PART I

Some Cases, Some Ground Clearing

1

Introduction

1.1 THREE CASES

1.1.1 *Big Bars Bad:* Loomis *and COMPAS*

A little after 2 a.m. on February 11, 2013, Michael Vang sat in a stolen car and fired a shotgun twice into a house in La Crosse, Wisconsin. Shortly afterward, Vang and Eric Loomis crashed the car into a snowbank and fled on foot. They were soon caught, and police recovered spent shell casings, live ammunition, and the shotgun from the stolen and abandoned car. Vang pleaded no contest to operating a motor vehicle without the owner's consent, attempting to flee or elude a traffic officer, and possession of methamphetamine. He was sentenced to ten years in prison.[1]

The state of Wisconsin also charged Loomis with five crimes related to the incident. Because Loomis was a repeat offender, he would face a lengthy prison sentence if convicted. Loomis denied being involved in the shooting, and he maintained that he joined Vang in the car only after the shooting. Nonetheless, Loomis waived his right to a jury trial and pleaded guilty to two less severe charges (attempting to flee a traffic officer and operating a motor vehicle without owner consent). The plea agreement dismissed the three most severe charges[2] but stipulated that they would be "read-in" such that the court would consider them at sentencing and would consider the underlying, alleged facts of the case to be true. In determining Loomis's sentence, the circuit judge ordered a presentence investigative report ("PSI" or "presentence report"), using a proprietary risk assessment tool called COMPAS that is developed by Northpointe, Inc.[3]

[1] Jungen, "Vang Gets 10 Years in Prison for Drive-by Shooting."

[2] First degree recklessly endangering safety, possession of a firearm by a felon, and possession of a short-barreled shotgun or rifle (all as party to a crime). See *Wisconsin v. Loomis*, 881 N.W.2d paragraph 11.

[3] The tool used is part of a suite of assessment tools developed for use at various stages in the criminal justice system with different algorithms and software packages geared toward (among others) defendants who are recently incarcerated or under state supervision (COMPAS Core), persons who will soon reenter their community after incarceration (COMPAS Reentry), young people (COMPAS Youth), and general case management (Northpointe Suite Case Manager). The tool used in Loomis is COMPAS Core (which we call "COMPAS" for simplicity).

COMPAS takes as inputs a large number of data points about a defendant's criminal behavior, history, beliefs, and job skills, and generates a series of risk scales. These include pretrial release risk (likelihood that a defendant will fail to appear in court or have a new felony arrest if released prior to trial), risk of general recidivism (whether a defendant will have subsequent, new offenses), and risk of violent recidivism.[4] Among the factors that COMPAS uses to assess these risks are current and pending charges, prior arrests, residential stability, employment status, community ties, substance abuse, criminal associates, history of violence, problems in job or educational settings, and age at first arrest.[5] Using information about these factors and a proprietary algorithm, COMPAS generates bar charts corresponding to degree of risk. According to Northpointe, "[b]ig bars, bad—little bars, good," at least as a first gloss.[6] Users can dig deeper, though, to connect particular risk factors to relevant supervisory resources.

Loomis's COMPAS report indicated that he presented a high risk of pretrial recidivism, general recidivism, and violent recidivism.[7] The presentence report recounted Northpointe's warning about the limitations of COMPAS, explaining that its purpose is to identify offenders who could benefit from interventions and to identify risk factors that can be addressed during supervision.[8] Likewise, the presentence report emphasized that COMPAS scores are inappropriate to use in determining sentencing severity.[9] Nonetheless, the prosecution urged the court to use Loomis's risk scores, and the circuit court referenced the scores at sentencing.[10] The presentence and COMPAS reports were not the only bases for the sentence: The other charges (i.e., those to which Loomis did not plead guilty) were read in, meaning that the trial court viewed those charges as a "serious, aggravating factor."[11] The court sentenced Loomis to "within the maximum on the two charges" amounting to two consecutive prison terms, totaling sixteen and a half years.[12]

