminal behavior such as theft. So why weren't economic factors considered in Kudriavtsev's framework? Halfway through his book on causation, he addresses this point:

> Under socialism, the main cause of crime characteristic of exploitative societies was eliminated: the exploitation of man by man, the need and poverty of the working masses, though there is still incomplete satisfaction of the material needs of the population. *These and other economic reasons, of course, do not cause people to have a direct desire to commit crimes.* Such an understanding of the social causes of crime would be superficial and incorrect. ... *The percentage of crimes committed due to material insecurity is very small.*[109]

The exclusion of economic variables, then, appears to rest on a strong assumption: that the "incomplete satisfaction" of individual economic needs has no significant influence on the commission of crimes in the Soviet Union. This could in principle hold in an ideal socialist system, but Kudriavtsev presents no evidence or argument in support of this assumption. Many contemporary criminologists reject this assumption, although the relationship between poverty and crime is admittedly complex.[110] Nevertheless, this value-laden assumption is prevalent and unquestioned in legal cybernetics publications that use this framework.

For instance, one such publication writes that the complex processes underlying crime can be elucidated using logical models; they demonstrate this with a causal diagram (Figure 17) relating low education, which is closely connected to "poverty of the spiritual world," which is "undoubtedly one of the common causes of alcoholism," which, in turn reinforces the moral decline of the personality and deepens its spiritual impoverishment, resulting in "a situation fraught with a real danger of criminal acts."[111] The value of models like these, the

[108] Matthews, 1986, 10–13.

[109] Kudriavtsev, 1968, 73–74; emphasis mine. "При социализме устранёна главная причина преступности, свойственная эксплуататорскому обществу: эксплуатация человека человеком, нужда и нищета трудящихся масс. Однако ещё имеет место неполное удовлетворение материальных потребностей населения. Эти и другие экономические причины, конечно, не вызывают у людей непосредственного стремления к совершению преступлений. Такое понимание действия социальных причин преступности было бы поверхностным и неверным. ... Процент преступлений, совершаемых из-за материальной необеспеченности, весьма невелик.

[110] For a discussion, see Sharkey et al., 2016.

[111] Bluvshtein, 1970, 113–114. "Возьмём соотношение нескольких важных для криминологии факторов,

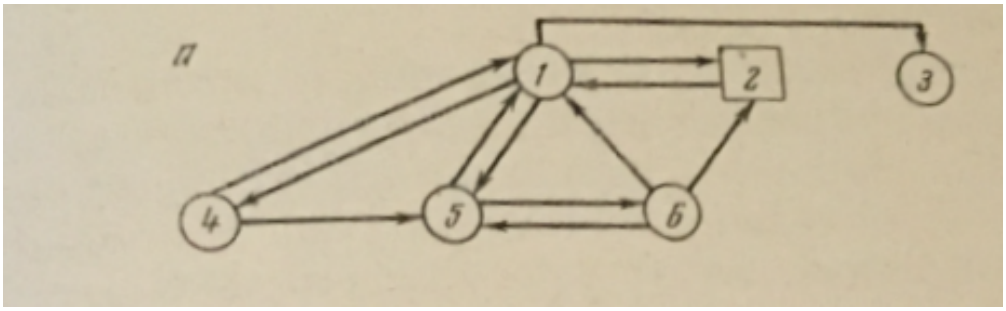


Figure 17: A diagram that "convincingly demonstrates" the causal relationship between various factors: (1) personality, (2) local social sphere, (3) criminal act, (4) low education, (5) poverty of spiritual life, (6) alcoholism. From Bluvshtein (1970), 113. Photographed at the Russian State Library in Moscow.

author writes, is to show the possible events that could lead to crime, and adds that future work will involve evaluating the strength of these causal connections by adding correlation coefficients. Consistent with Kudriavtsev's framework, economic considerations are absent from the model.

Other studies aim to tabulate data about crimes and analyze them using computers in order to generate statistical data for sociological models. As input fields, one study includes education level, age, profession, party membership, type of crime, way in which the crime was committed, criminal record, and state of health.[112] A later publication working on a similar project includes 22 input fields, including additional variables like geographical location, gender, number of children, employment, nature of past crimes and criminal sentences, and "danger of recidivism."[113] The aggregated data are then analyzed using statistical methods, such as correlation analysis and factor analysis. In both cases, variables relating to economic or material difficulty, beyond employment information, are not considered.

например, низкого образования личности и её низкого культурного уровня, алкоголизма и преступного поведения. Перед нами возникает граф связей, который убедительно демоснтрирует взаимодействие всех этих явлений: с ниским образованием (4) тесно связана бедность духовного мира (5), которая, весспорно, является одной из распространённых причин алкоголизма (6). В свою очередь алкоголизм, с одной стороны, оказывает обратное влияние, исиливающее моральное падение личности, углубляет и закрепляет её духовное обеднеине, с другой – ведёт к ситуации, чреватой реальной опасностью преступных деяний."

[112]Chugunov and Gorskii, 1967, 156–160.

[113]Polevoi and Shliakhov, 1977, 187–190.

The closest any of the surveyed legal cybernetics applications comes to considering economic factors as causal variables in their models is a paper describing simple linear models of crime based loosely on Kudriavtsev's framework. In addition to the standard social and demographic causal factors of crime, the authors discuss measuring the contribution of the "material well-being" of the population on the overall number of crimes:

$$\bar{R} = \frac{\alpha}{Z} + \epsilon$$

where $\bar{R}$ is the number of overall crimes, $Z$ is average material well-being (income in rubles per capita), $\alpha$ measures the contribution of average material well-being to the crime rate, and $\epsilon$ is measurement error. Of course, this is intended as a way of estimating the relationship of a single factor (average income) to the absolute number of crimes, rather than as a way of stratifying the population in their causal model; the authors remain silent on the effect of individual material well-being and wealth inequality on crime and thus have the same value-laden assumption as the papers discussed above.

## 3.4 Taking Stock

What did cybernetics contribute to this framework?

The mechanical objectivity of the cybernetics bandwagon may have raised the authority of Soviet criminology, but it did not bring with it the absolute objectivity that post-Stalinist criminologists sought. Instead, an emphasis on formal and quantitative methods served at times to produce the illusion of absolute objectivity in criminology: it reinforced existing dogmatic values in the field, such as the exclusion of economic causes of crime. This illustrates one way that the mechanical objectivity of quantitative methods falls short of excising ideological values – the increasing addition of formalism and quantitative metrics to Kudriavtsev's framework did not alter its underlying causal variable choice. Rather, it served to obscure and more deeply entrench this inherited value-laden assumption – an instance of domain distortion.

Nevertheless, packaging criminology in the palatable wrapper of cybernetics was an

effective rhetorical strategy for raising the authority of criminology, and was lauded as such. Kudriavtsev was awarded numerous state honors, including the Order of Lenin, one of the highest state honors,[114] and in 1984 Kudriavtsev, along with four other prominent criminologists, received the State Award of the USSR, a prestigious award given to honor exceptional scientific work.[115] Olga Semukhina, a Soviet historian, argues that: "For many criminology researchers, this award signified the final recognition of criminology as a legitimate science and attenuation from the label of pseudo- or harmful science that had been attached to it in the mid-1930s."[116] In 1985, Kudriavtsev became Vice-President of the Russian Academy of Sciences, headed for the first time in history by a criminologist.[117]

Rhetoric aside, it is unclear what the application of cybernetics contributed to criminology. Talk of cybernetics in criminology eventually disappeared; neither modern criminology textbooks[118] nor course lists[119] make any mention of cybernetics, and periodicals published about cybernetics and crime tapered out by the 1980s.[120] In 1987, the "All-Union Institute for Study of Causes of Crime and its Prevention" was renamed the "All-Union Research Institute for the Problem of Strengthening Law and Order."[121] Today, explanations of crime put forward by Russian criminologists continue to be at odds with those in Western countries, and until very recently, rates of violent crime in Russia surpassed those in the Western world.[122]

In Kudriavtsev's 2014 obituary, a former student of his remarks:

> Crime stifles the country, stifles the economy, stifles democracy, stifles the lives of many people. ... Scientific approaches to the fight against crime have been ignored by authorities and law enforcement agencies. We still have not developed evidence-based approaches, programs, or laws to counter crime. And this science, with the departure of V. N. Kudriavtsev, is on its last legs. It needs serious support.[123]

[114]Luneev, 2014, 50–53.
[115]Kudriavtsev and Eminov, 1997.
[116]Semukhina, 2017, 424.
[117]Luneev, 2014, 50.
[118]E.g., textbook on Criminology from 2004 and 2006 have no mention the word "cybernetics."Kuznetsova and Luneev, 2004, Malkov, 2006.
[119]It was absent on the long list of courses here: http://www.law.msu.ru/node/20366
[120]Based on my keyword searches in the two largest Russian libraries.
[121]Universitet Prokuratury Rossiiskoi Federatsii, 2018.
[122]See Goertzel et al., 2013 for a discussion of crime trends in Russia.
[123]Luneev, 2014, 54. "[П]реступность душит страну, душит экономику, душит демократию, душит жизни многих людей. ... [К] великому сожалению, правоохранительная система продолжала работать

Still, the story of legal cybernetics is not entirely black-and-white. In personal conversation, Slava Gerovitch, Soviet science historian, reminded me to keep in mind the bigger picture:

> To what extent the scientists overstated the objectivity of their algorithms is an interesting question, but we have to realize that they did so in the context of a very fierce struggle against the Stalinist legacy in Soviet science. There is bias, and then there is bias – theirs was a struggle against even larger bias. ... We should deconstruct them but understand that they were struggling against a greater evil, in a sense.[124]

Nevertheless, the strategies of attaining objectivity and authority through quantitative methods – both in Soviet and contemporary US contexts – carry, as Shannon prophetically understated, "an element of danger."[125] The findings in this chapter serve as a cautionary tale for the contemporary evidence-based drive to replace human judgment with quantitative crime prediction in US penal decision-making. It is easy to recognize the political aims of formal models in sociotechnical systems we are external to; our own systems require the same scrutiny.

---

главным образом на желаемые бумажные показатели и снижать учтенный уровень преступности и числа заключенных в местах лишения свободы не минимизацией криминогеннсти в стране и профилактикой преступлений, а изменением уголовного законодательства путем перевода традиционных преступлений в дисциплинарные и административные проступки. Научные подходы борьбы с преступностью властями и правоохранительными органами игнорировались. У нас до сих пор не выработано научно обоснованных подходов, программ и законов по противодействию преступности. И эта наука с уходом В.Н. Кудрявцева совсем осиротела. Она нуждается в серьезной поддержке."

[124] Quote from a Skype interview with Slava Gerovitch, a historian of Soviet cybernetics and science, on October 2nd, 2018.

[125] Shannon, 1956.

## 4.0 Judicial Resistance to a Risk Assessment Instrument

As we have seen in chapters 1, 2, and 3, algorithmic decision-making in the public sector – criminal law, policing, education, and public benefits – is often introduced as a reform measure intended to address institutional inefficiency and problems of legitimacy (Porter, 1995). Throughout the dissertation, we have repeatedly encountered claims that recidivism risk assessment instruments are more objective than human judgment and are an 'evidence-based' strategy for increasing consistency in sentencing, reforming cash bail, and reducing mass incarceration. At the same time, we have seen the ways in which risk assessment instruments can be value-laden and reinforce existing social values. In practice, however, the value-ladenness of risk assessment instruments is strongly mediated by the people that use them. In this chapter, I present novel empirical research about this judge-algorithm interaction, which shows that algorithm-centric reforms can simply add another layer to the sluggish, labyrinthine machinery of bureaucratic systems and are met with internal resistance.

Consider the Sentence Risk Assessment Instrument, a recidivism risk assessment instrument implemented in Pennsylvania in 2020.[1] The actuarial tool uses demographic factors such as age and number of prior convictions to estimate the risk that an individual will "reoffend and be a threat to society" – that is, be reconvicted within 3 years of release from prison (Pennsylvania Commission on Sentencing, 2019a). It was adopted on the premise that it would help judges identify candidates for alternative sentences, despite public criticism that the tool would exacerbate racial biases in sentencing (ACLU of Pennsylvania, 2019; Coalition to Abolish Death by Incarceration, 2019; Sassaman, 2019). Through a community-informed interview-based study of 23 criminal judges and other criminal legal bureaucrats in Pennsylvania, however, I find that judges overwhelmingly ignore the Sentence Risk Assessment Instrument, which they disparage as "useless," "worthless," "boring," "a waste of time," "a non-thing," and simply "not helpful."

Proponents and critics of risk assessment instruments alike tend to focus on the algorithms' technical aspects, such as their ability (or inability) to meet benchmarks of accuracy and

[1] I discuss the audit of this instrument in chapter 1.

algorithmic fairness, their proprietary nature, their predictive features, and their opacity. Many studies also assume, with no empirical basis, that bureaucrats such as judges, police officers, and government workers are prone to relying uncritically on predictive instruments – which are often advisory. Finally, studies and audits of risk assessment instruments are frequently conducted without the input or expertise of the communities most affected by, and most experientially knowledgeable about, the ongoing effects of their implementation – in the present context, communities impacted by incarceration.

This study takes a different approach to all three of these issues. It builds on the insights of previous empirical studies on the impacts of predictive technologies in the criminal legal system (Stevenson, 2018; Albright, 2019; Stevenson and Doleac, 2021; Sloan et al., 2018; Garrett and Monahan, 2020), ethnographic work on professional resistance in sociotechnical systems (Christin, 2017; Brayne, 2020), and input from community members to examine the impacts the Sentence Risk Assessment Instrument has had on judicial practice in Pennsylvania since its implementation in 2020.

My study has several key findings. I show that criminal court judges in Pennsylvania overwhelmingly ignore the recommendations of the Sentence Risk Assessment Instrument, a form of professional resistance to algorithmic systems. I argue, however, that this algorithm aversion cannot be accounted for by individuals' distrust of the tools or automation anxieties, per the explanations given by existing scholarship (Dietvorst et al., 2015; Brayne and Christin, 2020). Indeed, I find that even staunch supporters of risk assessment reform measures are critical of this particular tool. Instead, I identify three organizational factors that jointly explain the instrument's non-use: disparate county-level norms about pre-sentence investigation reports; alterations made to the instrument by the Pennsylvania Sentencing Commission in response to years of public and internal resistance; and problems with how information is disseminated to judges. My qualitative analysis thus provides an explanation of the Pennsylvania Sentencing Commission's own initial data analysis that the tool has had no impact on sentencing (Pennsylvania Commission on Sentencing, 2021), the inconsequential outcome of a decade-long process to satisfy a 2010 state legislative mandate for a sentencing risk assessment instrument. I also note two potential unexpected consequences of the tool's adoption: additional hidden labor for the probation department and longer pre-trial detention

times for defendants.

These findings shed new light on the important role of organizational influences on professional resistance to technology, which helps clarify one reason that algorithm-centric reforms can fail to have their desired effect. This study thus lends empirical support to a practical argument against the use of risk assessment instruments: they are resource-intensive and have not demonstrated positive on-the-ground impacts.

## 4.1 Background and Related Work

### 4.1.1 Risk Assessment Instruments and Human Discretion

Scholarship on predictive technologies in the public sector has exploded in recent years (Chouldechova, 2017; Brown et al., 2019; Fogliato et al., 2021; Levy et al., 2021; Akpinar et al., 2021; Meyer et al., 2022). The use of algorithmic decision-making in the criminal legal system has been particularly controversial, with reason. The claim that risk assessment instruments promote progressive criminal justice goals in practice is largely speculative – the few existing empirical studies suggest that risk assessment tools have had little to no impact (Stevenson, 2018; Sloan et al., 2018; Garrett and Monahan, 2020; Stevenson and Doleac, 2021) – and a vocal chorus of critics has stressed that such instruments could exacerbate racial disparities in pretrial, sentencing, and parole decisions because they base predictions on (and reproduce) structurally racist patterns in the US criminal legal system (Harcourt, 2008; Hannah-Moffat, 2013; Angwin et al., 2016).

To be sure, algorithmic bias is worth addressing seriously and can be reason alone to condemn the use of a particular instrument. But a key detail often neglected in discourse about risk assessment instruments and other public sector algorithmic systems is that their recommendations are advisory. Algorithmic systems are socially situated, interacting and entangling by necessity with people, institutional practices, and societal norms (Alkhatib and Bernstein, 2019; Mittelstadt et al., 2016; Pruss, 2021; Glaser et al., 2021). Individuals like judges and police officers make on-the-ground discretionary decisions – what Lipsky refers to

as 'street-level bureaucracy' (Lipsky, 1980) – that ultimately impact the lives of individual people, not the technical details of the algorithmic instruments on their own, and human judgment can interact with algorithmic decision-making systems in unexpected ways. The few studies of how risk assessment instruments are actually used have shown that judges differ widely in their adherence to recommendations and follow them inconsistently for different types of defendants (Stevenson, 2018; Garrett and Monahan, 2020; Stevenson and Doleac, 2021).

