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).

https://doi.org/10.1017/9781108895057.001 Published online by Cambridge University Press

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."

https://doi.org/10.1017/9781108895057.001 Published online by Cambridge University Press

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.

https://doi.org/10.1017/9781108895057.001 Published online by Cambridge University Press

"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.

https://doi.org/10.1017/9781108895057.001 Published online by Cambridge University Press

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."

https://doi.org/10.1017/9781108895057.001 Published online by Cambridge University Press

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).

https://doi.org/10.1017/9781108895057.001 Published online by Cambridge University Press

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."

https://doi.org/10.1017/9781108895057.001 Published online by Cambridge University Press

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."

https://doi.org/10.1017/9781108895057.001 Published online by Cambridge University Press

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."

https://doi.org/10.1017/9781108895057.001 Published online by Cambridge University Press

Accountability Act that would impose requirements to create impact assessments and address bias and security issues.[39]

The academic literature is also expanding in its criticism of algorithmic systems. Kirsten Martin, for instance, argues that big data generates negative externalities in the form of additional surveillance, which she calls "surveillance as pollution."[40] Tal Zarsky argues that automated decision-making introduces both efficiency- and fairness-based problems.[41] Danielle Citron and Frank Pasquale argue for imposing auditing procedures on the process of algorithmic development on the basis of those problems.[42] Karen Yeung argues that the inoffensive practice of nudging, which classically involves only subtle forms of behavioral modification through manipulation of "choice architectures," can be galvanized by predictive analytics to produce "hypernudging" platforms whose effects at scale wind up being radically paternalistic.[43] Cathy O'Neil groups such systems under the banner of "weapons of math destruction,"[44] arguing that they enjoy an aura of epistemic respectability that encourages us to use them beyond their actual capacities. Citron argues that addressing these problems requires nothing short of a new constitutional paradigm – a new "technological due process."[45]

Algorithmic systems (including both predictive systems and digital platforms) have come under substantial economic, political, and philosophical criticism. We agree with much of it. However, for a few reasons we do not defend any overall moral or ethical conclusion about the technologies themselves. First, the fact that they are rapidly advancing as part of an ongoing process means that the horizon for productive commentary on current technology is time delimited. Second, we acknowledge that these technologies – whatever their life spans might be – can be employed for useful aims as well as for pernicious ones. There are few global, all-things-considered moral judgments that can be made about, for instance, the governance of Facebook's News Feed or use of risk assessment algorithms. Third, we acknowledge that the algorithmic landscape of predictive analytics and digital platforms is here to stay in some form or other. It is possible to exert some influence on how these systems are employed and perhaps even develop new conceptions of fair play to cope with these changes, but predictive analytics and digital platforms will not be eliminated altogether.

For those reasons we aim to look beyond the particular features of the technologies as much as possible, treating the technologies themselves as case studies that are useful for making certain moral and social issues vivid and concrete, rather than as the sources of ontologically distinctive philosophical issues. In many cases, the philosophical issues have more to do with our psychological features and our social

[39] Algorithmic Accountability Act of 2019, H.R. 2231; Algorithmic Accountability Act of 2019, S. 1108.

[40] Martin, "Ethical Issues in the Big Data Industry," 75.

[41] Zarsky, "The Trouble with Algorithmic Decisions."

[42] Citron and Pasquale, "The Scored Society: Due Process for Automated Predictions."

[43] Yeung, "'Hypernudge': Big Data as a Mode of Regulation by Design." See also Lanier, *Ten Arguments for Deleting Your Social Media Accounts Right Now*.

[44] O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*.

[45] Citron, "Technological Due Process."

https://doi.org/10.1017/9781108895057.001 Published online by Cambridge University Press

structures than with the inherent hazards of technological systems considered abstractly. Hence, what unifies the systems discussed in the book is not strictly about the technologies per se, but rather about the human values that are implicated by the designs and modes of operation of all of those socio-technical systems. All these systems, we argue, raise philosophical issues about *autonomy*.

Focusing on autonomy is important for several reasons. Primarily, it connects unease about algorithmic systems to a good that has deep, stable, and plausible moral value. This is a rich seam to mine in the same way that autonomy issues have been fundamentally important in other domains of applied ethics. Arguments grounded in autonomy connect concerns about algorithmic systems to an area with a broad and well-developed philosophical literature.

