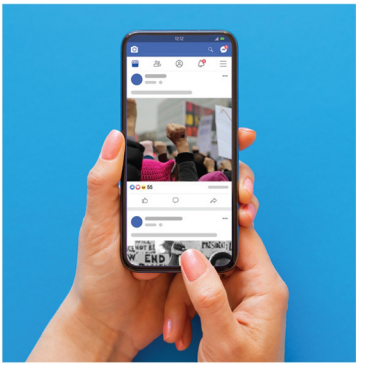
Josh Simons

Algorithms for the People

Democracy in the Age of AI











VERIFY

ALGORITHMS FOR THE PEOPLE

Algorithms for the People

DEMOCRACY IN THE AGE OF AI

JOSH SIMONS

PRINCETON UNIVERSITY PRESS
PRINCETON & OXFORD

Published by Princeton University Press
41 William Street, Princeton, New Jersey 08540
99 Banbury Road, Oxford OX2 6JX

press.princeton.edu

ISBN 978-0-691-24400-6
ISBN (e-book) 978-0-691-24491-4

British Library Cataloging-in-Publication Data is available

Editorial: Matt Rohal
Production Editorial: Jill Harris
Jacket Design: Chris Ferrante
Production: Erin Suydam
Publicity: Kate Hensley and Charlotte Coyne
Copyeditor: Cynthia Buck

Jacket images: Shutterstock / VectorStock / Suzy Bennett (Alamy Stock Photo)

This book has been composed in Arno

Printed on acid-free paper. ∞

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

For Leah

CONTENTS

ALGORITHMS FOR THE PEOPLE

Introduction

"WE DEFINITELY oversample the poor," explains Erin Dalton, deputy director of the Data Analysis Department in Allegheny County, Pennsylvania. "All of the data systems we have are biased. We still think this data can be helpful in protecting kids."[1] Erin is describing the Children, Youth, and Families (CYF) office's Allegheny Family Screening Tool (AFST). This machine learning algorithm mines a database to predict the risk of a child suffering abuse or neglect, producing a score from 1 (lowest risk) to 20 (highest risk). When CYF receives a call reporting possible abuse, a caseworker notes down the details and performs a screening on AFST. If the risk is deemed high enough, a social worker is sent to the child's home. The stakes are high. One in four children experience some form of abuse or neglect in their lifetime. Almost two thousand die across the country every year.[2]

Allegheny County wanted to use its impressive, integrated database to reduce the number of cases of violent maltreatment that were reported but mistakenly ignored and to tackle stubborn racial disparities in child welfare provision. Over several years, with exemplary care and consideration, the county engaged some of the world's best computer scientists, brought in local stakeholders and community leaders, and commissioned regular technical and ethical reviews. And yet AFST still seemed to replicate patterns of racial and economic inequality, disproportionately subjecting poorer, African American families to unwanted and often unnecessary supervision. In Allegheny County, 38 percent of all calls to the maltreatment hotline concern Black children, double the expected rate based on their population. Eight in every 1,000 Black children have been placed outside their home, compared to 1.7 in every 1,000 white children. As one mother explains, frequent visits from investigating authorities can be frustrating: "'Why are you so angry?'" they ask me, 'Because I am tired

of you being here! Leave me alone. I'm trying to get you to go away. We want you to go away.'"[3]

As more of our physical world is converted into numerical data, and more of our behavior is measured, recorded, and predicted, institutions will have strong incentives to widen the range of decisions supported or supplanted by predictive tools, imperceptibly narrowing the spheres in which judgment, empathy, and creativity are exercised and encouraged. As AFST has been fed more data, the "accuracy" with which it predicts "bad outcomes" has steadily increased. "Getting them to trust," explains Erin Dalton, "that a computer screen is telling them something real is a process." Caseworkers are now given less scope to exercise professional judgment and ignore AFST's risk predictions.[4]

In the real world, the design and use of predictive tools like AFST is often messier, more confused, and much less glamorous than the utopian or dystopian visions of AI in movies or novels. Officials find themselves frustrated by poor-quality data and the need to direct technical choices they do not fully understand. Computer scientists feel confused by vague rules and laws and are acutely aware that building predictive tools involves moral and political choices they are not equipped to make. Citizens subject to their predictions feel disempowered by predictive tools, unable to understand or influence their inner logic. Although you cannot always "teach people how you want to be treated," as Pamela Simmons explains of child welfare services, "sometimes you can change their opinion." As she points out, "there's the opportunity to fix it with a person," whereas with AFST, you "can't fix that number."[5]

Three important gaps often fuel these feelings of frustration, confusion, and disempowerment. There is an experience gap between those who build predictive tools and those who use them to make decisions: computer scientists rarely know what it is like to make decisions as a social worker or police officer, as a judge or parole board, as a content moderator or campaign manager. The accountability gap between those in positions of responsibility and those who actually design predictive tools leaves those with responsibility unable or unwilling to justify design choices to the citizens whose lives they shape. Finally, a language gap makes it harder to bridge the experience and accountability gaps: those in positions of responsibility, whether a CEO who wants to make hiring more efficient or a local government leader who wants to further the cause of racial justice, rarely understand the language of computer science in which choices that implicate values and interests are articulated.

