CONTENTS

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 im-
mediate 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 integrat-
ing 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 in-
stitutions 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 pro-
cess? How should Google justify ranking systems that control access to infor-
mation? 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 ma-
chine 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 be-
come 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.

INDEX