For example, human decision-makers can selectively follow algorithmic recommendations to the detriment of individuals already likely to be targets of discrimination. In Kentucky, a pretrial risk assessment tool – intended as a bail reform measure – increased racial disparities in pretrial releases and ultimately did not increase the number of releases overall because judges ignored leniency recommendations for Black defendants more often than for similar white defendants (Albright, 2019). Likewise, judges using a risk assessment instrument in Virginia sentenced Black defendants more harshly than others with the same risk score (Stevenson and Doleac, 2021).

In other contexts, human discretion can correct for algorithmic bias. In Pennsylvania, a recent study about racial bias in an algorithm that screens for child neglect showed that call screeners minimized the algorithm's disparity in screen-in rate between Black and white children by "making holistic risk assessments and adjusting for the algorithm's limitations" (Cheng et al., 2022) (see also De-Arteaga et al., 2020). Virginia's risk assessment instrument would have led to an increase in sentence length for young people had judges adhered to it; however, because judges systematically deviated from recommendations, some of the instrument's potential harms (and benefits) were minimized (Stevenson and Doleac, 2021).

Of course, another way that human discretion can interact with algorithms is not to interact with them. Algorithm aversion – the reluctance to follow algorithmic recommendations – is thought to arise from lack of confidence in algorithmic systems (Dietvorst et al., 2015); however, experimental research on algorithm aversion has focused on individual and algorithm factors, neglecting the role of social context and organizational factors (Mahmud et al., 2022). Sociological work shows that resistance to algorithms happens in contexts where individuals feel that their agency or power is being threatened by a new technology, as illustrated by Sarah

Brayne in her ethnography of LAPD officers using PredPol, as well as by Angèle Christin in her ethnography of prosecutors and judges using a pretrial risk assessment instrument (Brayne, 2020; Christin, 2017). Police officers and legal professionals alike felt threatened by how these new technologies could be used to surveil their performance and limit the role of their discretion, resulting in professional resistance to algorithmic systems in the form of adversarial data obfuscation – the process of manipulating a system's data to make it useless – and foot-dragging.

These dynamics can also intersect. In Virginia, judges had highly divergent attitudes toward (and literacy about) risk assessment and varied widely in whether and how they adhered to algorithmic recommendations (Garrett and Monahan, 2020). Understanding how these possible forms of human-algorithm interaction apply in a given case thus requires not only empirical research in a context of application but also attention to the social and organizational factors at play.

### 4.1.2 The Sentence Risk Assessment Instrument

In July 2020, criminal courts throughout Pennsylvania were instructed to begin consulting the Sentence Risk Assessment Instrument when sentencing crimes, with the aim of helping judges identify candidates for alternative sentences. The instrument applies to non-DUI defendants being sentenced following an open plea or trial. It generates a risk score of an individual's risk of recidivism based on demographic factors including age, gender, number of prior convictions, current conviction offense type, and prior juvenile adjudication.[2]

The tool recommends seeking 'Additional Information', typically a pre-sentence investigation report (PSI), for individuals with a low or high risk of recidivism "for whom additional information may assist the court in determining candidates for alternative sentencing" (Pennsylvania Commission on Sentencing, 2019a). The instrument is thus intended to influence a judge's decision to order a PSI for a given criminal defendant, with the presumption that information contained within PSIs will in turn influence a judge's decision to assign an alternate sentence. Currently, PSI-ordering rates in Pennsylvania vary substantially county-to-county,

[2]See chapter 1 for a detailed discussion of the tool's development and function.

as do the contents of the reports; one of the expected outcomes of the tool's adoption was thus to minimize county-level disparities in how often, and for which kinds of defendants, judges choose to order a PSI (Figure 18).

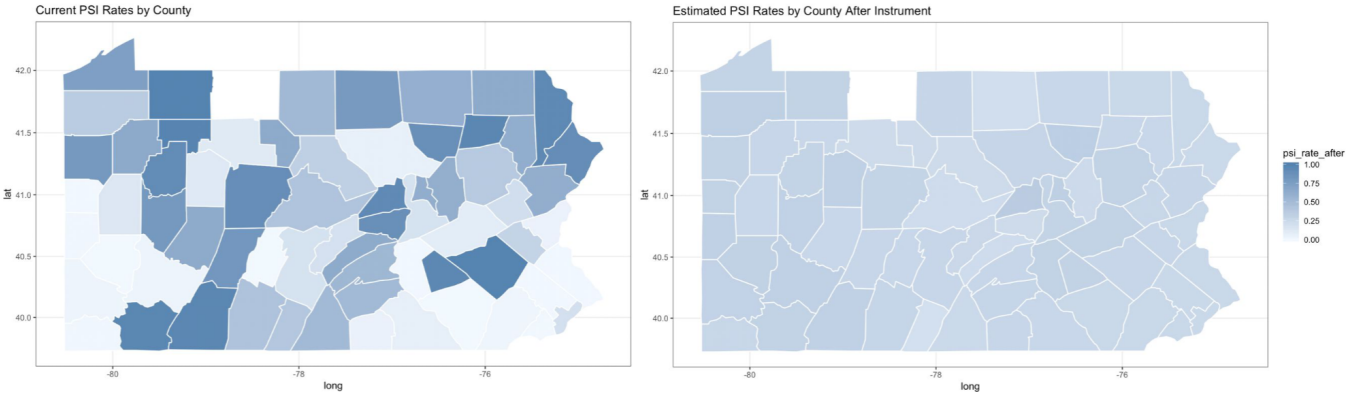


Figure 18: "Comparison of PSI Rates Before and After the Instrument," a figure from a third-party audit of the tool, which states that "if PSIs were to completed [sic] following the rate at which the instrument identifies high- or low-risk offenders, the PSI rates across counties will be more consistent" (Becerril et al., 2019).

The Pennsylvania Sentencing Commission – a legislative agency that advances "fairer and more uniform decisions at sentencing, resentencing, and parole" – was tasked with fulfilling a 2010 state legislative mandate to develop the instrument (Pennsylvania Commission on Sentencing, 2022). However, the Commission's members soon found themselves embroiled in controversy. From 2017–2019, the Commission received over 100 overwhelmingly negative public testimonies about the tool from sources including AI Now, the ACLU, high-profile academics, and local community organizations. Critics argued that the "racist tool" (ACLU of Pennsylvania, 2019) could "perpetuate the racial biases and stigmas inherent in our criminal legal system" (Coalition to Abolish Death by Incarceration, 2019). The instrument also met intense criticism from within the criminal legal system, particularly from probation officers, who argued that the Sentence Risk Assessment Instrument was "an unnecessary burden in time, effort, and resources" and would increase the workload of "already overwhelmed" county probation departments (County Chief Adult Probation and Parole Officers Association of Pennsylvania, 2019).

In informal interviews, Commission staff explained that they were legally required to

implement the legislative mandate despite these criticisms, lamenting that "from the start ... there has been no public support for the development and use of risk assessment at sentencing" (Pennsylvania Commission on Sentencing, 2019a). Commission staff, to their credit, engaged with the public through a transparent and iterative process of development, removing piece by piece the most controversial parts of the instrument and seeking further public comment each time. For instance, an earlier version of the instrument showed judges not only an individual's risk score but also a detailed risk distribution, indicating to the judge exactly where the defendant's numerical score falls relative to other individuals. The final version of the tool only shows judges a small text box with the words "Additional Information" (if the defendant is low- or high-risk), or "NA" (if the defendant is moderate-risk). The Commission also changed its outcome variable from rearrest to reconviction in response to public testimonies, which argued that arrest is not only a poor predictor of actual crime but also racially correlated due to racial profiling by police (Sassaman, 2018). The core concerns of the public and probation officers, however, went unaddressed – the tool was still implemented, and no additional resources were allocated to assist probation departments with the anticipated increase in ordered reports.

As part of the state's Evidence-Based Practices Strategic Plan, the Commission solicited an external review of the tool by Carnegie Mellon University researchers in 2019.[3] This audit focused on technical benchmarks of validity, accuracy, and fairness, and made several recommendations, including discarding the high risk category due to low accuracy; removing gender as a predictive factor; raising the high-risk category cutoff to increase its accuracy; and not deploying the violent crime risk scale component of the instrument due to an unacceptable level of false positives (Becerril et al., 2019). The Commission voted to follow the latter two recommendations. Notably, the audit does not consult relevant stakeholders or mention the tool's interactional effects with judicial discretion or other social factors, instead including projections (e.g., Figure 18) that assume complete uptake of the tool. Later in 2019, the Commission voted to adopt the tool, and the Sentence Risk Assessment Instrument was formally rolled out in July 2020.

[3]See chapter 1.

## 4.2 Methods

To understand how judges use and interpret the recommendations of the Sentence Risk Assessment Instrument, I conducted semi-structured interviews with 15 criminal court judges (Merriam and Tisdell, 2015), as well as unstructured interviews with three probation officers and four current and former Pennsylvania Sentencing Commission staff.

**Community Recommendations.** Drawing on feminist standpoint theory (Harding, 1992), I hired two justice-impacted individuals from the community organization Coalition to Abolish Death by Incarceration (CADBI) as consultants on the project in an effort to prioritize the affected community's interests and knowledge in developing my interview questions. One of the consultants was formerly incarcerated and the other works supporting incarcerated people and their families. Prior to conducting interviews, I met with both consultants to determine the scope of the project's research questions and later solicited their written and verbal feedback on a draft of an interview guide I produced based on this initial meeting; I compensated consultants for their time at a rate of $40/hour. One individual expressed concern that the new risk assessment tool would make judges more likely to ignore the humanity and personal circumstances of the people they sentenced and suggested gauging judges' awareness of this issue. Consultants also wanted to include interview questions about the personal nature and impacts of their sentencing decisions. Based on this feedback, I added questions to the interview guide to probe judges' concerns about the instrument and which personal factors judges consider in their sentencing decisions.

**Recruitment and Demographics.** I conducted interviews with judges from Allegheny, Philadelphia, Delaware, Dauphin, and York Counties. In each county, I initially recruited judges through emailed and physically mailed study invitations; I made follow-up phone calls and in-person visits regarding these invitations until I received a response or the time frame for my data collection passed. Other judges were recruited through snowball sampling from initial responders. I made an effort to select a sample of judges with variation (Weiss, 1995) across county, political orientation, favorability to risk assessment instruments, age, gender, race, and time served as a judge (see Table 1 for the results of a demographic survey given to interviewed judges; see Appendix B for the survey). Nevertheless, it is likely that the sample

Table 1: Features of interviewed judge population based on demographic survey (15 judges)

| **Sex** | **Frequency** | **%** |
|---|---|---|
| Male | 8 | 53.3% |
| Female | 7 | 46.6% |
| **Age** | | |
| 40–49 years | 1 | 6.7% |
| 50–59 years | 6 | 40% |
| 60–69 years | 5 | 33.3$ |
| 70–79 years | 2 | 13.3% |
| No response | 1 | 6.7% |
| **Years as a judge** | | |
| 0–2 years | 2 | 13.3% |
| 2–5 years | 2 | 13.3% |
| 5–10 years | 5 | 33.3% |
| 10–20 years | 5 | 33.3% |
| 20+ years | 1 | 6.7% |
| No response | 1 | 6.7% |
| **Race/Ethnicity** | | |
| White or Caucasian | 11 | 73.3% |
| Black or African American | 3 | 20% |
| No response | 1 | 6.7% |
| **County** | | |
| Allegheny | 4 | 26.7% |
| Philadelphia | 5 | 33.3% |
| Dauphin | 2 | 13.3% |
| Delaware | 2 | 13.3% |
| York | 2 | 13.3% |
| **Political Orientation** | | |
| Democrat | 7 | 46.7% |
| Republican | 4 | 26.7% |
| Non-Partisan/Independent | 2 | 13.3% |
| No response | 1 | 6.7% |

over-represents judges with higher-than-average familiarity with risk assessment instruments, since these individuals are more likely to agree to an interview about such instruments and in turn likely to refer study participants similar to themselves (Parker et al., 2019). I continued recruiting and interviewing judges until I achieved saturation, that is, I no longer heard new information in my interviews (Small, 2009). In total, I attempted to recruit 86 judges, resulting in a response rate of 17%.

Figure 19: The Juanita Kidd Stout Center for Criminal Justice in Philadelphia, Pennsylvania, which houses the county's Court of Common Pleas. Photographed in May, 2022.

**Interview Process.** Interviews with judges ranged from 30 minutes to 2 hours, with a median length of 50 minutes, and were conducted over video call, by phone, and in person; follow-up questions were answered over email and follow-up interviews were conducted with three judges. Interview topics included the career trajectories of judges; sentencing practices; training, impressions, and use of the risk assessment tool; and attitudes about risk assessment instruments more broadly (see Appendix C for interview questions). Interviews with probation officers and Sentencing Commission staff were unstructured and helped triangulate interview data from judges and inform the research project more broadly. This study received an IRB exemption and I made sure not to include any information from interviews that might contain identifying information in order to keep the identities of study participants anonymous.

**Qualitative Analysis.** I produced an analytic memo for each interview (Miles et al., 2014), reviewed interview transcriptions generated by OpenAI's Whisper 2-3 times, and relistened to audio recordings twice. I coded interviews iteratively to identify and label repeating ideas in the interviews, moving between inductive coding and data collection to refine themes and look for disconfirming evidence as further interviews were conducted (Miles et al., 2014). I converged on seven high-level themes, each with 4-10 sub-themes: sentencing practice; PSI ordering behavior; information and training about the tool; familiarity with and misconceptions about the tool; use of the tool; desires and concerns about the tool; and attitudes about risk assessment instruments more broadly (see Appendix D for a code table). In order to ensure internal validity, I used member checks and triangulated data from from multiple sources (Merriam and Tisdell, 2015), including participant observations with chambers and courthouse staff during two in-person site visits to the Allegheny and Philadelphia county courthouses (see Figure 19), public testimony documents, instrument development documentation, and recorded meetings of the Pennsylvania Sentencing Commission, all of which are publicly available on the Commission's website.

**Positionality.** As a white woman, an academic researcher, and a regular contributor to activist initiatives opposing the use of carceral technology in my local community, I acknowledge that my positionality shaped the research questions I was interested in pursuing as well as my interactions with interviewees. I have participated in rallies and other events

organized by the community organization I collaborated with, which helped me build rapport with my community consultants despite my privileged academic position, race, and lack of personal contact with the criminal legal system; nevertheless, these differences likely shaped the feedback my consultants were comfortable giving me. On the other hand, my privileged position as a white researcher from a respected local university helped me access and build rapport with judges, many of whom were also white and received their legal training at elite academic institutions.

## 4.3 Results

With respect to tool uptake, I rapidly achieved saturation in my findings: judges were not interested in, and did not consult, the Sentence Risk Assessment Instrument. Only two of the judges I spoke with reported regularly consulting the instrument, and even these individuals could not recall a single instance in which it had affected their decisions to order a PSI. In more populated counties (Allegheny and Philadelphia), I noticed repeating data by my third interview; I continued getting the same result from judges in smaller counties (Dauphin, Delaware, York), where political orientation and PSI-ordering behavior differed from the larger counties, which I expected to correlate with tool use. However, regardless of county size, judges almost unanimously did not use the instrument. This finding is further supported by the Pennsylvania Sentencing Commission's own quantitative analysis of the tool, which shows that there was no requisite change in PSI-ordering rates after the implementation of the tool. Moreover, my study sample likely over-represents individuals with atypically high interest in and knowledge about the tool; the fact that even these judges ignore the tool supports the generalizability of my finding.

Although I achieved saturation with respect to lack of tool uptake, I saw a wide range of responses for my other interview themes, especially PSI-ordering behavior, familiarity with and misconceptions about the tool, desires and recommendations about the tool, concerns about the tool, and attitudes about risk assessment instruments more broadly. That is, I saw variety in the reasons *why* judges ignored the tool. Here I present the main ones.