These gaps matter because our lives are increasingly structured by the moments in which people in institutions make choices about how to design and

use predictive tools. The lives of families in Allegheny County have been shaped by the moment when computer scientists responded to the county's request for proposals, and then by the moment when they sat with county leaders and CYF staff to make choices about AFST's design. The lives of criminal defendants across the country have been shaped by the moments when local officials decided whether to purchase tools that predict the likelihood that they will reoffend, then by the moment when those officials decided how those tools should be used to inform decisions. The lives of citizens who communicate on Facebook and access information on Google have been shaped by the moments when engineers and policy teams sat down to translate the requirements of the First Amendment or civil rights law into choices about the design of the machine learning systems used in ranking and content moderation. As predictive tools become ever more ubiquitous, the pursuit of justice and democracy will depend in part on how we bridge these gaps of experience, accountability, and language.

I have spent my career bridging these gaps, translating between computer scientists and those in positions of responsibility in technology companies, governments, and academia. Too often, choices about the design of predictive tools are driven by common misunderstandings about the fundamental terms of computer science, as well as by only a vague understanding of what existing laws and values mean for data analytics that often obscures deeper and more intractable political disagreements that ought to be surfaced and debated. If the effects of the widespread use of predictive tools on our society, economy, and democracy depend on how we design and deploy them, we must pursue a vision for technology regulation that goes beyond theorizing the "ethics of AI" and wrestles with fundamental moral and political questions about how technology regulation supports the flourishing of democracy. That is what this book aims to do.

The starting point is establishing a clearer understanding of predictive tools themselves. We need to get under the hood of prediction. I do this by exploring one kind of predictive tool: machine learning. Machine learning is a collection of techniques and methods for using patterns in data to make predictions: for instance, what kinds of allegations of child abuse turn out to be serious, what kinds of people tend to reoffend, or what kinds of advertisements people tend to click on. Wherever institutions can use predictions to inform decisions, or reframe decisions as exercises in prediction, machine learning can be a powerful tool. But the effects of machine learning depend on choices about the design of machine learning models and the uses of their

predictions to make decisions. Child welfare agencies can use machine learning in ways that unintentionally reinforce poverty and racial injustice, or they can use it to empower experienced staff and promote social equality. Internet platforms can use machine learning either to drive short-term engagement and fragment public debate or to encourage shared understanding and experiment with innovative forms of collective decision-making.

Unlike other works on the subject, this book does not assume that the challenges posed by machine learning are new just because the technology is. It articulates a different starting point, a fundamental truth buried in the language of statistics and computer science: machine learning is political. Choices about how to use data to generate predictions and how to use predictions to make decisions involve trade-offs that prioritize some interests and values over others. And because machine learning increases the scale and speed at which decisions can be made, the stakes of these choices are often immense, shaping the lives of millions and even billions of people at breakneck speed.[6]

Machine learning shifts the point at which humans control decisions. It enables people to make not just individual decisions but choices about how decision procedures are structured. When machine learning is used to rank applicants for a job and invite the top 50 percent for interviews, humans exercise control not in deciding which individual candidates to interview, but in designing the model—selecting the criteria it will use to rank candidates and the proportion it will invite to be interviewed. It is not call screeners' decisions about individual allegations of abuse and neglect that shape the lives of millions of families across Allegheny County, but choices about how AFST is designed and how call screeners are instructed to use it to make decisions.[7]

By forcing institutions to make intentional choices about how they design decision procedures, machine learning often surfaces disagreements about previously implicit or ignored values, goals, and priorities. In Allegheny County, the process of building and integrating AFST encouraged a debate about how call screeners should make decisions. Caseworkers felt that decisions should be based on the severity of the allegation, whether it was that a child had been left to play in the street unwatched or had been physically abused, whereas supervisors tended to think that one-off incidents could be misleading and were often misunderstood by those who made referral calls. They preferred to focus on patterns in administrative data that could be used to generate predictions of individual risk. CYF's managers realized that they wanted call screeners to approach their decisions differently, to focus less on the severity of the allegation in the referral and more on the risk to the people

involved. As Erin Dalton explains: "It's hard to change the mind-set of the screeners. . . . It's a very strong, dug-in culture. They want to focus on the immediate allegation, not the child's future risk a year or two down the line. They call it clinical decision-making. I call it someone's opinion."[8]

Similar debates revolve around many of the cases we explore. Whether in the provision of child welfare services, the criminal justice system, or policing, or in the ranking of content on Facebook and Google, designing and integrating machine learning models forces institutions to reflect on the goals of their decision-making systems and the role that prediction should play in them. As more and more decisions are made using prediction, we must engage in public arguments about what different institutions are for, what responsibilities they have, and how decision-making systems should reflect those purposes and responsibilities. This book offers a framework to guide that endeavor. I use the tools of political theory to sharpen our reasoning about what makes machine learning political and what its political character means for regulating the institutions that use it.

By starting with the political character of machine learning, I hope to sketch a systematic political theory of machine learning and to move debates about AI and technology regulation beyond theorizing the ethics of AI toward asking questions about the flourishing of democracy itself. Approaching machine learning through the lens of political theory casts new light on the question of how democracies should govern political choices made outside the sphere of representative politics. Who should decide if statistical tools that replicate racial inequalities in child welfare provision or gender inequalities in online advertising can be justified? According to what criteria? As part of what process? How should Google justify ranking systems that control access to information? Who should determine whether that justification is satisfactory? Should Facebook unilaterally decide how to use machine learning to moderate public debate? If not, who should, and how? By following the threads of machine learning models used in different kinds of organizations, we wrestle with fundamental questions about the pursuit of a flourishing democracy in diverse societies that have yet to be satisfactorily answered.