1.1.2 *School-wide Composite Scoring: Wagner and TVAAS*

In 2010, the state of Tennessee began requiring that school systems evaluate teachers based on value added models (VAMs). VAMs are algorithmic tools used to measure student achievement.[13] They seek to isolate and quantify teachers' individual

4 Northpointe, Inc., "Practitioner's Guide to COMPAS Core," 27–28.
5 Northpointe, Inc., 24.
6 Northpointe, Inc., 4.
7 *Wisconsin v. Loomis*, 881 N.W.2d paragraph 16.
8 *Wisconsin v. Loomis*, 881 N.W.2d paragraph 16.
9 *Wisconsin v. Loomis*, 881 N.W.2d paragraph 18.
10 *Wisconsin v. Loomis*, 881 N.W.2d paragraph 19.
11 *Wisconsin v. Loomis*, 881 N.W.2d paragraph 20.
12 *Wisconsin v. Loomis*, 881 N.W.2d paragraph 22.
13 Walsh and Dotter, "Longitudinal Analysis of the Effectiveness of DCPS Teachers."

contributions to student progress in terms of the influence they have on their students' annual standardized test scores.[14]

One VAM endorsed by the state legislature is the Tennessee Value-Added Assessment System (TVAAS), a proprietary system developed by SAS, a business analytics software and services company. The TVAAS system included standardized tests for students in a variety of subjects, including algebra, English, biology, chemistry, and US history. Roughly half of teachers at the time of the case taught subjects not tested under TVAAS. Nonetheless, because of the law requiring teacher evaluation on the basis of VAMs, teachers of non-tested subjects were evaluated on the basis of a "school-wide composite score," which is the average performance of *all* students on *all* subjects in that school. In other words, it is a score that is identical for all teachers in the school regardless of what subjects and which students they teach.

Teresa Wagner and Jennifer Braeuner teach non-tested subjects (physical education and art, respectively). From 2010 to 2013, each received excellent evaluation scores based on observations of their individual classes combined with their schools' composite scores. In the 2013–14 school year, however, their schools' composite scores dropped from the best possible score to the worst possible score, while their individual classroom observation scores remained excellent. The result was that Wagner's and Braeuner's individual, overall evaluations decreased from the highest possible to middling. This was enough to preclude Wagner from receiving the performance bonus she had received in previous years and to make Braeuner ineligible for consideration for tenure. Moreover, each "suffered harm to her professional reputation, and experienced diminished morale and emotional distress."[15] Nonetheless, the court determined that the teachers' Fourteenth Amendment equal protection rights were not impinged on the grounds that use of TVAAS passed the rational basis test.[16]

1.1.3 *"Exiting" Teachers:* Houston Fed of Teachers *and EVAAS*

In 2012, the Houston Independent School District ("Houston Schools") began using a similar SAS-developed proprietary VAM (EVAAS) to evaluate teachers. Houston Schools had the "aggressive goal of 'exiting' 85% of teachers with 'ineffective' EVAAS ratings."[17] And in the first three years using EVAAS, Houston Schools

[14] Isenberg and Hock, "Measuring School and Teacher Value Added in DC, 2011–2012 School Year."
[15] *Wagner v. Haslam*, 112 F. Supp. 3d.
[16] 112 F. Supp. 3d at 698. In reviewing government regulations under the Fourteenth Amendment's Equal Protection Clause, courts apply increasingly stringent levels of scrutiny (and are therefore more likely to find violations of the equal protection clause) based on types of classification used and how fundamental the right affected is. Where government regulation does not use a suspect class or affect a fundamental right, it is subject to the rational basis test. This is the least stringent level of scrutiny, and requires only that the regulation be rationally related to a legitimate government purpose. This is a high bar for plaintiffs to clear. See 16B Am Jur 2d Constitutional Law §§ 847–860.
[17] *Houston Fed of Teachers, Local 2415 v. Houston Ind Sch Dist*, 251 F. Supp. 3d at 1174.