### 4.3.1 "I find it to not be particularly, um... helpful."

The most common reason that judges did not use the tool was that they simply did not find it useful. This is due in part to the work of activists, lawyers, and academics who, over years of public testimony hearings, successfully pressured the Pennsylvania Sentencing Commission to remove the most controversial parts of the instrument, including directly showing judges risk scores and detailed risk distributions. The implemented version of the tool recommends ordering additional information about low- and high-risk defendants, in keeping with the original goal of helping judges identify candidates for alternative sentences. However, none of the judges I spoke with were looking to change their PSI-ordering behavior. Judges reported either ordering PSIs for all trial cases, ordering PSIs for more serious trial cases, or almost never ordering PSIs; this behavior reflected how useful judges found the PSIs themselves, whose contents vary by county. Nearly half of the judges I talked to also did not find the contents of PSIs helpful because in many counties, including the state's most populous Philadelphia and Allegheny counties, the reports contain information judges can get simply by talking to the defendant. In other words, the tool intervenes on a factor – PSI-ordering behavior – that judges are uninterested in changing, and falsely assumes that successfully influencing PSI-ordering behavior will in turn influence sentencing decisions.

Five judges explicitly used the words "useless" or "worthless" (sometimes with an expletive) to describe the Sentence Risk Assessment Tool. Over half of the judges also stated that they would have preferred to see different information presented to them at sentencing time, including the causal impacts of different sentencing practices on recidivism, a risk and needs responsivity risk assessment, information about how the risk assessment was derived ("Show me the math"), and information about risk categories ("It would be better if they said high, moderate, or low, to be honest"; one judge said they would only want to see information about low-risk defendants, while another said they would only want to see information about high-risk defendants).

### 4.3.2 "I have no idea where it is on the form; I don't recall looking at it at any point."

Another common reason that judges ignored the tool, which often overlapped with judges' perceptions of the tool's uselessness, was a lack of information about what the tool did or what form its recommendation appeared on. As one judge put it, "I never knew where that information was going to be provided for me. Was it going to come in an email? A news blog? A winter weather alert? I had no idea." Several judges explicitly asked me to show them where on the sentence guideline form that judges routinely receive at sentencing time – "the world's least user-friendly form" – the recommendation appears. Another judge called their supervising judge during my interview because they did not believe me that a sentencing risk assessment instrument was in use in their county. With two exceptions, every judge I spoke with revealed some degree of misconception about the tool during the course of my interview, such as the claim that the tool shows judges risk scores (it does not), that the tool applies to DUI cases (it does not), and that the judge has to do something in order to generate the risk assessment (they do not; it is automatically generated and appears on the sentence guideline form). Several interviewed judges were ashamed about being on the record about their lack of awareness of the tool, while others used their lack of knowledge about the tool as a reason to decline participation in my study.

Nearly all judges had low literacy of the tool, despite the Commission's claim that, effective January 1, 2020, it would "conduct a six-month training and orientation for judges and practitioners related to the use of the Sentence Risk Assessment Instrument, the purpose of the recommendation, and the type of information recommended" (Pennsylvania Commission on Sentencing, 2019a). Many judges and probation officers remarked that the tool – and how to use it – had been poorly publicized. In personal conversations, Commission staff explained that their information campaign had been derailed by the start of the pandemic coinciding with the rollout of the tool.

More broadly, however, my findings indicate systemic problems with how information is disseminated to judges in Pennsylvania. In one particularly revealing moment, a judge told me that they were attending a virtual Continuing Judicial Education session over video

call in the background of their computer – during our interview. The problem of judicial education was echoed to me by a chief probation officer, who lamented that even with respect to the risk assessment already included in PSIs in their county, the probation department had not done much in the way of educating judges about how to interpret risk assessment information, adding that many judges "didn't really understand how it applies to the work that they do" and that this was likely the case statewide.[4]

### 4.3.3 "It's unworkable. I don't know how you're building that into numbers."

In addition to misinformation and perceptions of uselessness, skepticism or concern about risk assessment instruments more broadly was often a complementary reason that judges cited for ignoring the Sentence Risk Assessment Instrument, though it was typically a secondary issue. These concerns fell roughly into three categories.

The most common concern, which roughly half of judges expressed, was that the tool ignored a defendant's humanity. Notably, this was a central issue raised by CADBI members in their feedback on my study design; one formerly incarcerated individual worried that the new risk assessment tool would make judges more likely to ignore the humanity and personal circumstances of the people they sentenced. "Each individual has a history that brought them to this space," this consultant told me. "There must be individualization." Judges echoed this point, raising concern about "having a formula that takes away my ability to see the humanity of the people in front of me"; another judge argued that "cookie cutter justice doesn't work" and that risk assessment was "merely labeling and boxing"; a third said, "I don't know how you can reduce all of the human factors that go into, you know, sentencing or making a bond decision and, and put it into a number, you know, I just, I just think that there are a world of human factors that need to be considered." These judges emphasized the crucial role that individual narratives and personal context played in their sentencing decisions. Most judges also indicated that they did not assign central importance to aggregated recidivism risk in their sentencing decisions (with the exception of recidivism risk for sex crimes). Rather, they were interested in the personal trajectories of criminal defendants, particularly escalation

[4]This issue extends beyond Pennsylvania; low literacy about risk assessment among judges has also been documented in Virginia (Garrett and Monahan, 2020).

toward violent behavior; whether a defendant was employed; and drug use.

Another common concern judges raised was about the tool's bias, especially racial bias. One judge, who identified as Black, was critical of the discriminatory potential of the tool: "Who's making the determinations? Who's interpreting the statistics? You can say anything with statistics." Another judge noted the third-party audit's finding that the tool's high-risk category was less accurate than the low-risk category, commenting that this could be "prejudicial to certain minority groups because there was an historically higher arrest rate, possibly related to things like race rather than actual criminal activity." Judges were concerned about other biases as well – a judge who was otherwise an advocate of risk assessment tools claimed that the tool was biased in favor of sex offenders (a claim that is not factually accurate), while two others commented that age was an unfair indicator of recidivism because minorities are statistically more likely to be stopped by police at a younger age. This concern about bias was not unanimously shared, however; other judges acknowledged that the tool had biases but maintained that these were still better than human biases: "You can never take all biases out. You can never take out – there's biases, people get arrested – what's in it, but you can continue to work on the tools to try to make them as fair as possible. But it's better than individuals." One judge even claimed that "[risk assessment tools] have been deliberately distorted as being racist, as being not accurate, as being using wrong statistics and things like that."

The third most common concern was that the tool was worse than the discretion of experienced judges. A common refrain from judges was that younger, less experienced judges might get more benefit from the risk assessment tool, but that for more experienced judges, such an instrument was unnecessary. There was also a general sentiment from judges that personal discretion was a centrally-defining feature of what it means to be a judge; one judge with over a decade of experience firmly announced in the first 10 seconds of our conversation that they were "elected to be a judge, not a robot." Nine judges independently brought up that judges "don't want to be told by anybody what to do;" however, those same judges did not view themselves as being in this category. Seven judges said their own sentencing practice was better than other judges, describing their sentencing using adjectives like "different," "atypical," or (pleasantly) "shocking" to defendants. Several judges were critical of any efforts

to limit their discretion, including sentencing guidelines, which are supposed to standardize sentence lengths based on an individual's prior record score and the gravity of their current offense. One judge aptly summarized this particular concern: "[The legislature] is trying to give us more narrow options on what we can do. And, and I don't like that, because I think that there's a reason that we're up there – we're up there because supposedly we've demonstrated some ability to think more broadly about the whole system and to make a better decision than just something that's electronically generated. You know if you're going to do it all based on a computer program, then you don't need me out there."

Importantly, however, judges' skepticism about risk assessment instruments should not be conflated with skepticism toward data-driven strategies in criminal justice more broadly. As already mentioned, many judges reported wanting access to more data at sentencing time – just not the kind of information provided by this risk tool. Moreover, most judges did in fact acknowledge the importance of consistency in sentencing and, with few exceptions, reported complying with sentencing guidelines. With the exception of two judges, the skeptical claims above were regularly expressed alongside pro-data and pro-science stances at other points within the same interview. One judge, who had expressed concerns about the tool's racial bias earlier in our interview, maintained that "I'm a believer in science. This [risk assessment] is science, so we need to use it."

### 4.3.4 "Anything that slows down processing will be met with resistance."

Several judges worried that the Sentence Risk Assessment Instrument could have unexpected downstream consequences, were it to be used. The Commission "expressly disavows the use of the sentence risk assessment instrument to increase punishment" (Pennsylvania Commission on Sentencing, 2019a). However, as one judge and public testimonies pointed out, judges can still infer risk levels from the 'Additional Information' label, and empirical evidence from other states suggests that judges are more likely to use risk information to detain individuals longer (Human Rights Watch, 2017). "People know it's called the risk tool. If it's 'Additional Information', there may be some concern about how dangerous the defendant is," a judge noted. Moreover, if judges followed the tool's recommendation to order

PSIs for low-risk defendants – who often have minor sentences – then the tool could have the unintended effect of detaining these defendants longer pre-trial, since ordering a PSI can take 60 days or longer, depending on the county. Another judge remarked, "I'm not letting them [the defendant] sit 8 more weeks in jail because some computer program said so."

Speaking about unintended impacts in other parts of the criminal legal system, two probation officers also shared worries about the tool creating unnecessary – and invisible – labor for their departments, which are tasked with generating the risk and needs responsivity assessments that go into the PSIs in some counties. One of these officers said they feared they were going to get "a flood of cases"[5] where judges were ordering PSIs, but that "thankfully that has not happened" because they did not have the resources to handle such a surge. They said they would like to see the system someday permit having such an assessment done for every defendant, but that this would "require a lot of resources, a lot of resources." The second officer raised the concern that the tool, if widely used, would "significantly slow down" the already-backlogged sentencing process, which they said could cause individuals to spend even more time awaiting trial in jail. To this probation officer, the risk assessment instrument was just "another unfunded mandate, the burden of which was going to fall on county probation."

#### 4.3.5 "We're past that train stop and a little bit further down the tracks."

A minority of judges I spoke with were knowledgeable, vocal advocates of other risk assessment instruments and the Pennsylvania Sentencing Commission's other projects. Even among these four judges, however, only one claimed to be regularly consulting the Commission's tool, with the caveat that it had never changed their PSI-ordering behavior. Judges in this group were either advocates of using risk assessment at other stages of the criminal legal pipeline, such as at preliminary arraignment, or were serving in counties where the Ohio Risk Assessment Instrument (ORAS), a significantly more detailed risk and needs responsivity risk assessment, is conducted by the probation department and is already a routine part of the

[5]This was something projected by CMU's audit, as well: "The total PSI number would increase to 36,336 for the same defendants from 2004 to 2006 and the overall PSI rate of Pennsylvania would be 27.7%, which indicates more labor hours" (Becerril et al., 2019, 24).

PSIs that judges receive. One self-described "cheerleader" for risk assessment instruments explained: "I like the [Commission's] tool, I just like our tool [the ORAS] better – it's shinier and faster." This was the position of two of the probation officers I spoke with as well.

In sum, although nearly all of the judges I interviewed reported ignoring the Sentence Risk Assessment Instrument, their reasons for this varied. This suggests a nuanced explanation for aversion to algorithmic systems in the criminal legal system that is neglected in existing discussions that are centered largely around lack of confidence in technology and fears of deskilling and surveillance. The rest of this paper discusses the implications of this finding for scholarship on algorithmic resistance and risk assessment instruments.

## 4.4 Discussion

'Evidence-based' sentencing strategically positions the objectivity and accuracy associated with algorithmic decision-making systems as a solution to institutional crises of mass incarceration and inefficiency.[6] But the few existing on-the-ground studies of risk assessment instruments – this study included – show that the tools' impacts are different than what either critics or proponents had anticipated. One reason for this is that, much like any institutional reform, the success of algorithm-centric reforms is contingent on the organizational conditions in which they are introduced. An algorithm that is intended to assist decision-makers but is developed without attention to their actual needs, or whether and how they will actually use it, is unlikely to have the anticipated effect, and whatever effect it does have will vary by individual. An algorithm that intervenes on a locus – PSI-ordering – that is highly variable by county and a largely settled behavioral pattern is unlikely to alter that behavior. An algorithm whose success relies on the effective dissemination of information in an institutional context in which judges can be interviewed at the same time as attending virtual training sessions is unlikely to have an effect. Crucially, none of these statements have anything to do with the algorithm's bias or accuracy, which are typically the focus of algorithmic audits and

[6]For discussions of the relationship between quantification, objectivity, and scientific authority, see Porter (1995); Galison (2019) and Espeland and Vannebo (2007) for a discussion of quantification in law specifically.

one of the main criticisms of risk assessment instruments.

The implications of this study can thus be distilled into two main points: an understanding of resistance to technology that considers organizational factors is better able to capture real-world cases of algorithm aversion; and empirical research on the inefficacy of risk assessment instruments supports an alternative argument for their abolition. I discuss these in turn.

### 4.4.1 Algorithm Aversion from an Organizational Perspective

As one judge aptly summarized it, "there was a lot of resistance to the tool" – not only from the community but also from public defenders, probation officers, criminal attorneys and, as I have shown, judges themselves. A standard algorithm aversion explanation for this could be individuals' lack of confidence in the tools (Dietvorst et al., 2015). Distrust is, no doubt, an important part of the story, particularly with respect to public resistance to the instrument. But does lack of confidence explain the resistance from judges? As I discussed in §4.3.3, some judges did cite lack of confidence in the instrument's predictions as one reason for not wanting to use them; however, this was not the primary reason but rather something that came up later in the interview once I started probing about their other concerns about the instrument. Moreover, a weak majority of judges I spoke with were supportive of using risk assessment instruments in some capacity – if not at sentencing, then at some other stage of the criminal legal pipeline – and often reported wanting *more* empirically-derived information to assist decision-making (§4.3.1). In general, I observed among most judges a strong pro-data mentality. In short, lack of confidence is a simplified, algorithm-centered explanation that does not provide an adequate explanation of this real-world case – such as why judges who are self-avowed "cheerleaders" of other risk assessment instruments used in their counties are still critical of the Commission's tool.

Brayne and Christin provide another, sociological explanation: judges may be engaging in behavior like foot-dragging due to fears of deskilling and managerial surveillance (Brayne and Christin, 2020). In their studies of predictive policing and pretrial risk assessment, Brayne and Christin found resistance to algorithms to be strongest in cases of function creep, where algorithmic tools served the added purpose of increasing managerial control and surveiling

bureaucrats' productivity. While this sort of function creep is, for now, absent in the present case, I did see some evidence of automation anxieties and fears of deskilling. Some judges – particularly in Philadelphia – expressed antagonism toward any mandates that were intended to limit judicial discretion, including sentencing guidelines. Almost all the judges described their own discretion as a strength, not a weakness, though they were also often critical of other judges practicing the wrong kind of discretion, and reductions in judicial discretion were sometimes perceived negatively – recall the judge who compared using risk assessment instruments to being "a robot" (§4.3.3). Despite their concerns, however, most judges still expressed agreement with the premise of data-driven sentencing, and few opposed the use of some discretion-limiting measures, such as sentencing guidelines. This makes it unlikely that judges' resistance to the risk assessment instrument is entirely "fueled by fears of deskilling and heightened managerial surveillance" (Brayne and Christin, 2020).[7]

In this case, a more adequate explanation for why judges ignore the tool has to do with the organizational influences that led to the tool's development, policies about the contents of PSIs, and problems in how information is disseminated to judges. One probation officer described the Pennsylvania Sentencing Commission as "trying to make certain groups happy" – that is, the public, the legislature, the judges, and probation officers. One of the outcomes of this negotiation process was the selection of a less-publicly-controversial locus of intervention for the tool: the decision to order a PSI.

However, judges did not report PSIs influencing their sentencing decisions except in very unusual situations, such as where a criminal record is stale.[8] Typically, judges said that PSIs are not very helpful and never "dramatically changed [their] mind" about a sentence; this was the case even for PSIs that contained the more detailed risk assessment. This means that the final version of the tool is, at best, useless for judges; not using the tool was largely a response to this fact, complemented by widespread low literacy about the tool. At worst, judges' adherence to the tool's recommendations could have produced 'ghost work' (Gray and

[7]Brayne and Christin also propose the thesis that predictive technologies displace discretion to less visible areas within organizations. I found some unexpected support for this thesis in counties where an additional risk and needs risk assessment instrument is included in PSIs. Judges revealed to me that probation officers have an enormous amount of discretion in how they prepare such reports; in one county, PSIs even include concrete recommendations for what an individual's sentence should be – a determination made by the probation officer preparing the report.

[8]In such cases, they may use the background information to go below sentencing guidelines.

Suri, 2019) for probation departments and detained individuals longer pre-trial (§4.3.4). But activists still see the final weakened tool as a win. As Hannah Sassaman, a Philadelphia-based community organizer, told news outlets the day after the tool's adoption, "the tool that the Commission instituted yesterday was massively changed over the past few years from one that actively centered racist factors in guessing the future of a sentenced person, to one that will be considerably less damaging" (Gross, 2019).