Moreover, by drawing out the importance of autonomy, our account can address concerns about algorithmic decision systems that are not captured by accounts that focus on fairness, harm, and bias. Algorithmic systems often reflect, harden, and create unfair structures; this is an enormous moral concern. However, that is only part of the moral importance of algorithmic systems. *Loomis* shows why: Loomis was plausibly wronged, but it is not clear that he has been treated unfairly (at least in the sense that COMPAS treats him differently from other, similarly situated defendants) and it does not appear that he has been materially harmed by use of COMPAS. Note, too, that while there are some scholars who have addressed whether certain kinds of algorithmic systems conflict with autonomy via manipulation,[46] our view is that autonomy is grounds for a much broader evaluation of algorithmic systems. *Loomis*, *Wagner*, and *Houston Schools* do not appear to involve manipulation in any strong sense.

Our focus on autonomy also provides a foundation for moral concerns that are often under-explained (e.g., transparency, filter bubbles). Specifically, a focus on autonomy can serve to route around at least some disputes about bias and fairness. It has become clear that there are different ways that a single system can be plausibly unfair to members of different groups.[47] Determining which facets of fairness matter the most requires considering different values. We will argue that an appeal to autonomy can help in that regard. Finally, note that our approach is consistent with other important critiques; that is, a concern about autonomy for the most part *adds to* rather than contradicts extant critiques.

[46] For accounts addressing algorithmic systems and autonomy, see Yeung, "'Hypernudge': Big Data as a Mode of Regulation by Design"; Lanzing, "'Strongly Recommended' Revisiting Decisional Privacy to Judge Hypernudging in Self-Tracking Technologies"; Danaher, "The Threat of Algocracy: Reality, Resistance and Accommodation"; Danaher, "Toward an Ethics of AI Assistants"; Susser, Roessler, and Nissenbaum, "Online Manipulation: Hidden Influences in a Digital World." They most often address direct effects of automated systems on individuals' decision procedures. Certainly, such cases are important from the standpoint of autonomy. However, our aim here is to address a broader range of issues surrounding autonomy. We address this issue further in Chapter 5.

[47] Binns, "Fairness in Machine Learning: Lessons from Political Philosophy"; Binns, "Algorithmic Accountability and Public Reason"; Corbett-Davies and Goel, "The Measure and Mismeasure of Fairness."

https://doi.org/10.1017/9781108895057.001 Published online by Cambridge University Press

1.4 OVERVIEW OF THE BOOK

To get at the different ways in which algorithmic systems bear upon autonomy, we have divided the book into four main parts. Part I is introductory and ground-clearing. Chapters 1 and 2 serve as introductory chapters, outlining the conceptual foundations and philosophical commitments that ground our arguments throughout the book. The primary task of Chapter 2 is providing an account of autonomy and its importance. It begins with a high-level explanation of autonomy itself and then canvasses several key conceptions of autonomy and the philosophical concerns underlying them. We then advance our ecumenical account of autonomy. Specifically, we draw on procedural, psychological accounts of autonomy as well as social accounts of personal autonomy. We argue that while they have important differences in how they explain key facets of autonomy, in practice they are substantially overlapping. Hence, we can draw on both as a foundation for the arguments we make in the book. That is because fully realized autonomy demands both procedural independence (which includes both epistemic competence conditions and authenticity conditions) and substantive independence (which includes conditions of reciprocal support and non-domination from society more broadly).

Part II builds upon our account of autonomy and our polestar cases to address the kinds of moral claims and the nature of the respect that persons are owed in virtue of their autonomy. In Chapter 3 we argue that people are owed systems of rules and practices that they can reasonably endorse. We begin the chapter with a closer consideration of VAMs used for K-12 teacher assessment. We argue that the problem with such tools cannot be reduced to concerns about their reliability or their potential for bias. Rather, teachers have a claim to incorporate their values into their lives as they see fit. And respecting teachers requires recognizing them as value-determiners, neither thwarting nor circumventing their ability to act according to those values without good reason. Moreover, as agents they are capable of abiding fair terms of social agreement (so long as others do too), and hence "good reasons" will be reasons that they can abide as fair terms of cooperation. Teachers can endorse those reasons either as consistent with their own values or as a manifestation of fair social agreement.

We argue that VAMs fail to respect teachers as agents because they are used in a way that teachers cannot reasonably endorse, for four interrelated reasons. First is their reliability, which many have questioned. Second is that their results can be based on factors for which teachers are not morally responsible (e.g., student performance may correlate with teachers' age, ethnicity, or gender). Third is stakes. The fact that an algorithmic system is unreliable or measures factors for which persons are not responsible is important primarily as a function of the stakes involved. Fourth is the relative burdens placed upon people subject to them. We conclude by applying our framework to our polestar cases (*Loomis*, *Wagner*, and *Houston Schools*).

https://doi.org/10.1017/9781108895057.001 Published online by Cambridge University Press

One important, oft-cited criticism of algorithmic systems is that they lack transparency. Such systems can be opaque because they are complex, protected by patent or trade secret, or deliberately obscure. In the EU, there is a debate about whether the General Data Protection Regulation (GDPR) contains a "right to explanation," and if so what such a right entails. Our task in Chapter 4 is to address this informational component of algorithmic systems. We argue that information access is integral for respecting autonomy, and transparency policies should be tailored to advance autonomy.