Above all, my aim is to explore how to make democracy work in the coming age of machine learning. Our future will be determined not by the nature of machine learning itself—machine learning models simply do what we tell them to do—but by our commitment to regulation that ensures that machine learning strengthens the foundations of democracy. Our societies have become too unequal and lack an appreciation of the political goals of laws and

regulations designed to confront entrenched divisions of race, gender, class, and geography. Fear of the uncertainties involved in empowering citizens in processes of participatory decision-making has drained public institutions and public spaces of power and agency. How we govern machine learning could exacerbate these ills, but it could also start to address them. By making visible how and why machine learning concentrates power in courts, police departments, child welfare services, and internet platforms, I want to open our imaginations to alternative futures in which we govern institutions that design and use machine learning to support, rather than undermine, the flourishing of democracy.

The Structure of the Argument

This book is structured in two halves. Each half follows a similar structure but explores machine learning systems used in two different contexts: I examine the political character of machine learning, critique existing proposals for governing institutions that design and use it, and outline my own constructive alternative. In both halves, I argue that existing proposals restrict our capacity to wrestle with the connections between political values and choices in machine learning, and that to govern machine learning to support the flourishing of democracy we must establish structures of political oversight that deliberately keep alive the possibility of revision and experimentation.

The first half of the book explores the machine learning systems used to distribute social benefits and burdens, such as in decisions about child protection, loan applications, bail and parole, policing, and digital advertising. In chapter 1, I describe the specific choices involved in designing and integrating machine learning models into decision-making systems, focusing on how AFST is designed and used in CYF's decisions about investigating allegations of abuse and neglect. I show that the choices involved in machine learning require trade-offs about who wins and who loses, and about which values are respected and which are not. When patterns of social inequality are encoded in data, machine learning can amplify and compound inequalities of power across races, genders, geographies, and socioeconomic classes. Because predictions are cloaked in a veneer of scientific authority, these inequalities can come to seem inexorable, even natural, the result of structures we cannot control rather than social processes we can change. We must develop structures of governance that ensure the design and use of machine learning by institutions to advance equality rather than entrench inequality.

Common responses to this problem are to impose mathematical formalizations of fairness, which I explore in chapter 2, or to apply the law and concept of discrimination, the subject of chapter 3. Underpinning both responses is the idea that if characteristics like race and gender are not morally relevant to the distribution of benefits and burdens, decision-making systems should be blind to those characteristics. Despite its superficial appeal, this idea can lead us to avoid political arguments about when and why people should be treated differently to address structural disadvantages that are corrosive of equal citizenship. In chapter 4, I propose a structure for governing decision-making that, animated by the ideal of political equality, invites us to confront rather than ignore questions about the moral relevance of difference and disadvantage.

The second half of the book explores the machine learning systems used to distribute ideas and information. In chapter 5, I look at the design of ranking systems that use machine learning to order the vast quantities of content or websites that show each time you load Facebook or searches on Google. Because people are more likely to engage with content ranked higher in their newsfeed or search results, ranking systems influence the outcomes they are meant to predict: you engage with content that Facebook predicts you are likely to engage with because that content is displayed at the top of your newsfeed, and you read websites that Google predicts you are likely to read because those websites are displayed at the top of your search results. Building these ranking systems involves choices about the goals that should guide the design of the public sphere and the civic information architecture.

In chapter 6, I argue that Facebook's and Google's machine learning systems have become part of the infrastructure of the digital public sphere, shaping how citizens engage with one another, access information, organize to drive change, and make collective decisions. Their unilateral control over these ranking systems involves a distinctive kind of infrastructural power. Unlike railroads or electricity cables, Facebook's newsfeed and Google's search results not only enable people to do what they want to do but shape what people want to do. Ranking systems mold people in their image, commandeering people's attention and shaping their capacity to exercise collective self-government. We must develop structures of governance within which corporations design infrastructural ranking systems that create a healthy public sphere and civic information architecture.

The common response to the infrastructural power of Facebook and Google is to invoke competition and privacy law. I argue that the goals of protecting competition and privacy are of instrumental, not intrinsic, importance: they

matter because and insofar as they support the flourishing of democracy. We should instead begin by analyzing the distinctive kind of power that Facebook and Google exercise when they build ranking systems powered by machine learning. I propose that structures of participatory decision-making should be built into every stage of Facebook's and Google's design of machine learning systems, allowing for deliberate experimentation and social learning about how best to support the flourishing of democracy in the design of infrastructural ranking systems. I call this the democratic utilities approach.

The two halves of the book connect two debates in political philosophy, law, and computer science that are too often considered separately: fairness and discrimination in machine learning and competition policy and privacy law in the regulation of Facebook and Google. Those interested only in debates about fairness and discrimination in machine learning can read chapters 1 through 4, and those interested only in debates about regulating Facebook and Google can read chapters 5 through 8, but anyone interested in how democracy can flourish in the age of AI should read both.