### 4.4.2 A Resource Argument Against Risk Assessment Instruments

A defender of evidence-based sentencing could make the case that, had the Sentence Risk Assessment Instrument (or the legislative mandate it was built to satisfy) been designed differently – and had the public been more receptive to its development – then perhaps judges would not have been so resistant to it, and perhaps more people would have gotten alternatives to prison sentences as a result. But empirical research on risk assessment instruments used in other states – mostly for pretrial detention decisions – suggests that the tools' impacts have been minimal, unfairly distributed, and have tended to wash out over time (Stevenson, 2018; Sloan et al., 2018; Albright, 2019; Garrett and Monahan, 2020; Stevenson and Doleac, 2021). This has been the trend even for tools with greater uptake and more significant loci of intervention than the decision to order a PSI. Empirical research also suggests that risk assessment instruments introduce an element of arbitrariness to decision-making, such as sharp differences in sentencing decisions for individuals with risk scores that fall near the low-risk category cutoff (Stevenson and Doleac, 2021). As economist and legal scholar Megan Stevenson starkly puts it, "Somehow, criminal justice risk assessment has gained the near-universal reputation of being an evidence-based practice despite the fact that there is virtually no research showing that it has been effective" (Stevenson, 2018).

This research thus contributes another case study to an alternative, empirically-informed argument for abolishing recidivism risk assessment instruments: in practice, these algorithm-centric reforms have no significant impacts on sentencing, are resource-intensive to develop and implement (in a context in which resources are highly limited), and merely pay lip service to addressing the crisis of mass incarceration. Grassroots organizations such as CADBI have

been promoting low-tech liberatory policy changes for decades, including abolishing cash bail, releasing elderly populations from prison, and reinvesting money in schools and communities. Unlike risk assessment instruments, such measures do not rely on individual judges' alignment with policy goals and have robust empirical support for reducing prison populations.[9]

## 4.5 Summary

In this chapter, I presented a qualitative study of criminal court judges, probation officers, and Pennsylvania Sentencing Commission staff; the study's interview questions were designed with the assistance of the community organization CADBI. I found that judges ignored the tool, a result of the tool's lack of utility and shortcomings in how information is disseminated to judges, rather than a mere distrust of the tool or a fear of automation. This lack of utility, in turn, was the interplay of organizational factors and competing interests, which illustrates the importance of an organizational perspective on scholarship on algorithm aversion and resistance. This study adds to the empirical scholarship on risk assessment instruments' on-the-ground impacts and invites a departure from the speculative discourse around AI-centric criminal justice reforms. Evidently, algorithmic decision-making systems are not immune to the shortcomings of other bureaucratic reforms.

The Sentence Risk Assessment Instrument was the locus of considerable time and taxpayer dollars; it was in development for nearly a decade following a 2010 state legislative mandate for adopting a risk assessment tool for sentencing. Despite having no impact, the final version of the tool satisfies this mandate, producing the false impression that some evidence-based measure has been taken to address Pennsylvania's crisis of mass incarceration and racial disparities in sentencing. But as Megan Stevenson puts it, "A practice should not be considered evidence-based because it references big data sets and sophisticated techniques – it should be considered evidence-based because its impacts have been carefully researched and understood" (Stevenson, 2018, 311). The evidence, in the present case, is that the risk assessment tool has had no positive impact. This study adds to an empirically-informed

[9]See Zhou et al. (2021) and Note (2018) for empirical discussions of bail reform.

argument against reforms like these, which can help direct attention toward decarceration efforts that are less costly – and actually work.

# 5.0 Conclusion

Over the last four chapters, this dissertation investigated the diverse ways in which social values influence, and are influenced by, algorithmic decision-making systems. I illustrated this bidirectional relationship throughout using two case studies from the quantitative methodology of crime prediction: contemporary recidivism risk assessment instruments promoted by the US evidence-based sentencing movement, and cybernetic models of crime promoted by Soviet criminologists in the 1960s. In both contexts, the adoption of the methodology emerged from a distrust in human decision-makers and a professed need for mechanical objectivity.

Following a high-profile audit by ProPublica (Angwin et al., 2016), the main concern about the use of recidivism risk assessment instruments in the US has been the tools' algorithmic bias – the systematic deviation of an algorithm's predictions from a normative standard. Critics and proponents of risk assessment instruments alike have focused on the algorithms' technical features, particularly their ability to meet quantitative benchmarks of predictive accuracy and fairness. In chapter 1, I presented the main philosophical and computational discussions of algorithmic bias – what it is, where it comes from, and how to measure and minimize it. I argued that the formal fairness rules used in algorithmic fairness can be fruitfully understood as producing meta-mechanical objectivity, in that they minimize the contribution of algorithmic (rather than human) bias and emerge from suspicion toward algorithmic (rather than human) decisions. Building on work from the philosophical and historical literature on mechanical objectivity, I described the shortcomings of technical audits of risk assessment instruments that depend on the meta-mechanical objectivity of formal measures of fairness. As an illustration, I analyzed Carnegie Mellon University's audit of the Sentence Risk Assessment Instrument, which combined value-laden fairness measures in a balancing act to make policy recommendations that were either rejected by the Pennsylvania Sentencing Commission or had little effect on the consistency of judges' decisions in practice, contrary to the audit's ambitious projections.

In chapter 2, I used the case study of sentence risk assessment instruments as an example of *domain distortion*, in which scientific methods are both impacted by and reinforce certain

social values, distorting how we reason about their domain of application. I argued that sentence risk assessment instruments promote, and are influenced by, social values of control. I showed that risk assessment instruments presuppose a version of legal formalism, which is widely rejected by legal scholars, and require a consequentialist position on sentencing. Risk assessment instruments also prioritize a narrow set of risk-oriented interventions, as opposed to structural interventions that address the root cause of crime, and therefore promote social values of control rather than liberatory or abolitionist social values.

In chapter 3, I examined the domain distortion of crime prediction the 1960s Soviet Union, when Soviet criminologists adopted methods from cybernetics in attempts to predict crime and raise the scientific authority of criminology. Based on archival material I accessed and translated at the Moscow State Library in 2018, I showed how Soviet political values about crime and punishment became embedded in and gained scientific authority through formal modeling choices. I showed that value-laden variable choices in V. N. Kudriavtsev's cybernetic models of crime both inherited and reinforced broader Marxist assumptions about the sources of crime. I argued that legal cybernetics helped revive the authority of Soviet criminology, which had lost its legitimacy during the Stalinist period; legal cybernetics contributed to the legitimisation of political crime-reduction campaigns focused on 'moral rehabilitation', such as Gorbachev's anti-alcohol campaign.

In chapter 4, I showed that social values also become salient through the interactions between algorithmic systems, individuals, and organizational influences. This means that understanding the social and ethical implications of AI requires attention to the social contexts in which it is deployed. I empirically evaluated how sentencing decisions were affected by judges' interactions with the Sentence Risk Assessment Instrument, which was introduced with the aim of increasing consistency in sentencing decisions and reducing the prison population. In interviews with judges, sentencing commission members, and probation officers throughout Pennsylvania, I found that the new instrument's effects were minimal because it was overwhelmingly ignored by judges. I argued that this algorithm aversion was due to organizational factors: county-level norms about pre-sentence investigation reports; alterations made to the instrument by the Pennsylvania Sentencing Commission in response to years of public and internal resistance; and problems with how information is disseminated

to judges. This chapter expands on the important role of organizational influences on professional resistance to algorithms. I designed my study in consultation with formerly incarcerated individuals from a justice reform organization to ensure I did not omit issues of critical importance to impacted communities.

The empirical, historical, and theoretical findings in this dissertation explain why algorithm-centric reforms like risk assessment instruments can fail to live up to their hype. This work holds normative upshots for policy-level justice issues and algorithmic auditing practices.

My dissertation work identifies troubling value entrypoints that are often neglected in the conversation on algorithmic risk assessment in the US criminal legal system. The findings in Chapter 2 not only show that risk assessment instruments have jurisprudential problems but also illustrate the need for openness and public deliberation about the social values of control and incarceration-centric intervention strategies that come with the algorithms' use. Chapter 4 also adds to the growing body of empirical work showing that despite their 'evidence-based' label and high price tag, risk assessment instruments in practice have had little to no positive impact on the high-stakes decisions made in courtrooms. Policy-makers should reconsider the utility and conceptual validity of these tools and reinvest resources into low-tech interventions that would immediately reduce prison populations, a first but necessary step toward the broader project of prison abolition.

Chapters 1 and 4 jointly illustrate the shortcomings of third-party audits of algorithmic decision-making systems: the limitations of formal definitions of algorithmic fairness; the lack of attention to human decision-makers; and the omission of input from stakeholder communities. As auditing becomes an increasingly standard tool for algorithmic accountability – NYC Local Law 144, for instance, made yearly independent audits of hiring algorithms in the city legally mandatory in January – it is crucial to develop a conceptual model of algorithmic auditing that identifies harmful social impacts in a way that is both epistemically and ethically robust. In future work, I aim to do just this: to develop a conceptual toolkit for re-imagining algorithmic auditing practices, with a focus on auditing in resource-stricken contexts like the criminal legal system. I am especially interested in how algorithmic auditing practices can fruitfully center community interests and the role of human discretion. I aim to draw

on tools and insights from community-based participatory research – in which communities generate research questions and are key participants in carrying out the research – as well as the philosophical tools from social epistemology and feminist standpoint theory. In addition to ethical and epistemic benefits, I expect to show that this research approach is a sustainable and tractable way of addressing resource limitations characteristic of public sector audits.

As a proof of concept of this model of algorithmic auditing, I envision expanding my analysis of the impact of the Pennsylvania Sentence Risk Assessment Instrument to a multi-site participatory, human-centered study of similar tools in other states and other stages of the criminal legal pipeline. This would be the first large-scale study of the impacts of risk assessment instruments to date, crucial as these 'evidence-based' tools are widely adopted despite the dearth of evidence of their positive impacts. I expect this future research to carry important criminal justice policy implications, ideally helping offset AI hype around algorithm-centric reforms that pay lip service to redressing mass incarceration and shifting attention to liberatory social values.

I will end on the note this dissertation began with: a call to resist AI hype and to recognize and redress the algorithmic harms already experienced by vulnerable groups, including communities impacted by incarceration, poor people, and Black, indigenous people of color. As I have shown, the adoption of algorithmic decision-making tools emerges from a distrust of social institutions and the prioritazation of social values of control and efficiency; in the US, these developments are presently inseparable from the neoliberal logics of privatization, profit-maximization, and management of 'risky' groups (Fourcade and Healy, 2013). The development and broader social impacts of AI must be understood – and resisted – within this broader context.

## 6.0 Bibliography

98th Congress (1984). Sentencing Reform Act of 1984. *H.R. 5773*.

ACLU of Pennsylvania (2019). Testimony for the Pennsylvania Commission on Sentencing Regarding the July 12, 2019, Proposed Sentence Risk Assessment Tool. *2019 08 Testimony (49 PaB 3718)*. 2019 08 Testimony (49 PaB 3718), Pennsylvania Commission on Sentencing.

Advancing Pretrial Policy and Research (2023). About the Public Safety Assessment: How It Works. https://advancingpretrial.org/psa/factors/.

Akademiia Nauk SSSR: Nauchnyi sovet po kibernetike [USSR Academy of Science: Scientific Council on Cybernetics] (1967). *Voprosy Kibernetiki i Pravo [Questions in Cybernetics and Law]*. Nauka. Accessed at the Russian State Library in Moscow.

Akpinar, N.-J., M. De-Arteaga, and A. Chouldechova (2021, March). The effect of differential victim crime reporting on predictive policing systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, New York, NY, USA, pp. 838–849. Association for Computing Machinery.

Albright, A. (2019). If You Give a Judge a Risk Score: Evidence from Kentucky Bail Decisions.

Alexander, M. (2012). *The New Jim Crow: Mass Incarceration in the Age of Colorblindness*. The New Press.

Alkhatib, A. and M. Bernstein (2019, May). Street-Level Algorithms: A Theory at the Gaps Between Policy and Decisions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, New York, NY, USA, pp. 1–13. Association for Computing Machinery.

Allen, L. E. (1959). Toward a procedure for detecting and controlling of syntactic ambiguity

in legal discourse. pp. 6–12. International Conference for standard on a common language for machine searching and translation.

Anderson, E. (2004). Uses of Value Judgments in Science: A General Argument, with Lessons from a Case Study of Feminist Research on Divorce. *Hypatia 19*(1), 1–24.

Andreev, N. D. and D. A. Kerimov (1961). Vozmozhnosti Ispol'zovaniia Kiberneticheskoi Tekhniki Pri Reshenii Nekotorykh Pravovykh Problem [The Possibility of Using Cybernetics Technology to Solve Certain Legal Problems]. In *Kibernetiku Na Sluzhbu Kommunizmu*, pp. 234–241. Gosudarstvennoe Energeticheskoe Izdatel'stvo. Accessed at the Russian State Library in Moscow.

Andrews, D. A. and J. Bonta (2010). Rehabilitating criminal justice policy and practice. *Psychology, Public Policy, and Law 16*(1), 39–55.

Andrews, D. A., J. Bonta, and J. S. Wormith (2016). The Recent Past and Near Future of Risk and/or Need Assessment:. *Crime & Delinquency*.

Angwin, J. and J. Larson (2016, July). ProPublica Responds to Company's Critique of Machine Bias Story. *ProPublica*. https://www.propublica.org/article/propublica-responds-to-companys-critique-of-machine-bias-story.

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

Arnold Ventures (2017). Public Safety Assessment: A Risk Tool That Promotes.... https://www.arnoldventures.org/stories/public-safety-assessment-risk-tool-promotes-safety-equity-justice.

Aïvodji, U., H. Arai, O. Fortineau, S. Gambs, S. Hara, and A. Tapp (2019). Fairwashing: the risk of rationalization. *arXiv:1901.09749 [cs, stat]*. arXiv: 1901.09749.

Babbage, C. (1833). *On the Economy of Machinery and Manufactures*. Charles Knight.

Baker, T. and J. Simon (2002). *Embracing Risk: The Changing Culture of Insurance and Responsibility*. University of Chicago Press.

Balkin, J. M. and R. B. Siegel (2003, May). The American Civil Rights Tradition: Anticlassification or Antisubordination:. *Issues in Legal Scholarship 2*(1). Publisher: De Gruyter.

Barabas, C. (2019, April). Beyond Bias: Re-Imagining the Terms of 'Ethical AI' in Criminal Law. SSRN Scholarly Paper ID 3377921, Social Science Research Network, Rochester, NY.

Barabas, C., C. Doyle, J. Rubinovitz, and K. Dinakar (2020, January). 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, Barcelona, Spain, pp. 167–176. Association for Computing Machinery.

Barocas, S. and A. Selbst (2016). Big Data's Disparate Impact. *California Law Review 104*(3), 671.

Becerril, D. M., C. Bell, K. LeFevre, L. Lin, W. Mui, K. Shah, and M. Hannigan (2019, May). Validation and Assessment of Pennsylvania's Risk Assessment Instrument, Pennsylvania Commission on Sentencing. *Heinz College System Synthesis Project*.

Benjamin, R. (2019, July). *Race After Technology: Abolitionist Tools for the New Jim Code*. John Wiley & Sons.

Berk, R. (2012). *Criminal Justice Forecasts of Risk: A Machine Learning Approach*. SpringerBriefs in Computer Science. New York: Springer-Verlag.

Berk, R. (2019). *Machine learning risk assessments in criminal justice settings*. Springer.

Berk, R., H. Heidari, S. Jabbari, M. Kearns, and A. Roth (2018). Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research*, 0049124118782533.

Betz, G. (2013, May). In defence of the value free ideal. *European Journal for Philosophy of Science 3*(2), 207–220.

Bhattacharya, J., C. Gathmann, and G. Miller (2013). The Gorbachev Anti-Alcohol Campaign and Russia's Mortality Crisis. *American economic journal. Applied economics 5*(2), 232–260.

Biddle, J. B. (2022, April). On Predicting Recidivism: Epistemic Risk, Tradeoffs, and Values in Machine Learning. *Canadian Journal of Philosophy 52*(3), 321–341. Publisher: Cambridge University Press.

Biddle, J. B. and R. Kukla (2017, June). The Geography of Epistemic Risk. In *Exploring Inductive Risk: Case Studies of Values in Science*, pp. 215–237. Oxford University Press.

Bluhm, R. (2017, June). Inductive risk and the role of values in clinical trials. In K. C. Elliott and T. Richards (Eds.), *Exploring Inductive Risk: Case Studies of Values in Science*, pp. 193–212. Oxford University Press.

Bluvshtein, U. D. (1970). Logicheskoe modelirovanie prichin antiobshchestvennogo povedeniia [Logical modeling of the causes of antisocial behavior]. In *Pravovaia Kibernetika [Legal Cybernetics]*, pp. 105–114. Nauka. Accessed at the Russian State Library in Moscow.