To make this argument we distinguish two facets of agency (i.e., capacity to act). The first is *practical* agency or the ability to act effectively according to one's values. The second is what we call *cognitive* agency, which is the ability to exercise what Pamela Hieronymi calls "evaluative control" (i.e., the ability to control our affective states such as beliefs, desires, and attitudes). We argue that respecting autonomy requires providing persons sufficient information to exercise evaluative control and properly interpret the world and one's place in it. We draw this distinction out by considering algorithmic systems used in background checks, and we apply the view to our polestar cases.

While Part II of the book considers what we owe people *given that they have autonomy*, Part III addresses our responsibility to secure the conditions under which people can *act autonomously*. Chapter 5 considers the relationship between algorithmic systems and freedom. There is substantial dispute about the concept and moral value of freedom. A key area of dispute is whether freedom is best understood in terms of negative, positive, or republican freedom. We offer an account according to which freedom is *ecological* and includes both republican freedom (which is to say, freedom from others' exercise of arbitrary power) and positive freedom, properly understood (i.e., where positive freedom is a function of quality of agency). We argue that algorithmic systems in several ways conflict with ecological freedom.

Chapter 6 addresses a specific condition of autonomy from Chapter 2, viz., epistemic competence. It has been clear since the early 2000s that internet communication technologies generally and algorithmically driven information systems in particular create epistemically noxious environments. These include phenomena like filter bubbles and echo chambers as well as more insidious phenomena like toxic recommendation systems.[48] Most every media platform now employs content moderation systems to screen for the truly awful content that would make sites like YouTube, Facebook, Reddit, and others unnavigable.[49] That practice is relatively

[48] Among the most odious of these are YouTube algorithmic recommendation systems that serve disturbing (including violent and sexualized) content on channels specifically designed and marketed for children. See Maheshwari, "On YouTube Kids, Startling Videos Slip Past Filters"; Orphanides, "Children's YouTube Is Still Churning out Blood, Suicide and Cannibalism."

[49] For an overview of the enormous labor and the labor practices involved in content moderation, see Roberts, *Behind the Screen*.

https://doi.org/10.1017/9781108895057.001 Published online by Cambridge University Press

uncontroversial. We argue that there are further obligations on sites to exercise a kind of epistemic paternalism.

While Parts II and III begin from the premise that people have certain claims as agents (specifically, claims to the conditions that foster autonomy), Part IV shifts focus to the obligations *of* agents. Chapter 7 considers the autonomy and responsibility of those who deploy information technologies (as collectors of big data, users of algorithmic decision systems, developers of social media sites, and so on). Specifically, we argue that there is a type of wrong that arises when autonomous agents obscure responsibility for their actions, which we call "agency laundering." At root, agency laundering involves a failure to meet one's moral responsibility for an outcome by attributing causal responsibility to another person, group, process, or technology, and it does so by undermining a key component of responsibility itself, viz., accountability. We apply our conception of agency laundering to a series of examples, including Facebook-automated advertising suggestions, Uber driver interfaces, as well as to our polestar cases.

We then turn to the ways in which autonomy underwrites democratic governance. Political authority, which is to say the ability of a government to exercise power, may be justifiable or not. Whether it is justified and how it can come to be justified is a question of political *legitimacy*. In Chapter 8 we consider several views of legitimacy and argue for a hybrid version of normative legitimacy based on one recently offered by Fabienne Peter.[50] We explain that each facet of the hybrid requires a legitimation process that is itself grounded in autonomy. We argue that the autonomy view is a basis for criticism of the legitimation processes related to a predictive policing technology and to actions of Cambridge Analytica and the Internet Research Agency in recent elections. In Chapter 9 we offer some conclusions and caveats.

1.5 A HEURISTIC

It is worth stepping back for a moment before launching into the main substance of the book's arguments in order to explain its purpose and approach. This book is about ethics and algorithmic decision systems, such as COMPAS, EVAAS, YouTube and other recommendation systems, Facebook Ad services, and others. Its aim is to better understand moral concerns about those systems. So we will consider questions such as the following: Is it morally justifiable for a court to use an algorithmic system such as COMPAS in determining whether, and if so for how long, to sentence a defendant to incarceration? Should school systems use algorithmic systems such as EVAAS to promote, reward, and fire teachers? Does YouTube have an obligation to better police the videos that are suggested to viewers?