My motivating question connects these two debates: If our aim is to secure the flourishing of democracy, how should we govern the power to predict? Because machine learning is political, the pursuit of superficially neutral, technocratic goals will embed particular values and interests in the decision-making systems of some of our most fundamental institutions. The regulatory structures that we build must enable deliberate experimentation and revision that encourage us to wrestle with the connections between fundamental political values and choices in machine learning, rather than prevent us from doing so, for it is those connections that will determine the kind of future we build using machine learning. As the legal scholar Salomé Viljoen argues, machine learning raises "core questions [of] democratic governance: how to grant people a say in the social processes of their own formation, how to balance fair recognition with special concern for certain minority interests, what level of civic life achieves the appropriate level of pooled interest, how to not only recognise that data production produces winners and losers, but also develop institutional responses to these effects."[9]

A book about the politics of machine learning therefore becomes an argument about making democracy work in a society of immense complexity. To ensure that we pay unwavering attention to the political choices buried in technical systems, we must avoid forms of political oversight that constrict our capacity to discuss and make decisions together about value-laden

choices and instead embed forms of participatory decision-making every step of the way: in designing machine learning models, in setting standards and goals, and in governing the institutions that set those standards and goals. My proposals for reforming civil rights and equality law and for regulating Facebook and Google are not meant to be definitive statements about regulatory policy, but rather prior arguments about how to structure the institutions and processes we develop to regulate machine learning *given* its unavoidably political character. My goal is to show how democracies should regulate the power to predict if the overarching aim is to secure and promote the flourishing of democracy itself.

A political theory of machine learning illuminates how to think about uses and abuses of prediction from the standpoint of democracy. Attempts to govern the power to predict through technocratic regulations that aspire to exercise state power with neutrality, such as by conceiving of the state as the arbiter of fair decision-making, or by conceiving of the state as the protector of economic competition and personal privacy, will make the governance of prediction a matter not for public argument but for expert decree.

Only by wrestling with the political character of machine learning can we engage with the political and morally contestable character of debates about how to use prediction to advance equality and create a healthy public sphere and civic information architecture. There is no way to design predictive tools that can get around these moral and political debates; in other words, there is no technological solution to how we should govern the power of prediction. Instead of asking questions about the implications of technology for democracy, as if we were passive agents who need protection from the inexorable forces of technology and the institutions that build it, this book asks what a flourishing democracy demands of technology regulation.

My Approach

When I started reading philosophy and political theory, I often wished that scholars would explain how their experience has shaped their arguments. It seemed obvious that political theory was shaped by experience and emotion as well as by analytic rigor, so why not be reflective and open about it? My work in an unusual combination of spheres is central to the argument and approach of this book, so I want to explain, briefly, where I am coming from.

I started thinking about how to regulate data mining while working in the UK Parliament. In 2016, Parliament was scrutinizing the Investigatory Powers

(IP) Bill, the United Kingdom's legislative framework for governing how the intelligence agencies collect and process personal data. Alongside Sir Keir Starmer MP, Tom Watson MP, and Andy Burnham MP, I was working to ensure that judges as well as politicians signed off on requests by intelligence agencies for data collection and analysis. The more I spoke to people in intelligence agencies the more I saw the enormous gulf between what was happening in practice—mass data collection and processing, with limited oversight or evidence about how effective it was—and the public debate about the legislation. It became clear that identifying and articulating political questions about how data are used to make decisions required understanding predictive tools themselves.[10]

After I moved to the United States for my PhD, I quickly enrolled in an introductory machine learning class. Much of what I read went over my head, but a basic training in statistics was enough to help me appreciate the moral and political stakes of debates in computer science about the design of machine learning models. And yet, when I looked around, almost everyone writing about it was either a computer scientist or a lawyer. Few political theorists were seriously engaging with questions about what prediction is, how predictive tools should be designed, or how institutions that build and use them should be governed. So I set about reading all the computer science I could.

Soon after, I began working at Facebook. There I was a founding member of what became the Responsible AI (RAI) team, which needed people with multidisciplinary backgrounds that included ethics and political theory. Over four years at Facebook, I worked with the teams that built many of Facebook's major machine learning systems, including the newsfeed ranking system and the advertising delivery system. The second half of the book uses this experience to explore what makes Facebook's and Google's machine learning systems political and the concrete choices that Facebook and Google make in designing them.[11]

These experiences convinced me of three things. First, the salient moral and political questions about prediction depend on choices made by computer scientists in designing predictive tools. Second, those choices are shaped by the institutional context in which they are made: the policies and culture of a company or public body, the temperament of those who lead it, and the processes established to run it. Third, this institutional context is itself shaped by law and regulation. Any compelling and principled account of how to regulate institutions that use predictive tools must start by reckoning with how they work in practice and are built.

This combination of experience in politics and policy, AI teams in big technology companies, and scholarly training in political theory motivates the argument of this book. If I had lacked any one of these experiences, I doubt I would have thought in quite the same way about the connections between the design of predictive tools, institutional context, and law. To the extent that my approach is illuminating, it is because I have been fortunate enough to see through the eyes of those who build predictive tools, those who lead the companies that build them, and those who are responsible for regulating them.