Bonta, J., T. Rugge, T.-L. Scott, G. Bourgon, and A. K. Yessine (2008). Exploring the Black Box of Community Supervision. *Journal of Offender Rehabilitation 47*(3), 248–270. Publisher: Routledge _eprint: https://doi.org/10.1080/10509670802134085.

Boudette, N. E. (2016, September). Autopilot Cited in Death of Chinese Tesla Driver. *New York Times*.

Bratich, J. (2018). Observation in a surveilled world. *The Sage handbook of qualitative research 5*. Publisher: SAGE Publications London, UK.

Brayne, S. (2020, October). *Predict and Surveil: Data, Discretion, and the Future of Policing*. Oxford University Press.

Brayne, S. and A. Christin (2020). Technologies of Crime Prediction: The Reception of Algorithms in Policing and Criminal Courts. *Social Problems*.

Brown, A., A. Chouldechova, E. Putnam-Hornstein, A. Tobin, and R. Vaithianathan (2019, May). Toward Algorithmic Accountability in Public Services: A Qualitative Study of Affected Community Perspectives on Algorithmic Decision-making in Child Welfare Services. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, New York, NY, USA, pp. 1–12. Association for Computing Machinery.

Brown, M. J. (2013, December). Values in Science beyond Underdetermination and Inductive Risk. *Philosophy of Science 80*(5), 829–839.

Bulatov, S. Y. (1929). Vozrozhdenie Lombroso v sovetskoi kriminologii [The rise of Lombroso in Soviet Criminology]. *Revolution of Law 1*(48).

Burgess, E. W. (1936). Protecting the public by parole and parole prediction. *Journal of Criminal Law and Criminology 27*, 491–502.

Bushway, S. and J. Smith (2007, December). Sentencing Using Statistical Treatment Rules: What We Don't Know Can Hurt Us. *Journal of Quantitative Criminology 23*(4), 377–387.

Calabresi, M. (2014, July). Exclusive: Attorney General Eric Holder to Oppose Data-Driven Sentencing. *Time*. https://time.com/3061893/holder-to-oppose-data-driven-sentencing/.

Cartwright, N. (2002). Against Modularity, the Causal Markov Condition, and Any Link Between the Two: Comments on Hausman and Woodward. *British Journal for the Philosophy of Science 53*(3), 411–453.

Cheng, H.-F., L. Stapleton, A. Kawakami, V. Sivaraman, Y. Cheng, D. Qing, A. Perer, K. Holstein, Z. S. Wu, and H. Zhu (2022, April). How Child Welfare Workers Reduce Racial Disparities in Algorithmic Decisions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, pp. 1–22. Association for Computing Machinery.

Chouldechova, A. (2016). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *arXiv:1610.07524 [cs, stat]*. arXiv: 1610.07524.

Chouldechova, A. (2017). Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data 5*(2), 153–163.

Chouldechova, A. and A. Roth (2018, October). The Frontiers of Fairness in Machine Learning. arXiv:1810.08810 [cs, stat].

Christin, A. (2016, March). From daguerreotypes to algorithms: machines, expertise, and three forms of objectivity. *ACM SIGCAS Computers and Society 46*(1), 27–32.

Christin, A. (2017, December). Algorithms in practice: Comparing web journalism and criminal justice. *Big Data & Society 4*(2), 2053951717718855. Publisher: SAGE Publications Ltd.

Chugunov, V. E. and G. F. Gorskii (1967). Ispol'zovanie kiberneticheskikh i shchiotno-analiticheskikh mashin v konkretno-sotsiologicheskikh issledovaniiakh prichin prestupnosti i lichnosti prestupnika [The use of cybernetics and computers in sociological studies into the causes of crime and the personality of the criminal]. In *Voprosy Kibernetiki i Pravo [Questions of Cybernetics and Law]*, pp. 150–163. Nauka. Accessed at the Russian State Library in Moscow.

Coalition to Abolish Death by Incarceration (2019). Proposed Risk Assessment Instrument Public Hearing. *08 Testimony (49 PaB 3718)*. Quizz Cozzens, in testimony to the Pennsylvania Commission on Sentencing, https://pcs.la.psu.edu/guidelines-statutes/risk-assessment/sentence-risk-assessment-proposals-and-testimony/.

Cohen, F. S. (1944). Transcendental Nonsense and the Functional Approach. *ETC: A Review of General Semantics 2*(2), 82–115.

Corbett-Davies, S. and S. Goel (2018). The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *arXiv:1808.00023 [cs]*. arXiv: 1808.00023.

Corbett-Davies, S., S. Goel, and S. González-Bailón (2017, December). Even Imperfect Algorithms Can Improve the Criminal Justice System. *The New York Times*.

County Chief Adult Probation and Parole Officers Association of Pennsylvania (2019). Re: Proposed Sentence Risk Assessment Instrument. *2019 08 Testimony (49 PaB 3718)*. 2019 08 Testimony (49 PaB 3718), Pennsylvania Commission on Sentencing.

Danks, D. and A. J. London (2017). Algorithmic Bias in Autonomous Systems. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, pp. 4691–4697. AAAI Press.

Danziger, S., J. Levav, and L. Avnaim-Pesso (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences 108*(17), 6889–6892. Publisher: National Academy of Sciences Section: Social Sciences.

Daston, L. and P. Galison (2007). *Objectivity*. New York: Zone Books.

De-Arteaga, M., R. Fogliato, and A. Chouldechova (2020, April). A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, New York, NY, USA, pp. 1–12. Association for Computing Machinery.

Desmarais, S. L., K. L. Johnson, and J. P. Singh (2018). Performance of Recidivism Risk Assessment Instruments in U.S. Correctional Settings. In *Handbook of Recidivism Risk/Needs Assessment Tools*, pp. 1–29. John Wiley & Sons, Ltd. Section: 1 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119184256.ch1.

Dieterich, W., C. Mendoza, and T. Brennan (2016). COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity: Performance of the COMPAS risk scales in Broward County. *Northpointe Inc.*.

Dietvorst, B. J., J. P. Simmons, and C. Massey (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General 144*, 114–126. Place: US Publisher: American Psychological Association.

Dobson, M. (2009, April). *Khrushchev's Cold Summer: Gulag Returnees, Crime, and the Fate of Reform after Stalin*. Ithaca, NY: Cornell University Press.

Douglas, H. (2000, December). Inductive Risk and Values in Science. *Philosophy of Science 67*(4), 559–579.

Douglas, H. (2017). Why Inductive Risk Requires Values in Science. In *Current Controversies in Values and Science*. Routledge. Num Pages: 13.

Douglas, H. E. (2009). *Science, Policy, and the Value-Free Ideal*. University of Pittsburgh Press.

Dowling, R. (2013, July). Explaining and Preventing Crime in the Soviet 1970s: the Institute of Criminology and Problems in the (American) War on Crime.

Dressel, J. and H. Farid (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances 4*(1).

Dworkin, R. (1986). *Law's Empire*. Harvard University Press.

Eisman, A. A. (1967). Nekotorye voprosy otsenki kak kolichestvennoi kharakteristiki dostovernosti dokazatel'stv [Some questions regarding assessment as a quantitative characteristic of the reliability of evidence]. In *Voprosy Kibernetiki i Pravo [Questions of Cybernetics and Law]*, pp. 164–179. Nauka. Accessed at the Russian State Library in Moscow.

Elliott, K. C. (2011, March). *Is a Little Pollution Good for You?: Incorporating Societal Values in Environmental Research*. Oxford University Press, USA.

Elliott, K. C. (2017). *A Tapestry of Values: An Introduction to Values in Science*. Oxford University Press.

Ellman, M. (2002, November). Soviet Repression Statistics: Some Comments. *Europe-Asia Studies 54*(7), 1151–1172.

Espeland, W. N. and B. I. Vannebo (2007). Accountability, quantification, and law. In J. Hagan, K. L. Scheppele, and T. Tyler (Eds.), *Annual Review of Law and Social Science*, Annual Review of Law and Social Science, pp. 21–43.

Estelle, S. M. and D. C. Phillips (2018). Smart sentencing guidelines: The effect of marginal policy changes on recidivism. *Journal of Public Economics 164*, 270–293.

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

Fazelpour, S. and D. Danks (2021). Algorithmic bias: Senses, sources, solutions. *Philosophy Compass 16*(8), e12760. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/phc3.12760.

Fazelpour, S. and Z. C. Lipton (2020, January). Algorithmic Fairness from a Non-ideal Perspective. *arXiv:2001.09773 [cs, stat]*. arXiv: 2001.09773.

Federal Bureau of Investigation (2011). NIBRS Offense Codes. *National Incident-Based Reporting System (NIBRS) Uniform Crime Reporting (UCR) Program*. U.S. Department of Justice.

Feeley, M. M. and J. Simon (1992). The New Penology: Notes on the Emerging Strategy of Corrections and Its Implications. *Criminology 30*(4), 449–474.

Feeley, M. M. and J. Simon (1994). Actuarial justice: The emerging new criminal law. In *The futures of criminology*, pp. 172–201. Thousand Oaks, CA: Sage Publications.

Fleisher, W. (2021, July). What's Fair about Individual Fairness? In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, New York, NY, USA, pp. 480–490. Association for Computing Machinery.

Fodor, J. A. (1974). Special Sciences (Or: The Disunity of Science as a Working Hypothesis). *Synthese 28*(2), 97–115. Publisher: Springer.

Fogliato, R., A. Chouldechova, and Z. Lipton (2021, October). The Impact of Algorithmic Risk Assessments on Human Predictions and its Analysis via Crowdsourcing Studies. *Proceedings of the ACM on Human-Computer Interaction 5*(CSCW2), 428:1–428:24.

Forward, J. (2017). The Loomis Case: The Use of Proprietary Algorithms at Sentencing. *InsideTrack, State Bar of Wisconsin 9*(14).

Foucault, M. (1975). *Discipline and Punish: The Birth of the Prison*. Knopf Doubleday Publishing Group.

Fourcade, M. and K. Healy (2013). Classification situations: Life-chances in the neoliberal era. *Accounting, Organizations and Society 38*(8), 559–572.

Friedman, B. and H. Nissenbaum (1996, July). Bias in computer systems. *ACM Transactions on Information Systems 14*(3), 330–347.

Galison, P. L. (2019). Algorists Dream of Objectivity. In J. Brockman (Ed.), *Possible Minds: 25 Ways of Looking at AI*. Penguin Publishing Group.

Garland, D. (2002, August). *The Culture of Control: Crime and Social Order in Contemporary Society*. Chicago, IL: University of Chicago Press.

Garland, D. (2012). *The Culture of Control: Crime and Social Order in Contemporary Society*. University of Chicago Press.

Garrett, B. and J. Monahan (2020, January). Judging Risk. *California Law Review 108*(2), 439–493.

Gavrilov, O. A. and V. A. Kolemaev (1970). Matematicheskie Modeli v Kriminologii [Mathematical models in criminology]. In *Pravovaia Kibernetika [Legal Cybernetics]*. Nauka. Accessed at the Russian State Library in Moscow.

Gerchick, M., T. Jegede, T. Shah, A. Gutiérrez, S. Beiers, N. Shemtov, K. Xu, A. Samant, and A. Horowitz (2022). The Devil is in the Details: Interrogating Values Embedded

in the Allegheny Family Screening Tool. https://www.aclu.org/the-devil-is-in-the-details-interrogating-values-embedded-in-the-allegheny-family-screening-tool.

Gernet, M. H. (1922). *Moralnaya statistika [Moral statistics]*. Publishing House of Central Investigative Department.

Gerovitch, S. (2002). *From Newspeak to Cyberspeak: A History of Soviet Cybernetics*. Cambridge, MA, USA: MIT Press.

Giere, R. N. (2006). *Scientific Perspectivism*. University of Chicago Press.

Gilinskiy, Y. (2017, July). Soviet and post-Soviet Russian criminology – an insider's reflections. *International Journal of Comparative and Applied Criminal Justice 41*(3), 113–122.

Glaser, V. L., N. Pollock, and L. D'Adderio (2021, April). The Biography of an Algorithm: Performing algorithmic technologies in organizations. *Organization Theory 2*(2), 26317877211004609. Publisher: SAGE Publications Ltd.

Glymour, C. and M. R. Glymour (2014). Commentary: race and sex are causes. *Epidemiology (Cambridge, Mass.) 25*(4), 488–490.

Goel, S., R. Shroff, J. Skeem, and C. Slobogin (2021, May). The accuracy, equity, and jurisprudence of criminal risk assessment. *Research Handbook on Big Data Law*, 9–28. ISBN: 9781788972826 Publisher: Edward Elgar Publishing Section: Research Handbook on Big Data Law.

Goertzel, T., E. Shohat, T. Kahn, A. Zanetic, and D. Bogoyavlenskiy (2013, February). Homicide Booms and Busts: A Small-N Comparative Historical Study. *Homicide Studies 17*(1), 59–74.

Graham, L. R. (1987). *Science, Philosophy, and Human Behavior in the Soviet Union*. Columbia University Press.

Gray, M. L. and S. Suri (2019, May). *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. HarperCollins.

Green, B. (2022, October). Escaping the Impossibility of Fairness: From Formal to Substantive Algorithmic Fairness. *Philosophy & Technology 35*(4), 90.

Green, B. and S. Viljoen (2020). Algorithmic realism: expanding the boundaries of algorithmic thought. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, Barcelona, Spain, pp. 19–31. Association for Computing Machinery.

Greiner, D. J. and D. B. Rubin (2011). Causal Effects of Perceived Immutable Characteristics. *The Review of Economics and Statistics 93*(3), 775–785.

Gross, P. (2019, September). Pennsylvania's controversial risk-assessment tool was just approved.

Hacking, I. (1995). The looping effects of human kinds. In *Causal cognition: A multidisciplinary debate*, Symposia of the Fyssen Foundation, pp. 351–394. New York, NY, US: Clarendon Press/Oxford University Press.

Hacking, I. (2015). Let's Not Talk About Objectivity. In J. Y. Tsou, A. Richardson, and F. Padovani (Eds.), *Objectivity in Science*. Springer Verlag.

Hannah-Moffat, K. (2005). Criminogenic needs and the transformative risk subject: Hybridizations of risk/need in penality. *Punishment & Society 7*(1), 29–51.

Hannah-Moffat, K. (2013, April). Actuarial Sentencing: An "Unsettled" Proposition. *Justice Quarterly 30*(2), 270–296.

Hannah-Moffat, K. (2019). Algorithmic risk governance: Big data analytics, race and information activism in criminal justice debates. *Theoretical Criminology 23*(4), 453–470.

Hannah-Moffat, K., P. Maurutto, and S. Turnbull (2009). Negotiated Risk: Actuarial Illusions and Discretion in Probation. *Canadian Journal of Law and Society / La Revue Canadienne Droit et Société 24*(3), 391–409.

Harcourt, B. E. (2007). *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age*. University of Chicago Press.

Harcourt, B. E. (2008, September). *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age*. University of Chicago Press.

Harcourt, B. E. (2010). Risk as a Proxy for Race. SSRN Scholarly Paper ID 1677654, Social Science Research Network, Rochester, NY.

Harcourt, B. E. (2015). Risk as a Proxy for Race: The Dangers of Risk Assessment. *Federal Sentencing Reporter 27*(4), 237–243.

Harding, S. (1992). Rethinking Standpoint Epistemology: What is "Strong Objectivity?". *The Centennial Review 36*(3), 437–470.

Hausman, D. and J. Woodward (1999). Independence, invariance and the causal Markov condition. *The British Journal for the Philosophy of Science 50*(4), 521–583.

Hellman, D. (2019). Measuring Algorithmic Fairness. SSRN Scholarly Paper ID 3418528, Social Science Research Network, Rochester, NY.

Hempel, C. G. (1965). Science and Human Values. In *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, pp. 81–96. Free Press.

Hinton, L. Henderson, and C. Reed (2018, July). An Unjust Burden. https://www.vera.org/publications/for-the-record-unjust-burden.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association 81*(396), 945–960.

Horwitz, S. (2014, August). Eric Holder: Basing sentences on data analysis could prove unfair to minorities. *Washington Post*.

Hosmer Jr., D. W., S. Lemeshow, and R. X. Sturdivant (2013, April). *Applied Logistic Regression*. John Wiley & Sons.

Human Rights Watch (2017). "Not in it for Justice": How California's Pretrial Detention and Bail System Unfairly Punishes Poor People. Technical report.

Jee-Lyn García, J. and M. Z. Sharif (2015). Black Lives Matter: A Commentary on Racism and Public Health. *American Journal of Public Health 105*(8), e27–30.