"exited" between 20 percent and 25 percent of the teachers rated ineffective. Moreover, the district court determined that the EVAAS scores were the sole basis for those actions.[18]

As in *Wagner*, the *Houston Schools* court determined that the teachers did not have their substantive due process rights violated because use of EVAAS cleared the low rational basis standard.[19] However, the court determined that the teachers' *procedural* due process rights were infringed. Because the system is proprietary, there was no meaningful way for teachers to ensure that their individual scores were calculated correctly. The court noted that there were apparently no mechanisms to correct basic clerical and coding errors. And where such mistakes did occur in a teacher's score, Houston Schools refused to correct them because the correction process disrupts the analysis. In response to a "frequently asked question," the school district states:

> Once completed, any re-analysis can only occur at the system level. What this means is that if we change information for one teacher, we would have to run the analysis for the entire district, which has two effects: one, this would be very costly for the district, as the analysis itself would have to be paid for again; and two, this re-analysis has the potential to change all other teachers' reports (emphasis in original).[20]

That last point is worth stressing. Each teacher's individual score is dependent on all other teachers' scores. So a mistake for one teacher's score affects all others' scores. As the court states, "[T]his interconnectivity means that the accuracy of one score hinges upon the accuracy of all."[21]

1.1.4 *So What?*

Taking a step back from the specifics of the three cases, it is worth considering the impetus for decision-makers to adopt proprietary, algorithmic systems such as COMPAS, TVAAS, or EVAAS. Using sophisticated algorithms based on large datasets to help anticipate needs and better manage complex organizations like criminal justice systems and school systems makes a certain degree of sense. Human decision-makers have significant epistemic limitations, are prone to many kinds of biases, and at times act arbitrarily. And there are enormous advantages to using data-driven systems in lots of domains, generally. However, such systems have substantial problems.

A best-selling book by Cathy O'Neil describes similar systems as "Weapons of Math Destruction" because they hide harms, biases, and inadequate models behind

[18] *Houston Fed of Teachers, Local 2415 v. Houston Ind Sch Dist*, 251 F. Supp. 3d at 1175.
[19] *Houston Fed of Teachers, Local 2415 v. Houston Ind Sch Dist*, 251 F. Supp. 3d at 1183.
[20] Houston Independent School District, "EVAAS/Value-Added Frequently Asked Questions."
[21] 251 F. Supp. 3d 1168, 1178.

complicated and inscrutable veneers.[22] In another widely popular book, mathematician Hannah Fry offers a series of cautionary tales about over- and misuse of algorithmic systems, even while being optimistic about the power of such systems to do important work.[23] In a series of articles for the news organization *ProPublica*, Julia Angwin and others make the case that risk assessment algorithms used in criminal justice are racially biased.[24] Others have argued that algorithmic systems are harmful, oppressive, opaque, and reflect and perpetuate discrimination.[25]

Despite the growing literature on algorithmic harm, discrimination, and inscrutability, there remain several puzzles related to the cases we have described. Consider, for instance, *Loomis*. It is plausible that Loomis was not harmed in that he received exactly the sentence he would have received without the PSI. After all, he had a violent criminal history; the charges in the case were related to a violent, dangerous crime; and he admitted to the underlying conduct on which the charges were based. The circuit court specifically concluded that he had been driving the car when Vang fired the shotgun, that the shooting might have resulted in killing one or more people, and that Loomis had not taken full responsibility for his role. Moreover, because he is White, and the COMPAS algorithm appears to disadvantage Black[26] defendants (as we will discuss in Chapter 3), the judge's use of the COMPAS report likely did not expose Loomis to racial discrimination. Nonetheless, something seems off about using COMPAS in the case, and we will argue that he was wronged, regardless of whether his sentence was ultimately appropriate. But just how so is a difficult question.