By using these experiences to imagine what things would look like if political theorists were steering debates about technology regulation, I hope to generate new questions for political theorists, computer scientists, and lawyers. For political theorists and philosophers, my goal is to offer a clear sense of the central moral and political questions about prediction and a strong argument about how to answer them. For computer scientists, my goal is to pose new questions for technical research based on a sharp sense of how technical concepts connect to familiar political ideals. And because my goal is to reframe concepts that underpin current legal approaches to the governance of technology, I should acknowledge to lawyers that many of the legal and policy implications of my argument are often orthogonal to, and sometimes at odds with, existing fields of discrimination, competition, and privacy law. Future work will develop more finely tuned policy interventions.[12]

My approach to this subject is also the result of my background. Although this book is a work of political theory and philosophy, it is also intended as a work of political strategy. My life is devoted to the practice and study of politics, and proposals for political reform succeed when the right coalitions can be built around them. At several junctures, my goal is not to advance a definitive argument about a particular law or concept, but to clarify the stakes and pitfalls of particular strategies for reform by interrogating the concepts and arguments that underpin them. I hope to show what the world might look like if we pursue this or that path, and how each path might affect the flourishing of democracy.

Technology regulation is an opportunity, but one we could easily miss. Grasping that opportunity will require computer scientists, political theorists, and lawyers to collaborate to ensure that powerful institutions are explicit about the values and interests they build into their decision-making processes. That will require that politicians and policymakers confront the ambiguities and limits of some fundamental concepts, laws, and institutions that govern public bodies and private companies. By showing how technology regulation

and democratic reform are connected, my aim is to offer a compelling approach to one of the great challenges of our time: governing organizations that use data to make decisions—whether police forces or child welfare services, Facebook or Google—in a way that responds to some of the challenges our democracies are facing. Regulating technology and reenergizing democracy are entirely connected. Thinking hard about how we regulate technology sharpens some of what feels anemic and constricted about our democracies. And conversely, technology regulation is an opportunity to reimagine and reanimate democracy in the twenty-first century. Above all, I hope this book offers some compelling ideas about how we might grasp that opportunity with both hands.

1

The Politics of Machine Learning I

> No idea is more provocative in controversies about technology and society than the notion that technical things have political qualities. At issue is the claim that the machines . . . can embody specific forms of authority.[1]
>
> —LANGDON WINNER, "DO ARTIFACTS HAVE POLITICS?" (1980)

ALLEGHENY IS A MEDIUM-SIZED Pennsylvania county of about 1.2 million people; Pittsburgh is the county seat. The county has a history of working-class revolt, beginning with the Whiskey Rebellion of 1791, and it was home to the world's first billion-dollar corporation, J. P. Morgan and Andrew Carnegie's U.S. Steel. In 1997, Marc Cherna was hired to run Allegheny County's Children, Youth, and Families (CYF) office, which, as he put it, was "a national disgrace": CYF was processing just 60 adoptions a year, leaving 1,600 children waiting for adoption. Cherna recommended creating a single Department of Human Services (DHS) that would merge several services and house a centralized administrative database. Built in 1999, the database now holds more than a billion records, an average of 800 for each person in the county.[2]

CYF wanted to use these data to improve its decision-making. Too many dangerous cases were being missed, and the stark racial disparities found in cases were deemed worthy of further investigation. When officers receive a call reporting possible abuse, the "callers [often] don't know that much" about the people involved in the allegation, explains Erin Dalton, leaving call screeners with limited information to assess the risk to the child. Prejudice and bias can creep in as callers make unsupported assumptions about Black parents or the neighborhoods in which they live. CYF hoped that, by using data about each person's "history" from the administrative database, call screeners could

make "more informed recommendation[s]" to better protect vulnerable children.[3]

After it decided to build a predictive tool, CYF did everything it could to structure a fair and transparent process for designing and adopting this tool, offering an exemplary lesson in bridging the gaps of experience, accountability, and language. The office empowered call screeners to explain to computer scientists designing the tool how they weighed different factors when making decisions. CYF also commissioned academics to develop transparent explanations of the tool, completed an ethical review of the entire decision-making system, and worked closely with community stakeholders.

None of these measures could address underlying racial inequalities in child welfare provision. Across the United States, child protection authorities are disproportionately likely to investigate Black families and disproportionately likely to remove Black children from their homes. When Cherna joined DHS in 1997, Black children and youths made up 70 percent of those in foster care, but only 11 percent of the county's children and youth population. These disparities remain stubbornly high. In 2016 Black children and youths made up 48 percent of those in foster care, but 18 percent of the county's population. CYF found that its predictive tool simply reproduces these disparities and, when it is used to make real-world decisions, compounds them.[4]

This finding prompted CYF to reflect on how decisions were made to investigate allegations of abuse and neglect and on the goals of child protection itself. Caseworkers felt that decisions should be based on the severity of the allegations, whereas supervisors felt that, because one-off incidents are often misunderstood by those who observe them, it would be better to estimate the risk of individuals involved in allegations using past administrative data. Although they appear purely technical, choices involved in machine learning, by prioritizing the interests of some social groups over others and protecting some fundamental values while violating others, raise fundamental questions about the purpose of decision-making. Machine learning is political.[5]

This chapter uses the Allegheny Family Screening Tool (AFST) to explore what machine learning is and why it matters. I begin by examining the appeal of machine learning's two promises of fairness and efficiency, which incentivize institutions to use prediction in decision-making. I then explore what machine learning is and describe the discrete choices involved in designing and using machine learning models. I then argue that machine learning is irreduc-

ibly and unavoidably political. Machine learning is a process embedded within institutions that involves the exercise of power in ways that benefit some interests over others and prioritize some values over others. Data mining can map, and machine learning can reflect, the multiple dimensions of inequality with unmatched precision. This exploration of the political character of machine learning sets the foundations for the rest of the book.