Johnson, G. (2020a). Algorithmic Bias: On the Implicit Biases of Social Technology. *Synthese*.

Johnson, G. M. Are algorithms value-free? Journal Moral Philosophy.

Johnson, G. M. (2020b). The Structure of Bias. *Mind*.

Kerimov, D. A. (1962). Kibernetika i Pravo [Cybernetics and Law]. *Sovetskoe gosudarstvo i pravo 11*, 98–104.

Kilbertus, N., M. Rojas Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf (2017). Avoiding Discrimination through Causal Reasoning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30*, pp. 656–666. Curran Associates, Inc.

Kleinberg, J., S. Mullainathan, and M. Raghavan (2016). Inherent Trade-Offs in the Fair Determination of Risk Scores. *arXiv:1609.05807 [cs, stat]*. arXiv: 1609.05807.

Klingele, C. (2016, February). The Promises and Perils of Evidence-Based Corrections. *Notre Dame Law Review 91*(2).

Knapp, V. (1965). *O vozmozhnosti ispol'zovaniia kiberneticheskikh metodov v prave [On the possibility of using cybernetics methods in law]*. Progress.

Kotljarchuk, A. and O. Sundström (Eds.) (2017). *Ethnic and Religious Minorities in Stalin's Soviet Union: New Dimensions of Research*. Södertörns högskola (Södertörn University).

Kruse, K. R. (2011). Getting Real About Legal Realism, New Legal Realism and Clinical Legal Education. *New York Law School Law Review 56*, 26.

Kudriavtsev, V. N. (1960). *Ob"ektivnaia storona prestupleniia [Objective side of crime]*. Iuridicheskaia Literatura. Accessed at the Russian State Library in Moscow.

Kudriavtsev, V. N. (1965, September). Rassudi, Mashina! [Judge, Machine!]. *Literaturnaia Gazeta*.

Kudriavtsev, V. N. (1967). *Sovetskaia Kriminologiia – Nauka o Preduprezhdenii Prestuplenii [Soviet Criminology - the Science of Crime Prevention]*. Znanie. Accessed at the Russian State Library in Moscow.

Kudriavtsev, V. N. (1968). *Prichinnost' v Kriminologii: o strukture individual'nogo prestupnogo povedeniia [Causation in Criminology: on the structure of individual criminal behavior]*. Moscow: Iuridicheskaia Literatura. Accessed at the Russian State Library in Moscow.

Kudriavtsev, V. N. and A. A. Eisman (1964). *Kibernetika v Bor'be s Prestupnost'iu [Cybernetics in the Fight Against Crime]*. Znanie. Accessed at the Russian State Library in Moscow.

Kudriavtsev, V. N. and V. I. Eminov (1997). *Kriminologiia [Criminology]*. Lawyer.

Kuhn, T. S. (1977). *The Essential Tension: Selected Studies in Scientific Tradition and Change*. University of Chicago Press. Google-Books-ID: ByjNzh2YgMgC.

Kusner, M. J., J. Loftus, C. Russell, and R. Silva (2017). Counterfactual Fairness. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30*, pp. 4066–4076. Curran Associates, Inc.

Kuznetsova, N. F. and V. Luneev (2004). *Kriminologiia: Uchebnik [Criminology: A Textbook]*. Wolters Kluwer Russia.

Lacey, H. (1999). *Is Science Value Free?: Values and Scientific Understanding*. Psychology Press.

Lacey, H. (2021, August). The methodological strategies of agroecological research and

the values with which they are linked. *Studies in History and Philosophy of Science 88*, 292–302.

Large Soviet Encyclopedia (1940). Bol'shaia sovetskaia entsiklopediia.

Latessa, E. J. and B. Lovins (2014). Risk Assessment, Classification, and Prediction. In G. Bruinsma and D. Weisburd (Eds.), *Encyclopedia of Criminology and Criminal Justice*, pp. 4457–4466. New York, NY: Springer.

Lazer, D., R. Kennedy, G. King, and A. Vespignani (2014, March). The Parable of Google Flu: Traps in Big Data Analysis. *Science 343*(6176), 1203–1205. Publisher: American Association for the Advancement of Science Section: Policy Forum.

Lenhard, J. and E. Winsberg (2010). Holism, Entrenchment, and the Future of Climate Model Pluralism. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics 41*(3), 253–262.

Levi, I. (1960, May). Must the Scientist Make Value Judgments? *The Journal of Philosophy 57*(11), 345.

Levy, K., K. E. Chasalow, and S. Riley (2021). Algorithms and Decision-Making in the Public Sector. *Annual Review of Law and Social Science 17*(1), 309–334.

Lipsky, M. (1980). *Street-Level Bureaucracy: The Dilemmas of the Individual in Public Service*. Russell Sage Foundation.

Llewellyn, K. N. (1950). Remarks on the Theory of Appellate Decision and the Rules or Canons about How Statutes Are to Be Construed. *Vanderbilt Law Review 3*, 395.

Longino, H. (2019). The Social Dimensions of Scientific Knowledge. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2019 ed.). Metaphysics Research Lab, Stanford University.

Longino, H. E. (1990, February). *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton University Press.

Longino, H. E. (1995, September). Gender, politics, and the theoretical virtues. *Synthese 104*(3), 383–397.

Lopez, G. (2017, November). Report: black men get longer sentences for the same federal crime as white men. *Vox*.

Lowenkamp, C. T., B. Lovins, and E. J. Latessa (2009). Validating the Level of Service Inventory—Revised and the Level of Service Inventory: Screening Version With a Sample of Probationers. *The Prison Journal 89*(2), 192–204.

Luneev, V. (2014). Pamiati Akademika Vladimira Nikolaevicha Kudriavtseva [In memory of Academic Vladimir Nikolaevich Kudriavtsev]. *Zakon i Zhizn'*. This is an obituary written by one of his former students.

Maggs, P. B. (2017). Soviet law. *Encyclopedia Britannica*.

Mahmud, H., A. K. M. N. Islam, S. I. Ahmed, and K. Smolander (2022, February). What influences algorithmic decision-making? A systematic literature review on algorithm aversion. *Technological Forecasting and Social Change 175*, 121390.

Malkov, V. D. (2006). *Kriminologiia*. Iu. D. Yustitsinform.

Mallon, R. (2007). A Field Guide to Social Construction. *Philosophy Compass 2*(1), 93–108.

Marcellesi, A. (2013). Is Race a Cause? *Philosophy of Science 80*(5), 650–659.

Martinson, R. (1974). What Works? Questions and Answers About Prison Reform. *The Public Interest 35*.

Matthews, M. (1986, October). *Poverty in the Soviet Union: The Life-styles of the Underprivileged in Recent Years*. Cambridge University Press.

Merriam, S. B. and E. J. Tisdell (2015, August). *Qualitative Research: A Guide to Design and Implementation*. John Wiley & Sons.

Messerschmidt, J. W. (2007). Masculinities, Crime and. In *The Blackwell Encyclopedia of Sociology*. American Cancer Society.

Meyer, M., A. Horowitz, E. Marshall, and K. Lum (2022, June). Flipping the Script on Criminal Justice Risk Assessment: An actuarial model for assessing the risk the federal sentencing system poses to defendants. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, New York, NY, USA, pp. 366–378. Association for Computing Machinery.

Miles, M. B., A. M. Huberman, and J. Saldana (2014). *Qualitative Data Analysis*. SAGE.

Milgram, A. (2014). Why smart statistics are the key to fighting crime. TED@BCG San Francisco, https://www.ted.com/talks/anne_milgram_why_smart_statistics_are_the_key_to_fighting_crime.

Miller, K. (2023, January). IRS Disproportionately Audits Black Taxpayers. https://hai.stanford.edu/news/irs-disproportionately-audits-black-taxpayers.

Mitchell, S. D. (2004). The prescribed and proscribed values in science policy. In P. Machamer and G. Wolters (Eds.), *Science, Values, and Objectivity*, pp. 245–255. University of Pittsburgh Pre.

Mitchell, S. D. (2009). *Unsimple Truths: Science, Complexity, and Policy*. University of Chicago Press.

Mittelstadt, B., S. Wachter, and C. Russell (2023, January). The Unfairness of Fair Machine Learning: Levelling down and strict egalitarianism by default.

Mittelstadt, B. D., P. Allo, M. Taddeo, S. Wachter, and L. Floridi (2016, December). The ethics of algorithms: Mapping the debate. *Big Data & Society 3*(2), 2053951716679679. Publisher: SAGE Publications Ltd.

Molchanov, A. M. (1998). Limitiruiushchie faktory (po I. A. Poletaevu) i printsyp Le-Shatel'e. In *Ocherki istorii informatiki v Rossii*. OIGGM SO RAN.

Monahan, J. (2006). A Jurisprudence of Risk Assessment:. *Virginia Law Review 92*, 45.

Monahan, J. and J. L. Skeem (2016). Risk Assessment in Criminal Sentencing. *Annual Review of Clinical Psychology 12*(1), 489–513.

Moore, M. (1986). Purblind Justice: Normative Issues in the Use of Prediction in the Criminal Justice System. In *Criminal Careers and "Career Criminals"*, Volume 2. Washing, D.C.: National Academy Press.

Movement Alliance Project (2023). Where Are Risk Assessments Being Used? https://pretrialrisk.com/national-landscape/where-are-prai-being-used/.

Murphy, D. E. (2002). Cocaine and Federal Sentencing Policy. *United States Sentencing Commission*.

Nabi, R. and I. Shpitser (2018). Fair Inference on Outcomes. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Nagel, T. (1989, February). *The View From Nowhere*. Oxford University Press.

Narayanan, A. (2019). How to recognize AI snake oil.

Nauchnyi sovet po kibernetike [Scientific Council on Cybernetics] (1961). *Kibernetiku – na sluzhbu kommunizmu [Cybernetics in service of communism]*. Gosudarstvennoe Energetich-eskoe Izdatel'stvo.

Nilforoshan, H., J. D. Gaebler, R. Shroff, and S. Goel (2022, June). Causal Conceptions of Fairness and their Consequences. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 16848–16887. PMLR. ISSN: 2640-3498.

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York University Press.

NorthPointe (2015). *Practitioners Guide to COMPAS*. NorthPointe. http://www.northpointeinc.com/files/technical_ documents/FieldGuide2_081412.pdf.

Note (2018). Note, Bail Reform and Risk Assessment: The Cautionary Tale of Federal Sentencing. *Harvard Law Review 131*(1125).

Ochigame, R. (2019, December). The Invention of "Ethical AI": How Big Tech Manipulates Academia to Avoid Regulation. https://theintercept.com/2019/12/20/mit-ethical-ai-artificial-intelligence/.

O'Malley, P. (2010). *Crime and Risk*. SAGE Publications.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

Pankratov, V. V. (1967). Voprosy uluchsheniia ugolovnogo uchiota [Questions about improving criminal records]. In *Voprosy Kibernetiki i Pravo [Questions of Cybernetics and Law]*. Nauka. Accessed at the Russian State Library in Moscow.

Parasuraman, R. and D. H. Manzey (2010, June). Complacency and bias in human use of automation: an attentional integration. *Human Factors 52*(3), 381–410.

Park, A. L. (2019, February). Injustice Ex Machina: Predictive Algorithms in Criminal Sentencing. *UCLA Law Review*.

Parker, C., S. Scott, and A. Geddes (2019, September). Snowball Sampling. *SAGE Research Methods Foundations*. Publisher: SAGE.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Pekelis, V. D. (1986). *Kiberneticheskaia Smes' [Cybernetic Medley]*. Mir Publishers.

Pennsylvania Commission on Sentencing (2019a). Adopted Sentence Risk Assessment Instrument. *Title 204, Part VII, Chapter 305*.

Pennsylvania Commission on Sentencing (2019b). Risk Assessment Update: Staff's Response to Carnegie Mellon University's External Review. Commission Quarterly Meeting: June 13, 2019. https://pcs.la.psu.edu/guidelines-statutes/risk-assessment/.

Pennsylvania Commission on Sentencing (2019c, June). Risk Update. Commission Meeting, Research & Analysis Unit, Risk Assessment Project. https://pcs.la.psu.edu/guidelines-statutes/risk-assessment/.

Pennsylvania Commission on Sentencing (2020). Sentence Risk Assessment Instrument. *Title 204, Part VIII Criminal Sentencing, Chapter 305*.

Pennsylvania Commission on Sentencing (2021, September). Commission Policy Meeting, Annual Planning Meeting Slides (Part 2, Sentencing Risk Assessment Instrument Initial Analysis). https://pcs.la.psu.edu/policy-administration/previous-commission-policy-meetings/.

Pennsylvania Commission on Sentencing (2022). Pennsylvania Commission on Sentencing Website, https://pcs.la.psu.edu/.

Pennsylvania Sentencing Commission (2020). Risk Assessment. https://pcs.la.psu.edu/guidelines-statutes/risk-assessment/.

Phelps, M. S. (2011, March). Rehabilitation in the Punitive Era: The Gap between Rhetoric and Reality in U.S. Prison Programs. *Law & society review 45*(1), 33–68.

Pierson, E., C. Simoiu, J. Overgoor, S. Corbett-Davies, D. Jenson, A. Shoemaker, V. Ramachandran, P. Barghouty, C. Phillips, R. Shroff, and S. Goel (2020, July). A large-scale analysis of racial disparities in police stops across the United States. *Nature Human Behaviour 4*(7), 736–745. Number: 7 Publisher: Nature Publishing Group.

Polevoi, N. S. (1970). Analitichekii metod identifikatsii lichnosti po fotoizobrazheniiam [An analytic method for photo identification]. In *Pravovaiia Kibernetika [Legal Cybernetics]*, pp. 228–241. Nauka. Accessed at the Russian State Library in Moscow.

Polevoi, N. S. and A. R. Shliakhov (1977). *Osnovy Pravovoi Kibernetiki [Foundations of Legal Cybernetics]*. Izdatel'stvo Moskovskogo Universiteta. Textbook for the course "Osnovy Pravovoi Kibernetiki" [Foundations of Legal Cybernetics]. Accessed at the Russian State Library in Moscow.

Porter, T. M. (1995). *Trust in numbers: the pursuit of objectivity in science and public life*. Princeton, N.J: Princeton University Press.

Poshkiavichius, V. A. (1974). *Primenenie Matematicheskikh i Logicheskikh Sredstv V Pravovykh Issledovaniiakh [The Adoption of Mathematical and Logical Methods in Legal Research]*. Mintis. Accessed at the Russian State Library in Moscow.

Pound, R. (1908). *Mechanical Jurisprudence*. Columbia University Press.

Pound, R. (1910). Law in Books and Law in Action. *American Law Review 44*, 12–36.

Prins, S. J. and A. Reich (2017). Can we avoid reductionism in risk reduction? *Theoretical Criminology*.

Pruss, D. (2021, December). Mechanical Jurisprudence and Domain Distortion: How Predictive Algorithms Warp the Law. *Philosophy of Science 88*(5), 1101–1112. Publisher: Cambridge University Press.

Public Safety Risk Assessment Clearinghouse (2023). History of Risk Assessment. https://bja.ojp.gov/program/psrac/basics/history-risk-assessment.

Rachlinski, J. J. and A. J. Wistrich (2017). Judging the Judiciary by the Numbers: Empirical Research on Judges. SSRN Scholarly Paper ID 2979342, Social Science Research Network, Rochester, NY.

Ratinov, A. R. (1967). Voprosy sledstvennogo myshleniia v svete teorii informatsii [Questions in investigative thinking in light of information theory]. In *Voprosy Kibernetiki i Pravo [Questions of Cybernetics and Law]*. Nauka. Accessed at the Russian State Library in Moscow.

Reiss, J. and J. Sprenger (2017). Scientific Objectivity. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2017 ed.). Metaphysics Research Lab, Stanford University.

Rose, G. (2001). Sick individuals and sick populations. *International Journal of Epidemiology 30*(3), 427–432.

Rosenberg, A., A. K. Groves, and K. M. Blankenship (2017). Comparing Black and White Drug Offenders: Implications for Racial Disparities in Criminal Justice and Reentry Policy and Programming. *Journal of drug issues 47*(1), 132–142.

Rosental', M. and P. Iudin (1954). Kibernetika [Cybernetics]. In *Kratkii filosofskii slovar' [Short philosophical dictionary]*, pp. 236–237. Gospolitizdat.

Ross, C. (2023, March). Denied by AI: How Medicare Advantage plans use algorithms to cut off care for seniors in need.

Ross, J. E. (2002). What Makes Sentencing Facts Controversial - Four Problems Obscured by One Solution. *Villanova Law Review 47*, 25.