Likewise, something seems off in the *Wagner* and *Houston Schools* cases, but it is not straightforward to pin down whether the teachers were wronged (and, if so, why). It is certainly true that some teachers were harmed in each case, but that is not enough to conclude that they were wronged. After all, any teacher that does not receive a bonus, becomes ineligible for tenure, or is laid off is harmed. But such harms are wrongful only if they are unwarranted. Moreover, it is an open question whether the VAMs used in those cases were either unfair or unjust. We will argue that the use of algorithmic systems in these cases *is* wrongful. But again, that conclusion requires substantial explanation.

[22] O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*.
[23] Fry, *Hello World: Being Human in the Age of Algorithms*.
[24] Angwin et al., "Machine Bias," May 23, 2016.
[25] Citron, "Technological Due Process"; Sweeney, "Discrimination in Online Ad Delivery"; Citron and Pasquale, "The Scored Society: Due Process for Automated Predictions"; Sweeney, "Only You, Your Doctor, and Many Others May Know"; Barocas and Selbst, "Big Data's Disparate Impact"; Calo and Rosenblat, "The Taking Economy: Uber, Information, and Power"; Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*; Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information*; Noble, *Algorithms of Oppression*; Rosenblat, *Uberland*.
[26] Regarding capitalization of "Black" and "White," we are persuaded by the arguments in Appiah, "The Case for Capitalizing the 'B' in Black."

Answering these questions is the central task of this book. And our central thesis is that understanding the moral salience of algorithms requires understanding how they relate to the autonomy of persons. Understanding this, in turn, requires that we address three broad issues: what we owe people as autonomous agents (Chapters 3 and 4), how we preserve the conditions under which people are free and autonomous (Chapters 5 and 6), and what the responsibilities of autonomous agents are (Chapters 7 and 8).

Before we go any further, let's clarify our target.

1.2 WHAT IS AN ALGORITHM?

The academic literature and wider public discourse about the sorts of systems we have been discussing involve a constellation of concepts such as "algorithms," "big data," "machine learning," and "predictive analytics."[27] However, there is some ambiguity about these ideas and how they are related, and any discussion of emerging technologies requires some ground-clearing about the key concepts. There are, however, some general points of overlap in the literature. We won't attempt to settle any taxonomical debates here once and for all, but we will fix some of the important concepts for the sake of clarity.

Among the key concepts we will use, "algorithm" is among the most important, but its usage also invites confusion. At its most basic, an algorithm is just an explicit set of instructions for solving a problem. The instructions may be for a digital computer, but not necessarily so: a recipe for chocolate chip cookies, a set of instructions for operating a combination lock, and even the familiar procedure for long division are all algorithms. In contrast to this broad concept, we are considering algorithms in terms of their functional roles in complex technological systems.[28] The term "algorithm" is also ambiguous in this more specific setting. It can be used to refer either to a set of instructions to complete a specific task or to a system that is driven by such algorithms. This distinction makes a difference in patent law. Inventions built upon an abstract mathematical algorithm (such as a special mechanical process for molding synthetic rubber) can be patented, while the algorithm itself (meaning the equations used to guide the process or system) cannot.[29]

Our focus here, however, is algorithms in the more applied, systematic sense. That is, we are concerned with algorithms that are incorporated into decision systems. These systems take a variety of forms. Some are parts of mechanical systems, for example, sensor systems in modern cars that activate warnings (e.g., for obstacles nearby) or control safety features (e.g., emergency brakes). Others are parts of information systems, for example, recommendation systems for videos (e.g.,

[27] Mittelstadt et al., "The Ethics of Algorithms."
[28] Select Committee on Artificial Intelligence, "AI in the UK: Ready, Willing and Able?" 15; Fry, *Hello World: Being Human in the Age of Algorithms*.
[29] See *Diamond v. Diehr*, 450 U.S. 175 (1981).