The Promise of Machine Learning

Decisions are hinges that connect the past to the future, a point of indeterminacy where, for a brief moment, the future hangs in the balance. We especially experience that indeterminacy when we make big decisions: the stomach flutter when deciding whether to marry someone, or the pang of anxiety when deciding whether to quit a job and move to a different town. Even minor decisions—deciding to fix that persistent warning light in the car, or deciding not to have that extra beer—shape the connection between the past and the future. The capacity to make unexpected decisions in full knowledge of the past, without allowing those decisions to be determined by it, is part of what makes us human.[6]

Machine learning holds two fundamental promises for decision-making: the promise of efficiency and the promise of fairness. The consulting company McKinsey & Company estimates the global value of the efficiency gains offered by machine learning to be worth as much as $6 trillion. McKinsey explores using machine learning for "predictive maintenance, where deep learning's ability to analyze large amounts of high-dimensional data from audio and images can effectively detect anomalies in factory assembly lines or aircraft engines"; or in logistics, to "optimize routing of delivery traffic, improving fuel efficiency and reducing delivery times"; or in retail, where "combining customer demographic and past transaction data with social media monitoring can help generate individualized product recommendations."[7]

Machine learning offers efficiency gains in the public sector too. Machine learning can help government bodies be "more efficient" in "terms of public sector resources and shaping how services [are] delivered," and it can even "play a role in addressing large-scale societal challenges, such as climate change or the pressures of an aging population," which often require the processing of large volumes of information. Machine learning could also "improv[e] how services work, sav[e] time, and offer meaningful choice in an environment of 'information overload.'"[8]

The great obstacle to these efficiency gains is the slow and uneven pace at which machine learning is adopted in practice. Just 21 percent of the businesses that McKinsey surveyed had embedded machine learning in "several parts of the business," and just 3 percent had integrated it "across their full enterprise workflows." There is a growing gap between companies that build their own predictive tools, often large firms in financial services or the technology sector, and the typically smaller firms in education, construction, and professional services that purchase off-the-shelf tools. This gap is fast becoming a significant driver of economic inequality.[9]

The second promise of machine learning is fairer decision-making. In a town hall debate in Boston, Massachusetts, Andrew McAfee, a professor at the Massachusetts Institute of Technology (MIT), argued that an app that uses machine learning to grade students' exams is a fairer way to assign grades than a teacher grading individual exams. "If you think teachers are grading the one-hundredth exam with the same attention as they graded the first," argued McAfee, "I have hard news for you. . . . And if you think that if you gave teachers the exact same exam five years in a row and they would give you the same grade on it, I have really hard news for you." He argued that, instead of having teachers assign grades, subject to irrelevant factors like tiredness, the kind of day they have had, or how much they like a student, machine learning would remove human biases and make decisions with perfect consistency. "Let me assure the students in this room," he concluded, "if you want to be evaluated fairly and objectively, you desperately want that app."[10]

Consistency connects the efficiency and fairness promises of machine learning. Whereas people treat cases differently for all kinds of irrelevant reasons, machine learning models generate predictions with complete consistency, treating cases differently only if they are in fact statistically different for some prediction task. And according to one common view, consistency is what makes decision-making fair. The United Kingdom's Royal Society, for instance, argues that, as well as being "more accurate," machine learning can "be more objective than human[s]," helping to "avoid cases of human error," like issues that "arise where decision-makers are tired or emotional."[11] Or as Erin Dalton explains about AFST, "Humans just aren't good at this. They have their own biases. And so having a tool like this that can help to provide that kind of information to really talented staff really does just change everything."[12] Machine learning promises decision-making that is not only more efficient, but fairer too.

What Is Machine Learning?

Many decisions we make are based on regularities or patterns. I wear my raincoat because there are dark and ominous clouds (it will probably rain). The United States has just declared war on Iran, so I head to the store to stock up on gas (the price of oil will probably go up). Most of those who make decisions about a child's safety, releasing a defendant on bail, issuing a mortgage, or hiring someone—or even about whether to call an election or go to war—do so in one way or another based on an assessment of probabilities, regularities, and patterns.

Machine learning automates the process of discovering patterns and regularities by training a model to make predictions about an outcome of interest based on structures and patterns in data sets. An algorithm learns from data in which combinations of statistically related attributes serve as reliable predictors of an outcome of interest. Where people are concerned, the aim "is to provide a rational basis upon which to distinguish between individuals and to reliably confer to the individual the qualities possessed by those who seem statistically similar."[13]

I use the term "machine learning" in this book deliberately in order to distinguish my focus on predictive tools from the somewhat slippery, mythical term "artificial intelligence." AI is better thought of as a scientific field rather than as a single technology that aims to build smart machines to achieve particular goals. Machine learning is better thought of, not as a single technology, but as a set of techniques and methods for prediction.[14] Thinking in terms of techniques and methods draws attention to the human choices involved in designing and using predictive tools. As the computer scientist Cynthia Dwork explains, while many "foster an illusion" that algorithmic "decisions" are "neutral, organic, and even automatically rendered without human intervention—reality is a far messier mix of technical and human curating" because data and algorithms reflect choices: "about data, connections, inferences, interpretations, and thresholds."[15]

How predictive tools work depends on how we design and use them. Machine learning is a set of techniques developed by humans that address problems defined by humans, and training is done on data sets that are assembled by humans and reflect the structures, opportunities, and disadvantages of a very human world. This way of thinking about predictive tools helps make visible discrete human choices that shape how machine learning models work.