[50] Peter, "The Grounds of Political Legitimacy."

https://doi.org/10.1017/9781108895057.001 Published online by Cambridge University Press

These questions are multifaceted, and there are lots of different ways we might construe them. There is a meta-ethical issue of what the "should" and "justifiable" in the questions mean or refer to.[51] There are questions of legal doctrine such as whether the use of a system like COMPAS impinges upon due process rights and whether algorithmic systems violate contract terms or statutory protections. Those kinds of questions have ethical implications and will come up now and again in the book, but they are not the focus.

Instead we will focus on ethics (which we use interchangeably with "morals"). How, though, does one do that? There is no univocal recipe for answering moral questions. However, we can offer a heuristic for evaluating moral questions.[52] It is not the only way to think through problems, and it is useful in part because it shows just how difficult resolving some moral problems actually is. Nonetheless, it is a way to keep oneself in check and (perhaps more importantly) letting others understand what exactly one is doing. The hope is that following this kind of heuristic will help readers recognize what (if anything!) an ethical argument contributes and what its limitations are.

The first step is clarifying relevant concepts. In considering whether it is justifiable to use algorithmic systems in criminal justice decisions, for example, we will need to specify a number of concepts in order to make progress. An obvious one is "algorithmic system," which we have tried to clarify in Section 1.2. Do we mean literally *any* algorithm, including an ink-and-paper set of sentencing guidelines? Or are we talking about only sophisticated, big data–driven systems? Or even machine learning–based systems? Another concept to clarify is "criminal justice decisions." That could mean decisions about sentencing, supervision, early release, or something else altogether. As we will see in our discussion of COMPAS throughout the book, those issues matter. We will also spend substantial time working to clarify moral concepts. These include "autonomy," "agency," "fairness," "freedom," "legitimacy," and others.

The next step is to get one's empirical facts[53] straight. Of course, this is easier said than done. If we want to address the question of whether some use of algorithmic

[51] For example, one might ask whether the claim that X should happen is merely expressive of support for a view, or whether it seeks to say something true about the world, but is mistaken because there are no such facts, or the like. We won't take up that debate here, though we invite anyone who *would* like to have that debate to attend the Madison Metaethics Workshop at the University of Wisconsin-Madison each fall.

[52] Here, we offer a modified version of one Tom Regan outlines in Regan, "Introduction to Moral Reasoning." He offers what he calls a process for "ideal moral judgment," which strikes us as overly optimistic. We'll call it instead a "heuristic for better moral judgment." That may also be overly optimistic, even if substantially less so than Regan.

[53] One might call these "non-moral facts," though that has the potential to derail us. At least one of the authors is adamant that all facts are non-moral. At least one tentatively believes that moral facts are types of natural facts. And at least one has argued that there are moral facts and those facts are nonnatural. It is possible that some of these describe multiple authors and that some authors are described by more than one of these sentences. What matters for our purposes here is that there are

https://doi.org/10.1017/9781108895057.001 Published online by Cambridge University Press

systems in criminal justice decisions is justifiable or not, we will want to know something about how such systems operate, what effects they have, and what the underlying picture of the criminal justice system looks like. We might want to know something about how people are likely to use such systems; how those systems interact with social structures like courts, prisons, parole boards, and the like. There is a world of facts surrounding any moral question, some of which are contested, some of which are unknowable, and some of which people get wrong. Regardless, any moral claim will rest on some understanding of the facts on the ground, and getting those straight, getting a clear-eyed view of what facts are tenuous and unknown, and having a sense of what we would *like* to know is vital in addressing moral questions.

Step three is discerning and applying the correct moral theory to the questions at hand. A few thousand years of religious and philosophical disputes have not answered the question of which moral theories are correct. It is debatable as to whether (and to what extent and by what metric) there has been progress, and there is an important debate about the degree of cultural convergence and divergence on fundamental questions of right and wrong. However, the entire premise of the project of debating the justifiability of anything – using animals for food, abortion, policing, capital punishment, universal suffrage, civil disobedience, mandating vaccinations, and, our topic, applying algorithmic systems in various contexts – is that there can be better and worse (i.e., more and less morally justifiable) answers. And hence, we need some basis on which to find our way on these kinds of questions.