Netflix, YouTube), music (Spotify, Pandora), books (Amazon, Good Reads), and maps (Google maps). Still others are incorporated into complex social structures (supply chain logistics, benefits services, law enforcement, criminal justice). These systems have become ubiquitous in our lives; everything from border security to party planning is now managed by algorithms of one sort or another. When we discuss COMPAS, EVAAS, and the Facebook News Feed in one breath, we are discussing algorithms in this broad sense. Moreover, algorithms in this sense are best understood as constitutive parts of *socio-technical systems*. They are not purely sets of instructions for carrying out a task and they are not mere technological artifacts. Rather, they are used by individuals and groups and affect other individuals and groups such that they constitute an interrelated system that is both social and technological. For the remainder of the book we will refer to these kinds of systems in several ways, including "automated decision systems," "algorithmic decision systems," and (for the sake of terseness) simply "algorithms."

Another key concept is "big data." This term is often used to describe any data-mining approach to a problem using large datasets, but this washes over much of what makes such datasets a distinctive ingredient of modern technological systems. Datasets that are "big" in the sense of big data are usually enormous and high dimensional; often they consist of hundreds of thousands of rows and thousands of columns. However, a dataset that is merely big in this sense will not render the statistical magic often discussed under the rubric of predictive analytics. Rather, the systems and datasets that underlie algorithmic decision systems also have a number of other special properties.[30] These additional properties are often summarized in terms of the "three V's": *volume*, *velocity*, and *variety*. In other words, datasets that are big in the relevant sense are not only big in volume. They also have high velocity, meaning that they are often continuously updated or are created in real time, for example, systems offering driving route instructions that are updated to account for traffic conditions. Finally, they are diverse in variety, meaning that they encompass both data that is structured (i.e., organized in a predefined format), in the sense of being organized and comprehensible for analysis, and data that is unstructured (i.e., not organized in a predefined format).

As with the concepts of algorithms and big data, "predictive analytics" is not defined by a well-codified set of rules, systems, or practices. At root, the term describes the application of data-mining techniques in developing predictive models, but it is more than that. Many of the model-building techniques, such as linear regression, are standard statistical methods that have been known for hundreds of years.[31] The characteristic feature of modern predictive analytics is not its use of algorithms or even the size or complexity of its datasets, but rather the analytical possibilities offered by machine learning.

[30] Kitchin, "Big Data, New Epistemologies and Paradigm Shifts."
[31] Finlay, *Predictive Analytics, Data Mining and Big Data*, 3; Sloan and Warner, "Algorithms and Human Freedom."

Machine learning involves training computers to perform tasks according to statistical patterns and inferences rather than according to human-coded logical instructions. This approach incorporates different kinds of processes, the broadest categories of which are "supervised" and "unsupervised" learning. Supervised learning is the more straightforward and familiar of the two forms of machine learning. It involves systems that have been trained on large numbers of examples, either for classification (i.e., for classifying future examples) or for regression (i.e., for performing regression analysis). What makes the computer's learning supervised in these cases is that both classification and regression processes involve a "supervision signal," which is constructed from training on a set of pre-labeled examples and which defines the desired sort of output in advance. Classification, for instance, involves sorting novel examples into a known set of discrete values (e.g., determining whether a given image is of a cat, a dog, or a rabbit), given a set of pre-labeled training examples. Regression involves predicting some real-valued output (e.g., determining the value of a rental property in a complex market), given some set of examples.

In contrast to supervised learning, unsupervised learning involves analysis using large numbers of examples but lacks a supervision signal. Unsupervised learning algorithms, then, are not given right answers in advance for the purposes of future prediction; rather, they are designed to somehow discern or reduce the deep structure of the (often high dimensional) dataset for explanatory purposes. This can take the form of "clustering," in which the data is "naturally" grouped according to the distances between its data points, or "dimensionality reduction," in which the dataset is either compressed or broken down for intuitive visualization. In recent years, these techniques have found applications in data center regulation, social media sentiment analysis, and disease analysis based on patient clustering.