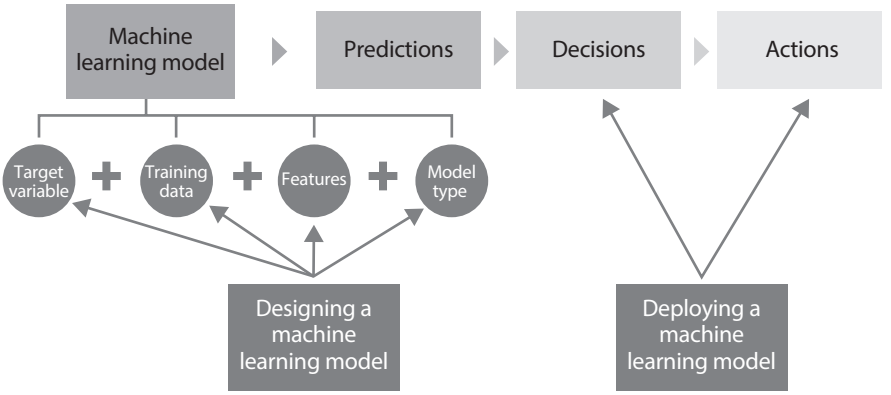


FIGURE 1.1. Building a decision-making procedure that uses machine learning

As Janine, one mother in Allegheny County, put it: "A computer is only what a person puts into it." Our moral, legal, and political analysis should focus on these human choices—which are the focus of much of this book.[16]

We can separate two kinds of choices involved in machine learning. First is a set of choices about the design of a machine learning model, or how data will be used to make predictions: the outcome the model will learn to predict, the data the model will learn from, the features the model will use to predict the outcome, and the training algorithm that will be used to generate the model. The second is a set of choices about the deployment of a machine learning model, or how predictions will be used to make decisions: whether the model will be used to support or supplant human decisions and what actions will result from those decisions.

Predictions

TARGET VARIABLE

The first choice in machine learning is the outcome that a model will learn to predict. An analyst (the person who builds a model) usually has something they want to know about, called the "outcome of interest." This can be simple, such as which emails are spam, or more complex, such as whether candidates for a job would be good employees. The analyst must define a precise proxy for that outcome of interest that can be quantified, measured, and predicted—the "target variable." The art of machine learning lies in turning vague problems in the real world into specific questions about the value of a target variable.[17]

Consider an easy case: building a model to detect spam. Suppose we define "spam," the outcome of interest, as "unwanted email." We need a target variable that serves as a reasonable proxy for unwanted email, something measurable that a model can be trained to predict. The easiest approach would be to use emails labeled as spam to train a model to predict whether new emails have features similar to those already labeled as spam. This is a proxy for the true outcome of interest, which is whether new emails are in fact spam. As definitions of unwanted email change or advertisers develop crafty new ways to make spammy emails look like regular emails, the proxy too must be changed and updated.[18]

Translating a vague problem into a target variable is often complex. Banks must decide whether an individual is sufficiently creditworthy to be offered a loan and what interest rate to attach to that loan. Creditworthiness is not an objective concept that captures something out there in the world, but a concept defined by banks, regulators, and the credit industry that changes with financial conditions and varying appetites for risk. As such, financial institutions exercise considerable discretion in defining the target variable used to predict creditworthiness. The choice of exactly what target variable is predicted by credit default models, and how those predictions are used in loan decisions, will shape who gets what loans.

Defining target variables always involves judgment. Consider how employers might use machine learning in hiring. An employer might define a good employee as someone who makes the most sales, produces the most in the least amount of time, stays in their job the longest, or contributes most to a team's work ethic. Predicting each of these outcomes implies a view about questions of value: the qualities of a good employee, and whether the purpose of employment is to generate revenue, increase production, decrease staff turnover, or boost a firm's morale. All are plausible candidates. Also implied is a prioritization among different interests. If an employer defines the target variable as the predicted length of time a candidate will be in the position, this could produce a model that tends to rank men above women, because, on average, men tend to stay in a position longer than women do.[19] An employer's use of the Myers-Briggs (MBTI) test to predict personality types could also impact genders unequally, since MBTI personality types are distributed unevenly across genders.[20]

Defining the target variable is often the most significant choice in machine learning. It can have profound effects on those subject to a model's predictions. Consider the AFST. The Child Abuse Prevention and Treatment Act, signed into law by President Richard Nixon in 1974, gives states the authority to

define abuse and neglect, above a certain minimum definition. There is no way to directly measure abuse and neglect, so AFST uses several proxies.[21]

The original version of AFST used two models that predicted different target variables. The first predicted the likelihood that an allegation of abuse and neglect deemed not to require further investigation (screened out) would be re-referred within two years—the probability of re-referral being conditional on being screened out. The second predicted the likelihood that an allegation of abuse and neglect deemed to require further investigation (screened in) would lead to a child being removed from their home and placed in foster care within two years—the probability of placement being conditional on being screened in. The original AFST system displayed the highest of the two risk scores.[22]