Of course, the question of *which* moral theory is correct comprises its own, vast area of inquiry. Among the possibilities are consequentialism, deontology, virtue ethics, and contractarianism, each of which has myriad versions. And even once one has some set of moral values one believes to be correct, it is yet another daunting task to figure out just how to apply them. If one is a thorough-going consequentialist who believes that an action is right just in case that act leads to the best overall consequences, it remains an open question of how to figure out how to apply that to actions on the ground.

There are, however, a number of views of morality that are critical of the notion that there are moral principles to apply in the first place, even while accepting that actions can be more or less justified. Some of these views are forms of pragmatism, holding that the practice of theorizing is inseparable from practical consideration. Others adhere to "casuistry," according to which one can find justification by comparing contested situations to similar "paradigmatic" cases. Still others are forms of moral particularism, according to which general moral principles (e.g., "lying is bad," "killing is wrong," "people should be treated equally") carry no moral weight and provide no role in explaining why actions are right or wrong. Rather,

issues about metaethics that lurk underneath the surface of any project in applied ethics. Nonetheless, we can make progress without resolving those underlying questions.

https://doi.org/10.1017/9781108895057.001 Published online by Cambridge University Press

particular acts are right or wrong, and any generalities we draw from particular acts are neither explanatory nor applicable to other cases. But even these "anti-theory" positions continue to insist that there is an account of moral justification (even if it is not one based on explanatorily prior moral principles). And so, if you are searching for the answer, as we are, to the question of whether using algorithmic systems under certain conditions is right or wrong, then you are committed at some level (and perhaps only implicitly) to there being better and worse accounts of moral justifiability on which to base such judgments. We should therefore understand "moral theory" in this third step of the heuristic (i.e., discern and apply the correct moral theory) to be quite broad. It is not only the discernment and application of moral principles but also (or, on the anti-theory views, rather) the discernment and application of the correct understanding of moral justification.

The final part of the heuristic is to apply principles of critical reflection to the concepts, empirical claims, and moral theories one has at one's disposal. This does *not* imply that one can simply apply rote, bloodless, "pure" reason to problems. That would be impossible, and it would be unwarranted in light of the inevitable conceptual questions, factual lacunae, and steep burdens of determining the correct moral theories. Rather, it requires a much more limited (but no less important) set of constraints. One should follow one's reasons where they lead. If the concepts, empirical claims, and moral theory entail that some action is impermissible, one should recognize that entailment and either accept the conclusion or aim to look for the flaw in the premises that led one there. Another element of good reasoning is to recognize just how fallible reasoning is; in other words, one should avoid dogmatism. Related is to accept that one's conceptual apparatus, understanding of facts, and moral theory are revisable. That's a good thing. After all, we are bound to have mistaken beliefs and commitments.

This is a project in applied moral philosophy and it aims to use the heuristic we have outlined. It considers some questions that are legal, and it adopts some factual claims. Its contribution comes in several parts. First, it clarifies some relevant concepts, including (among others) autonomy, agency, freedom, paternalism, responsibility, and legitimacy. Second, as to the empirical facts, we largely take our cues from others, drawing on court cases and related documents, other academics' empirical work on algorithmic systems, and journalists' investigations on related issues. Third, we advance some claims of normative moral theory. Most of these are grounded in autonomy. However, we do not advance the view that autonomy is the only moral value relevant to analyzing and evaluating algorithmic systems. Rather, our view is that autonomy is an important value, and many moral concerns about algorithmic systems are best understood as, at bottom, issues of autonomy.

Put slightly differently, our project is based on the premise that people are, to some degree, and within important limitations, able to govern themselves. They can determine what kinds of things matter to them, how to use those values to order their lives, and come to agreements with others about how to navigate differences in belief

https://doi.org/10.1017/9781108895057.001 Published online by Cambridge University Press

about values. The capacity for autonomy, we argue, gives rise to moral claims that others respect (to some degree and within important limits) persons' decisions about self-government. Our task, then, is to craft a set of considerations surrounding algorithmic systems based on autonomy. There are many other considerations. Some involve consequences. Some are about the law, including to what extent the law itself provides moral reasons. Yet another is virtue. There are also deeper questions of justice. Still others involve religious and quasi-religious issues. Others involve the proper scope of freedom for technologists to develop and implement systems. Still others involve trade secrets and capitalism. There is no way to adequately address all of these in a volume like this. But we submit that many potential readers will agree with our rock-bottom premise that autonomy matters. So we will start there, clarifying the concept and positing some moral principles that rest on it. On to Chapter 2.

https://doi.org/10.1017/9781108895057.001 Published online by Cambridge University Press