There is widespread recognition that there are ethical issues surrounding complex algorithmic systems and that there is a great deal of work to be done to better understand them. To some extent, concern about these issues is related to beliefs about the potential of unsupervised learning to help realize strong forms of AI.[32] The reality is more pedestrian.[33] Outside of cutting-edge AI labs such as OpenAI or DeepMind, machine learning is mainly a matter of employing familiar techniques such as classification, regression, clustering, or dimensionality reduction, at a big data scale. So rather than grappling with ghosts in machines that have not yet begun to haunt us, we aim to address the practical issues we already face.

[32] On its website, OpenAI describes its mission as "to ensure that artificial general intelligence (AGI) – by which we mean highly autonomous systems that outperform humans at most economically valuable work – benefits all of humanity." OpenAI, "About OpenAI." DeepMind, meanwhile, describes itself as "a team of scientists, engineers, machine learning experts and more, working together to advance the state of the art in artificial intelligence." DeepMind, "About DeepMind." For a somewhat recent book-length analysis of these issues, see Bostrom, *Superintelligence*.

[33] Marcus, "Deep Learning."

1.3 ALGORITHMS, ETHICS, AND AUTONOMY

We began this introduction by describing several recent legal disputes. *Loomis*, *Wagner*, and *Houston Teachers* will be polestar cases throughout the book. But at root, this book addresses *moral* questions surrounding algorithmic decision systems. Whether use of COMPAS violates legal rights is a distinct (though related) question from whether it impinges moral claims. Moreover, the proper scope of legal claims and how the law and legal systems ought to treat algorithmic systems are moral questions. Concerns about algorithmic systems have come from a range of sectors and include guidance from nongovernmental organizations, government agencies, legislators, and academics. For example, the UK's Nuffield Foundation published a road map for research on ethical and societal implications of algorithmic systems. They argue that there are important conceptual gaps that need to be facilitated by philosophical analysis. In their canvas of various sets of AI principles offered by scientific, engineering, corporate, and government groups, "most of the principles prosed for AI ethics are not specific enough to be action guiding."[34] Likewise, they point to a gap in the philosophical literature on ethics in algorithms, data, and AI.[35]

Government entities have also recognized moral concerns and the need for greater research on these issues as well. The US President's National Science and Technology Council's 2016 report, "Preparing for the Future of Artificial Intelligence," outlined a number of ethical concerns surrounding AI and algorithmic systems.[36] While the report focuses on transparency and fairness, the issues it raises have autonomy implications as well. The Ethics Advisory Group to the European Data Protection Supervisor (EDPS-EAG) issued a report in 2018 outlining a slate of ethical concerns surrounding digital technologies, including algorithmic decision systems. In particular, the advisory group explained the importance of linking foundational values – among them autonomy, freedom, and democracy – to digital technologies. The UK parliament appointed a Lords Select Committee on Artificial Intelligence in 2017 to examine a handful of issues in development and adoption of AI (within which they include algorithmic systems), one of which is "What are the ethical issues presented by the development and use of artificial intelligence?"[37] Among their recommendations are principles protecting "fairness and intelligibility" and prohibiting automated systems from having the power to "hurt, destroy, or deceive human beings."[38] Members of both houses of the U.S. Congress have introduced an Algorithmic

[34] Whittlestone et al., "Ethical and Societal Implications of Algorithms, Data, and Artificial Intelligence: A Roadmap for Research," 11.
[35] Whittlestone et al., 46–47.
[36] National Science and Technology Council, "Preparing for the Future of Artificial Intelligence."
[37] Select Committee on Artificial Intelligence, "AI in the UK: Ready, Willing and Able?" 12.
[38] Select Committee on Artificial Intelligence, 125. Related reports and recommendations have come from Japanese Society for Artificial Intelligence, "Ethical Guidelines"; Association for Computing Machinery, US Public Policy Council, "Statement on Algorithmic Transparency and Accountability"; Campolo et al., "AI Now 2017 Report."