The problem with the first target variable is that it built in discrimination. CYF's own research, in finding that Black families are disproportionately likely to be called in by other residents, identified referral calls as the major source of racial discrimination in the county's child protection system. The model defined "maltreatment" in terms of an activity that CYF knew to be racially biased. As Erin Dalton explains, "We don't have a perfect target variable. We don't think there are perfect proxies for harm."[23]

It is worth dwelling on why the risk of a child being placed in foster care is a better target variable than the risk of re-referral. Placement is an event that CYF directly observes: CYF always knows when a child has been placed in care. Placement is also a better proxy for abuse and neglect, because CYF removes children from their homes only in the most serious cases. Moreover, decisions about placement are made by different people than those making decisions about call screening. As Alexandra Chouldechova, the computer scientist who helped evaluate AFST, explains: "By predicting an outcome that cannot be directly determined by the staff, we reduce the risk of getting trapped in a feedback loop" in which workers "effect the outcome predicted by the model"—for instance, by gathering incriminating evidence about cases the model labels as high-risk. Allegheny County eventually removed the re-referral prediction model from AFST.[24]

TRAINING DATA

Since machine learning is about using data to make predictions, how we understand machine learning depends on how we understand data. Data are often assumed to represent something objective, as if each data point repre-

sents a fact: where someone lives, how much they earn, or which welfare programs they use.[25]

Yet data reflect not fixed representations of reality, but human choices about what to measure and how. Data are provisional information whose provenance, presentation, and context require further scrutiny. As the philosopher of statistics Ian Hacking writes, "Society became statistical [through] the enumeration of people and their habits. . . . The systematic collection of data about people has affected not only the ways in which we conceive of a society, but also the ways in which we describe our neighbour. It has profoundly transformed what we choose to do, who we try to be, and what we think of ourselves."[26]

Data reveal patterns about populations. States and corporations measure people not primarily because they want to know about each individual, but because they want to understand the behavior of social groups, societies, and countries. The more data an institution has, the more sophisticated the patterns they can detect and the more effectively they can use those patterns to predict, mold, and control. The power of the world's largest tech companies depends not on more sophisticated machine learning techniques, but on the volume of data they have and the speed and efficiency with which they can gather more. Google is good at detecting spam because it can assemble a data set of billions of labeled examples. The power of machine learning often depends on the volume of training data.[27]

The second step in machine learning is to assemble these training data. Choices about the target variable determine what a model learns to predict, and choices about training data determine what a model learns from. As with defining a target variable, assembling and interpreting data sets requires the exercise of judgment.

Consider the use of predictive tools in the Covid-19 crisis. As soon as the virus hit, scientists began to build models to predict how many could die. The range of predictions was enormous, from 200,000 to 2.2 million in the United States, and from 20,000 to 510,000 in the United Kingdom. Despite the often misleading reporting of these numbers, the range reflected an openness about the limits of what scientists understood about the disease and its spread. Imagine a simple version of a model predicting how many could die from Covid-19 in a country in which deaths are treated as a function of the number of those vulnerable multiplied by the infection rate multiplied by the fatality rate. Each of these variables incorporates a dizzying range of uncertainties.[28]

Take the fatality rate, calculated by dividing the total number of cases by the total number of deaths. Gathering data on these numbers is far from simple. At

the start of the pandemic, most countries were vastly underestimating their total number of cases. In the United States and the United Kingdom, where testing was constrained, the number of reported cases was anywhere from three to sixty times fewer than the number of actual cases. Then there were the false positives and false negatives produced by Covid-19 tests. A false positive rate of 4 percent might sound low, but for every one million tests, that could be forty thousand mistakes. Estimating the number of deaths is even more complex. Again, the problem was not just about partial data but about inherent uncertainties in the data-gathering processes. What it meant for a death to be "caused" by Covid-19 was not clear: Should the death of someone in a hospital who had been dying from terminal cancer and tested positive count? Because hospitals were among the first places to get tests, such deaths were among the first cases counted as Covid-19 deaths. But what about my grandma? She died in a care home in Bury, United Kingdom, in April 2020 aged ninety-four. She had a cough and difficulty breathing, yet because there were no tests available at the time, hers was not recorded as a Covid-19 death.[29]

Data represent not facts but judgments. The more you explore data sets, the clearer the judgments involved in constructing them become. One simple input, the death rate, requires countless choices about measuring the infection rate and deciding whose deaths count as Covid-19 deaths. Predictions can obscure the choices involved in assembling data.

Choices about what to measure—and what not to measure—are inextricably bound up with structures of power. Those least likely to produce data trails are often those most excluded by society, as institutions have less interest in gathering data about those who cannot engage in the formal economy. This results in "the non-random, systemic omission of people who live on big data's margins."[30] For instance, Street Bump is an ingenious app built in Boston that uses the accelerometers in smartphones to detect potholes. This can help cut the costs of keeping roads safe. Potholes are most effectively reported, however, in areas where most people have smartphones, that is, in generally wealthier neighborhoods that already have fewer potholes. Relying on the app would cause authorities to reduce services to already underserved, poorer communities. The widespread assumption that data accurately represent a population is more often wrong than right.[31]

AFST is also an example of partial data. In its original form, one-quarter of the variables in AFST's training data set were measures of poverty, while another quarter tracked the juvenile justice system. As a result, AFST was trained on data that disproportionately represented low-income, African American