Rudner, R. (1953). The Scientist Qua Scientist Makes Value Judgments. *Philosophy of Science 20*(1), 1–6. Publisher: [The University of Chicago Press, Philosophy of Science Association].

Sassaman, H. (2018). Testimony. *06 Testimony (48 PaB 2367)*. Pennsylvania Sentencing Commission, https://pcs.la.psu.edu/guidelines-statutes/risk-assessment/sentence-risk-assessment-proposals-and-testimony/.

Sassaman, H. (2019, September). Pennsylvania's proposed risk-assessment algorithm is racist. *The Inquirer*. https://www.inquirer.com/opinion/commentary/pennsylvania-sentencing-commission-rat-risk-assessment-20190904.html.

Schauer, F. (1988, January). Formalism. *Yale Law Journal 97*(4).

Selivanov, N. A., V. G. Tanasevich, A. A. Eisman, and N. A. Iakubovich (1978). *Sovetsaia Kriminalistika [Soviet Criminology]*. Iuridicheskaia Literatura.

Semukhina, O. (2017). Criminology in Russia. In *The Handbook of the History and Philosophy of Criminology*, pp. 422–436. John Wiley & Sons, Ltd.

Shannon, C. (1956). The bandwagon. *IRE Transactions on Information Theory 2*(1), 3–3.

Sharkey, P., M. Besbris, and M. Friedson (2016, May). Poverty and Crime. *The Oxford Handbook of the Social Science of Poverty*.

Shelley, L. (1979a, January). Soviet Criminology after the Revolution. *Journal of Criminal Law and Criminology 70*(3), 391.

Shelley, L. (1979b). Soviet Criminology: Its Birth and Demise, 1917-1936. *Slavic Review 38*(4), 614–628.

Shliakhov, A. R. (1967). Perspektivy Ispol'zovaniia Dostizhenii Kibernetiki v Deiatel'nosti Iuridicheskikh Uchrezhdenii [Perspectives on Using the Achievements of Cybernetics in the Activities of Legal Institutions]. In *Voprosy Kibernetiki i Pravo [Questions of Cybernetics and Law]*, pp. 7–19. Nauka. Accessed at the Russian State Library in Moscow.

Shliakhov, A. R. (1970). Pervye prakticheskie shagi pravovoi kibernetiki [First practical steps of legal cybernetics]. In *Pravovaia Kibernetika [Legal Cybernetics]*, pp. 5–12. Nauka. Accessed at the Russian State Library in Moscow.

Silver, E. and L. L. Miller (2002). A Cautionary Note on the Use of Actuarial Risk Assessment Tools for Social Control. *Crime & Delinquency 48*(1), 138–161.

Simon, J. (2005). Positively Punitive: How the Inventor of Scientific Criminology Who Died at the Beginning of the Twentieth Century Continues to Haunt American Crime Control at the Beginning of the Twenty-First. *Texas Law Review 84*, 2135.

Singer, J. (1988, March). Legal Realism Now. *California Law Review 76*(2), 465.

Skeem, J. L. and C. Lowenkamp (2016). Risk, Race, & Recidivism: Predictive Bias and Disparate Impact. SSRN Scholarly Paper ID 2687339, Social Science Research Network, Rochester, NY.

Sloan, C., G. Naufal, and H. Caspers (2018, December). The Effect of Risk Assessment Scores on Judicial Behavior and Defendant Outcomes.

Small, M. L. (2009, March). 'How many cases do I need?': On science and the logic of case selection in field-based research. *Ethnography 10*(1), 5–38. Publisher: SAGE Publications.

Sobolev, S. L., A. I. Kitov, and A. A. Liapunov (1955). Osnovnye cherty kibernetiki [Main features of cybernetics]. *Voprosy filosofii [Questions of Philosophy]* (4).

Solomon, P. H. (1974, January). Soviet Criminology: Its Demise and Rebirth, 1928-1963. *The Soviet and Post-Soviet Review 1*(1), 122–140.

Solum, L. B. (2005). Legal Theory Lexicon: Formalism & Instrumentalism.

Spamann, H. and L. Klöhn (2016). Justice is Less Blind, and Less Legalistic, Than We Thought: Evidence from an Experiment with Real Judges. *Journal of Legal Studies 45*.

Starr, S. B. (2014). Evidence-based Sentencing and The Scientific Rationalization of Discrimination. *Stanford Law Review 66*, 803–872.

Starr, S. B. (2015). The Risk Assessment Era: An Overdue Debate Guest Editor's Observations. *Federal Sentencing Reporter 27*(4), 205–206.

State v. Loomis (2016). 881 N.W.2d 749. *Wisconsin Supreme Court*.

Stevenson, M. T. (2018). Assessing Risk Assessment in Action. *Minnesota Law Review 103*, 303.

Stevenson, M. T. and J. L. Doleac (2021). Algorithmic Risk Assessment in the Hands of Humans. SSRN Scholarly Paper ID 3489440, Social Science Research Network, Rochester, NY.

Stewart, A. J., A. P. Copeland, N. L. Chester, J. E. Malley, and N. B. Barenbaum (1997). *Separating together: How divorce transforms families*. Separating together: How divorce transforms families. New York, NY, US: Guilford Press. Pages: viii, 293.

Stith, K. and J. A. Cabranes (1998, October). *Fear of Judging: Sentencing Guidelines in the Federal Courts*. University of Chicago Press.

Trusov, A. I. (1967). Sudebnoe Dokazyvanie V Svete Idei Kibernetiki [Forensic Evidence in Light of the Ideas of Cybernetics]. In *Voprosy Kibernetiki i Pravo [Questions of Cybernetics and Law]*, pp. 20–35. Nauka. Accessed at the Russian State Library in Moscow.

United States Sentencing Commission (1987). Sentencing Table. In *Chapter Five, Part A*. Guidelines Manual. https://guidelines.ussc.gov/chapters/5/parts/.

Universitet Prokuratury Rossiiskoi Federatsii (2018). Istoriia Instituta [History of the Institute]. http://www.agprf.org/instituty/nauchno-issledovatelskiy-institut/istoriya-instituta/.

Unspecified (2007). Pamiati Akademika V. N. Kudriavtsev [In memory of V. N. Kudriavtsev]. *Nauka i Zhizn'* (11), 25.

Vail', P. and A. Genis (1996). *Mir sovetskogo cheloveka [World of the Soviet person]*. Novoe literaturnoe obozrenie.

van Voren, R. (2010, January). Political Abuse of Psychiatry—An Historical Overview. *Schizophrenia Bulletin 36*(1), 33–35.

Wachter, S., B. Mittelstadt, and C. Russell (2021, July). Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review 41*, 105567.

Ward, Z. B. (2021, February). On value-laden science. *Studies in History and Philosophy of Science Part A 85*, 54–62.

Weisberg, M. (2006, December). Robustness Analysis. *Philosophy of Science 73*(5), 730–742. Publisher: Cambridge University Press.

Weiss, R. S. (1995, November). *Learning From Strangers: The Art and Method of Qualitative Interview Studies*. Simon and Schuster.

Werth, R. (2019). Risk and punishment: The recent history and uncertain future of actuarial, algorithmic, and "evidence-based" penal techniques. *Sociology Compass 13*(2), e12659.

Western, B. and C. Wildeman (2009). The Black Family and Mass Incarceration. *The ANNALS of the American Academy of Political and Social Science 621*(1), 221–242.

Winner, L. (1980). Do Artifacts Have Politics? *Daedalus 109*(1), 121–136. Publisher: The MIT Press.

Winsberg, E., B. Huebner, and R. Kukla (2014, June). Accountability and values in radically collaborative research. *Studies in History and Philosophy of Science Part A 46*, 16–23.

Woodward, J. (2010). Causation in biology: stability, specificity, and the choice of levels of explanation. *Biology & Philosophy 25*(3), 287–318.

Woodward, J. (2013). II—Mechanistic Explanation: Its Scope and Limits. *Aristotelian Society Supplementary Volume 87*(1), 39–65.

Woodward, J. (2015). Interventionism and Causal Exclusion. *Philosophy and Phenomenological Research 91*(2), 303–347.

Woodward, J. (2016). Causation and Manipulability. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2016 ed.). Metaphysics Research Lab, Stanford University.

Zhang, J. and E. Bareinboim (2018). Fairness in Decision-Making — The Causal Explanation Formula. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Zhou, A., A. Koo, N. Kallus, R. Ropac, R. Peterson, S. Koppel, and T. Bergin (2021, November). An Empirical Evaluation of the Impact of New York's Bail Reform on Crime Using Synthetic Controls.

Zhuravel', A. A., N. V. Troshko, and L. G. Edzhubov (1970). Ispol'zovanie algoritma obobshchennogo portreta dlia opoznavaniia obrazov v sudebnom pocherkovedenii [Using a generalized portrait algorithm for image recognition in forensic handwriting analysis]. In *Pravovaiia Kibernetika [Legal Cybernetics]*, pp. 212–227. Nauka. Accessed at the Russian State Library in Moscow.

Ægisdóttir, S., M. J. White, P. M. Spengler, A. S. Maugherman, L. A. Anderson, R. S. Cook, C. N. Nichols, G. K. Lampropoulos, B. S. Walker, G. Cohen, and J. D. Rush (2006). The Meta-Analysis of Clinical Judgment Project: Fifty-Six Years of Accumulated Research on Clinical Versus Statistical Prediction. *The Counseling Psychologist 34*(3), 341–382.

# Appendix A Counterfactual Fairness

This appendix begins with a brief overview of the causal modeling language and methodology used in counterfactual fairness, followed by a discussion of the ways in which they are partial and value-laden measures of fairness.

## A.1 Causal Models of Fairness

Building on Pearl (2009)'s counterfactual causal framework, Kusner et al. (2017)'s *counterfactual fairness* account rests on the intuition that protected attributes should not affect predictions unless they come from acceptable causal pathways.[1] Counterfactual reasoning is intended to capture the sentiment expressed in the legal definition of employment discrimination:

> The central question in any employment-discrimination case is whether the employer would have taken the same action had the employee been of a different race (age, sex, religion, national origin etc.) and everything else had been the same (7th Circuit Court, 1996; Pearl, 2009).

In other words, identifying discrimination requires a causal explanation of a decision. This is similar to Woodward's characterization of 'difference-making' explanations, which seek to answer 'what-if-things-had-been-different' questions (Woodward, 2013, 47).

Kusner et al. begin by assuming that they are given a causal model, as specified on Pearl (2009)'s framework. Briefly, a structural causal model is a tuple of sets $\langle U, V, F \rangle$, where

- $U$ is the set of latent (exogenous, unobserved) background variables, which are not caused by any variables in the set of observed variables $V$;
- $V$ is the set of observable (endogenous) variables in the model;
- $F$ is the set of structural functions $\{f_1, ..., f_n\}$, where there is an $f_i$ for each $V_i \in V$.

[1] Several other authors have taken similar approaches, notably Kilbertus et al. (2017), Nabi and Shpitser (2018), and Zhang and Bareinboim (2018). For brevity, I focus on a single account.

This model is assumed to be representable as a directed acyclic graph (DAG), whose nodes correspond to endogenous, observable variables ($V$) and directed edges represent causal relations between them. Each function (also known as a 'structural equation') in the causal model's set of structural functions assigns its corresponding variable $V_i$ a value, depending on the value of $V_i$'s parents and exogenous, unobservable variables.[2]

Assuming that these structural equations are each 'autonomous' (i.e., disrupting one causal mechanism in the model would not disrupt the others), a counterfactual intervention on the system can be modeled by replacing the function for some variable in the model with a constant. This is meant to simulate breaking the relationship between a variable and its parents – instead, the variable's value is just whatever constant the intervention sets it to (Pearl, 2009, 107).

Consider, for instance, the following system of structural equations:

$$A = U_A$$
$$X = U_X$$
$$Y = \alpha A + \beta X$$

Suppose we wanted to represent the counterfactual statement "the value of $Y$ if $A$ had been value $a$." Then we could intervene on $A$, which would replace the equation for $A$ with some value $a$ to form a new system of equations:

$$A = a$$
$$X = U_X$$
$$Y = \alpha A + \beta X$$

These two systems of equations are representable with the causal graphs in Figure 20.

Pearl's assumptions mean that any variables that change as a result of this intervention are caused by (are effects of) $V_i$.[3] In the example above, if $Y$ changes as a result of the intervention on $A$, then $A$ (race) causes $Y$ (risk score).

[2]Formally, $V_i = f_i(pa_i, U_{pa_i})$, where $pa_i$ refers to the parents (i.e., direct causes) of $V_i$ that are explicitly included in the model, and $U_{pa_i}$ refers to the impact of excluded variables, where $pa_i \subseteq V \backslash \{V_i\}$ ($pa_i$ cannot be its own parent), and $U_{pa_i} \subseteq U$. I modify Pearl's original notation slightly to match the notation in Kusner et al.'s paper.

[3]Woodward, 2016.

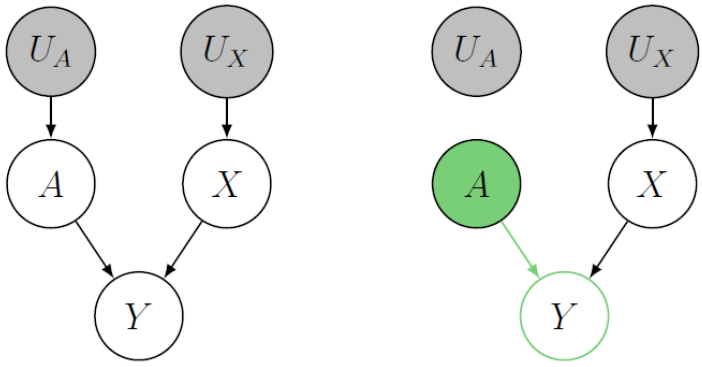


Figure 20: In the graph on the right, $A$ is intervened on, breaking the arrows to its parent $U_A$, and affecting $Y$. (adapted from an example in Pearl, 2009.)

Kusner et al. go on to define the following:

- $A$ is the set of an individual's protected features, which must not be discriminated against.
- $X$ is the set of an individual's other (unprotected) features.
- $Y$ is the outcome to be predicted (e.g., recidivism risk).
- $\hat{Y}$ is the 'predictor', a random variable that is produced by the algorithm as a prediction of $Y$.

Appealing to David Lewis, they suggest that a decision is fair toward an individual if it is no different between our actual world and a counterfactual world in which the individual belongs to a different demographic group. This means that changing protected attribute $A$ while holding anything not causally dependent on $A$ fixed will not change the distribution of $\hat{Y}$ – "$A$ should not be a cause of $\hat{Y}$ in any individual instance" (Kusner et al., 2017, 3). Formally, a predictor $\hat{Y}$ is counterfactually fair if, for any unprotected attribute $X = x$ and protected attribute $A = a$, for all predictions $y$ and for any protected attribute value $a'$ attainable by $A$,

$$\mathbb{P}(\hat{Y}_{A\leftarrow a}(U) \,|\, X = x, A = a) = \mathbb{P}(\hat{Y}_{A\leftarrow a'}(U) = y \,|\, X = x, A = a)$$

where $Y_{A\leftarrow a}(u)$ denotes the solution for $Y$ for some $U = u$ for which the equation for $A$ is replaced with $A = a$, i.e., $A$ is intervened on. An implication of this is that any predictor $\hat{Y}$

that is a function of the non-descendants of the protected attribute $A$ is counterfactually fair, by definition.

As an illustration of what counterfactual fairness assessments might look like, consider a scenario in which a car insurance company assigns insurance prices based on an individual's accident rate $Y$. Some unobserved factor $U$ (like aggression) causes drivers to be more likely to have an accident ($Y$), and also causes them to prefer red cars ($X$). In addition, individuals of some race $A$ are more likely to drive red cars, but they are no more likely to get into accidents than other individuals. Using $X$ to predict $Y$ is unfair because it charges individuals of a certain race higher prices, even though race does not cause driving behavior. This may not be captured by statistical measures, as discussed in the previous section. Counterfactual fairness does capture this intuition because it shows that intervening on $A$ would change $X$, but not $Y$ (Figure 21.).


Figure 21: $A$ represents race, $X$ represents driving a red car, $U$ represents aggressive driving, and $Y$ is the accident rate. In the right graph, $A$ is intervened on, and $Y$ is unchanged (adapted from Kusner et al., 2017).

Counterfactual fairness can also be used to identify cases in which causal pathways do exist between protected variables and outcomes, but that we may wish to remove from our decision-making procedure. In cases where $A$ and $Y$ are associated (such as the observed association between recidivism and race in training data) due to "a world that punishes individuals in a way that is out of their control," (5) Kusner et al. echo the finding that treatment parity and impact parity cannot always be reconciled. Counterfactual fairness suggests a reason for this: "*this is the result of A [race] being a cause of Y [recidivism risk]*" in the algorithm (Kusner et al., 2017, 6). Thus predictive instruments should strive not to use $Y$ (recidivism risk) as the basis for decision making, but rather some $\hat{Y}$ that estimates

another predictor that is "closest" (6) to $Y$ but independent of the protected variable $A$.

## A.2 Critiques of Causal Fairness Measures

Although the presence of causal connections between protected features and predictions can surely be an important indicator of unfairness, the counterfactual fairness approach makes assumptions and choices that either limit the range of cases they can model, or are value-laden in ways that suggest our prior values about fairness serve to identify cases of discrimination, rather than the models themselves. The upshot is that the ability of the causal modeling approach to detect algorithmic unfairness are more limited than might initially seem.

First the causal modeling framework requires several assumptions, which I argue are likely to be violated in the context of recidivism and other sensitive social contexts. In particular, two key dynamics of recidivism are unable to be represented using DAGs – its cyclic nature, and the close conceptual relationship between its variables.

In adopting Pearl's framework, the counterfactual fairness account assumes that the models in question are acyclic. This is not the case in recidivism. For instance, unemployment or economic instability is causally relevant to recidivism (the former is often an explicit input variable in risk assessment algorithms), but having a criminal record makes finding a job much more difficult. Incarceration thus contributes to a cycle of poverty and more incarceration, which is well documented by economists and criminologists. Furthermore, the output variable (the risk score) itself has a backward effect on other variables in the graph, including socioeconomic status, social relationships, and so on, and biases at other stages of the criminal legal pipeline, such as policing, make this cycle even more vicious. Finally, this problem is exacerbated by issues like looping effects – classifications can result in self-fulfilling prophecies, another (difficult to formalize) way in which recidivism predictions have a distinct cyclic effect.[4]

[4]See Hacking (1995) for a discussion of looping effects.

Even in cyclic graphs, interventions can sometimes be well-defined,[5] but in order to work, they require a second assumption called the Causal Markov Condition (CMC), to ensure that the effect of an intervention is not confounded by other variables. CMC states that, conditional on its parent vertices, each vertex in the graph will be conditionally independent of every other vertex in the graph, except of course its children (Hausman and Woodward, 1999, 523). As Hausman and Woodward point out, however, variables that are conceptually or logically connected to each other might have a relationship that is not causal, and in such cases we should expect CMC to fail. Causal models of proxy discrimination often have variables that are closely conceptually related to protected variables, so they are likely to run into this problem.[6]

While a failure to satisfy these assumptions does not imply that cases of discrimination cannot be causally modeled in principle, it does indicate that counterfactual fairness might be applicable only in some cases, suggesting that, like statistical fairness definitions, the choice to use these measures depends on context and an individual's values.

The second problem I will note with counterfactual fairness is the role that values play both in choices about variable inclusion and in determining which causal pathways are 'acceptable'. In particular, the decision to include protected variables in causal models suggests that (a) they are causes that (b) can be intervened on. Both of these points have puzzling implications for a feature like race. This is further complicated by whether a child of a protected feature should be considered a proxy that is meaningfully distinct from the protected feature, or grouped into one variable with the protected feature. Causal variable choice requires a stance on these issues, which carry with them a position on thorny issues like race essentialism. In many cases, these are precisely the questions that are most important and controversial in allegations of discrimination. Values also seem to play an important role in deliberating which causal paths are deemed acceptable, which is essential for identifying discrimination on the causal fairness account. The upshot is, once more, that applying counterfactual fairness in practice requires value-laden decisions.

Woodward (2016) suggests that, for something to be considered a cause in a counterfactual

[5]Thank you to Jim Woodward for this observation. See also Wysocki, in progress dissertation.

[6]Cartwright (2002) has also famously argued that CMC is unlikely to hold in real-world, indeterministic systems.

sense, there needs to be a coherent way to describe an intervention on it, even if such an intervention is not physically possible. Counterfactual fairness accounts, by including protected attributes in their causal models, presuppose such a possibility, but what such an intervention would look like in the case of a protected attribute like race is unclear. Indeed, some epidemiologists have argued that protected features should not be thought of causes, precisely because of this "impossibility of manipulating such traits as race in a way analogous to administering a treatment in a randomized experiment" (Greiner and Rubin, 2011, 775) and because protected features "are not the types of variables that lend themselves to plausible states of counterfactuality" (Holland, 1986, 14). On the contrary, they think that the causal agents in these cases are *perceptions* of race, i.e., racism (Jee-Lyn García and Sharif, 2015).

Proponents of counterfactual fairness agree that interventions on protected features are "often impossible in practice," whereas proxies "sometimes can be intervened upon" (Kusner et al., 2017, 2). Nevertheless, they write:

> [D]espite some controversy, we consider counterproductive to claim that e.g. race and sex cannot be causes. An idealized intervention on some $A$ at a particular time can be seen as a notational shortcut to express a conjunction of more specific interventions, which may be individually doable but jointly impossible in practice. It is the plausibility of complex, even if impossible to practically manipulate, causal chains from $A$ to $Y$ that allows us to claim that unfairness is real (Kusner et al., 2017, 7).

Some philosophers second this attitude. Marcellesi (2013), for instance, points out that the reasoning behind treating racism, rather than race, as a cause cannot account for why we should not make similar claims in other areas, e.g., why we should treat education level as a cause of employment success rather than an employer's *perception* of education level. He adds that "it seems intuitively correct" to think about discrimination as the claim that some individual would have been treated differently had they been a $B$ instead of an $A$.[7]

Supposing, then, that protected attributes can be causes, what counts as a 'proxy' of a protected attribute? This too is a controversial question. Criminal history, for instance, seems to be an important predictor of recidivism, but critics like Harcourt (2015) have pointed out that criminal history has in itself become a proxy for race. Further, variables in causal

[7]Glymour and Glymour (2014) make a related point in favor of thinking of race as a cause, though not in a counterfactual sense.

graphs are tacitly assumed to be independently fixable – an intervention on one should be possible without an intervention on the other (Woodward, 2015) – so when variables cannot be meaningfully intervened on in practice, it becomes unclear whether or not they should be grouped together as a single variable. These issues are tied up with difficult questions about the naturalness of kinds, but at bottom, causal variable choices are just that – choices. Moreover, these choices carry much of the normative force when it comes to distinguishing cases of discrimination.

As an illustration, consider two protected features, race and gender. Different risk distributions and base rates of recidivism are observed for both: on average, men reoffend at a different rate than women, and on average, Blacks reoffend at a different rate than Whites. In neither case would we want to say that base rates are caused by the protected features directly – presumably, differences in criminality between these groups can be explained largely by structural biases in society, different socialization patterns, arrest patterns, and so on. In the graph below, I call these social factors $S$, race $R$, gender $G$, and prediction $Y$ (Figure 22). $S$ seems, at a cursory glance, to be a proxy of both $R$ and $G$, and thus the pathway through $S$ should be disallowed in decision-making in both cases.

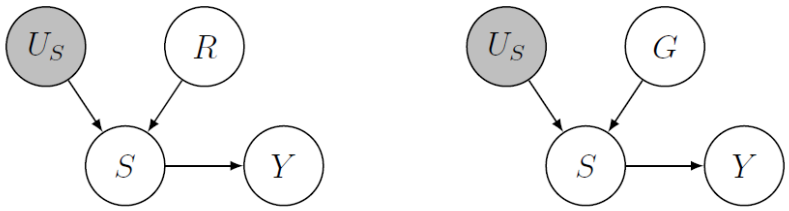


Figure 22: $R$ is race, $G$ is gender, $S$ is social factors, $U_S$ represents other latent variables affecting social factors, and $Y$ is recidivism.

So far, the gender and race situations seem analogous. Now, consider that separate risk assessment tools are often used to classify men and women. The justification for this is that, given the same unprotected features, women tend to have lower rates of recidivism, so grouping them with men would artificially inflate their risk scores and thus would discriminate against them (recall also that calibration and anti-classification cannot account for this). In light of this, the Wisconsin Supreme Court recently ruled that "if the inclusion of gender

[in risk assessment] promotes accuracy, it serves the interests of institutions and defendants, rather than a discriminatory purpose" (State v. Loomis, 2016). Of course, we could make a similar argument in the case of Whites and Blacks, saying that grouping members of both groups together would artificially inflate the risk scores of Whites and thus would discriminate against Whites. This problematic claim is widely rejected, as is the 'separate but equal' idea of separate risk assessment algorithms for Blacks and Whites. The causal models for race and gender are the same, so why are these cases treated differently?

One reason might be due to differences between the types of causal connections in each case. As Woodward (2010) points out, not all causes are the same – some are more stable or specific. The causal connection between race and recidivism, for instance, is highly contingent on social and economic abuses of Black populations in the US, while the causal connection between gender and recidivism is observed more robustly – in many cultures, gender accounts for more variance in crime than any other variable (Messerschmidt, 2007). This sort of information is not captured by a causal model.[8] Another reason these cases are be treated differently might be to avoid harming already marginalized groups. In the male-female case, females are already disproportionately harmed by structural biases in society, so harming them further by grouping them with men seems especially problematic. Conversely, Whites are in general not disproportionately harmed, so it does not seem necessary to group them separately. Either way, these factors are not captured by the causal models for race and gender in Figure 22, which appear from the outside to be identical. The counterfactual fairness accounts would consider $S$ to be a proxy in the case of race (and thus a problematic causal pathway), but not in the case of gender.

My point here is not to argue that these views are correct, nor to suggest that the influence of such social values is necessarily a bad thing – indeed, we should be suspicious of any definition of fairness that attempts to gloss over its social and contextual richness. Rather, I mean to point out that the choice to include protected attributes as variables implicitly encodes a position on controversial issues like the metaphysics of gender and race,[9] and that our values serve to identify cases of discrimination in the above example, not the

[8]In principle, some causal models could represent such information. Thank you to Colin Allen for this observation.

[9]Thank you Marina DiMarco and Katie Creel for pointing this out. See also Mallon (2007).

causal structure of the model per se. Pretending otherwise makes these metrics appear to be more value-neutral than they really are.

## Appendix B Demographic Survey

1. How long have you been a judge? _____
2. Please indicate your gender.
   a. Female
   b. Male
3. Please indicate your age.
   a. 18–29
   b. 30–39
   c. 40–49
   d. 50–59
   e. 60–69
   f. 70-79
   g. 80+
4. Which of the following best describes you?
   a. Asian or Pacific Islander
   b. Black or African American
   c. Hispanic or Latino
   d. Native American or Alaskan Native
   e. White or Caucasian
   f. Other: _____
5. Do you think of yourself as a Republican, a Democrat, an Independent, or something else?
   a. Republican
   b. Democrat
   c. Independent
   d. Other: _____

## Appendix C Interview Guide for Judges

(Start by introducing yourself and asking how their day is going.)
Thanks so much for taking the time to talk to me today. As I mentioned in the letter, I have been talking to judges statewide about how they are using the Pennsylvania Sentencing Commission's new Sentence Risk Assessment Instrument. The purpose of the study is to understand the impacts the tool has had on judicial practice.

I'm going to ask you some open-ended questions about your professional background, your sentencing process, and your experience with this specific tool. How does that sound to you?

[Offer to answer questions about the study, ask permission to record audio from the meeting, then start recording. Provide the following information on the tape:]

– court site/location
– judge name
– date
– interview number

1. Professional background [3 minutes]

   I'd like to begin by hearing a bit about your professional background. Could you please tell me how you became a judge? [Keep this as brief as possible]

   Probes:

   - Tell me about your prior work experiences related to being a judge.
   - How long have you been a judge?
   - How long in criminal?

2. Sentencing process [5-10 minutes]

   Before we talk about the risk assessment tool, I'd like to hear about the process you typically go through when deciding a sentence.

   Probes:

- Factors you consider most important for deciding a sentence. Ask for a specific example: "Can you tell me about a case from this past week?"
  - Demographic factors? E.g. age, past criminal history
  - Recidivism risk: considered/not considered, important/not important?
  - (If recidivism considered) Most important factors for assessing recidivism risk?
- When do you order a pre-sentence investigation report?
  - What information do you receive in the PSI?

3. Instrument implementation/training [10 minutes]

   Let's talk specifically about the Sentence Risk Assessment Instrument. Can you tell me how it was first introduced to you?

   Probes:

   - What were your first impressions of the tool?
   - Tell me about any training you received in using the tool.
   - Did you have any concerns about the tool's introduction?
   - Do you recall talking with your colleagues about the tool?
   - When was the first time you saw a case where the tool applied?

4. Instrument use [5-10 minutes]

   I'd like to turn now to your actual experiences using the instrument. Could you walk me through what happens when you receive the tool's recommendations?

   Probes:

   - What other information do you get at sentencing time?
   - Where is the Sentence Risk Assessment information presented to you?
   - How many cases have you seen so far?
   - Is the tool's recommendation something you typically make note of?
   - What happens when the tool recommends seeking "Additional Information"? (Ask for specific examples)

5. Examples of changes (or lack thereof) [5 minutes]

   (For judges who do not use the tool, skip this section.) I'm interested in hearing whether you've noticed any changes in your day-to-day work since the introduction of the tool. It would be helpful to hear specific examples of things the tool has and has not affected.

- When you see the "Additional Information" label, do you infer the defendant's risk level from this?
  - (If yes) do you think this inference about recidivism risk level affects how you think about a case?
- Can you give an example of a case with the "Additional information" label where you chose to order a pre-sentence investigation report?
- Can you give an example of an "Additional information" case where you did not order a pre-sentence investigation report?
- Can you give an example of when using the tool changed the sentence you assigned?
  - What was it about this additional information that changed your sentence?
- Can you give an example of when using the tool had no effect on the sentence you assigned?

6. Risk assessment in general [5-10 minutes] I'd like to hear what you think about risk assessment tools in general.
   - (If judge mentions racial bias/disparities) Do you think this risk assessment tool could help with the disparities/make them worse?
   - Do you feel that this tool helps judges identify appropriate candidates for alternative sentencing? Why or why not?
   - In general, have you found the tool useful?
7. Thank you and conclude [3 minutes]

   Thank you so much for taking the time to talk to me. This was very helpful.
   - Is there anything you'd like to add that I haven't asked that you think is relevant to this project?
   - Follow-up: After an interview I always find that I've forgotten to ask something. Would it be all right with you if I send you a follow-up question later via email?
   - Snowball: One last thing: I'm trying to learn as much as possible about the use of the Sentence Risk Assessment Instrument. I was wondering if you might be able to put me in touch with other judges to talk about the tool.
   - Reminder to fill out survey.

# Appendix D Code Table

| Sentencing Practice | Code |
|---|---|
| | Comments on using the sentencing guidelines |
| | Describe their sentencing process as different or unusual/better |
| | Emphasize importance of community safety |
| | Consider recidivism to be important in sentencing decision |
| | Consider seriousness of next offense more important than raw recidivism |
| | Comments on judicial discretion |
| | Mention importance of getting many cases through/efficiency |
| **PSI ordering behavior** | |
| | Ordering a PSI is correlated with the seriousness of the case |
| | Always orders PSIs for trial cases |
| | Never/almost never order PSIs |
| | Explicitly say that PSIs aren't helpful |
| | Say that PSIs are helpful in more serious cases |
| | Concerned about how slow generating a PSI is |
| | Order PSI to clarify 'stale' records |
| **Information and training** | |
| | Received training or attended CJE about the tool |
| | Never received training or attended CJE about tool |
| | Heard about tool primarily in email/documentation |
| | Generally not attending/paying much attention to CJEs |
| | CJEs are helpful |
| **Familiarity and misconceptions** | |
| | Misconceptions about what the tool was or how it worked |
| | Wasn't sure where the risk assessment information was presented |
| | Embarrassment or shame about lack of awareness of tool |
| **Use of the tool** | |
| | Do not use/pay attention to the risk assessment tool |
| | Pay attention to risk assessment tool |
| | Tool has never changed decision to order PSI |
| | Tool is not used in their county |
| **Desires and concerns** | |
| | PSI should be generated earlier |
| | Desire access to more information not provided by this tool |
| | Mention racial bias concerns with risk assessment |
| | Concern that tool ignores defendant's humanity |
| | Explicitly say that risk assessment tool isn't helpful |
| | Think judges infer risk level from the tool |
| | Complain about unintuitiveness of SGS/guidelines form |
| | Risk assessment isn't doing anything new |
| **Broader attitudes** | |
| | Positive view of risk assessment more broadly |
| | Skeptical or negative view of risk assessment |
| | Risk assessment useful in other areas of CJ but not for judges |
| | Risk assessment might be useful for less experienced judges |
| | Purpose of risk assessment is to increase consistency |
| | Purpose of risk assessment is to increase efficiency |