Canadian Journal of Philosophy (2022), 52: 3, 321–341
doi:10.1017/can.2020.27

ARTICLE

On Predicting Recidivism: Epistemic Risk, Tradeoffs, and Values in Machine Learning

Justin B. Biddle

School of Public Policy, Georgia Institute of Technology Atlanta, GA, USA
Email: justin.biddle@pubpolicy.gatech.edu

Abstract

Recent scholarship in philosophy of science and technology has shown that scientific and technological decision making are laden with values, including values of a social, political, and/or ethical character. This paper examines the role of value judgments in the design of machine-learning (ML) systems generally and in recidivism-prediction algorithms specifically. Drawing on work on inductive and epistemic risk, the paper argues that ML systems are value laden in ways similar to human decision making, because the development and design of ML systems requires human decisions that involve tradeoffs that reflect values. In many cases, these decisions have significant—and, in some cases, disparate—downstream impacts on human lives. After examining an influential court decision regarding the use of proprietary recidivism-prediction algorithms in criminal sentencing, *Wisconsin v. Loomis*, the paper provides three recommendations for the use of ML in penal systems.

Keywords: Values in science and technology; artificial intelligence; fairness; justice; equality; ethics; race

1. Introduction

In February 2013, Eric Loomis was arrested in Wisconsin while driving a car that had been used in a drive-by shooting. He was charged with five criminal counts, including involvement in the shooting. He denied involvement but pled guilty to two minor offenses of attempting to flee an officer and operating a vehicle without the owner's consent. As part of a plea agreement, the other charges would be dismissed but considered during sentencing. The court sentenced him to incarceration for sixteen and a half years. Loomis appealed his sentence (State v. Loomis 2016, 756).

This case might be unremarkable but for the fact that a consideration in the judge's sentencing decision was an output by an algorithm called the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) developed by Equivant (formerly Northpointe, Inc.), a for-profit company with offices in the United States and Canada. The algorithm inputs data about an individual who has been arrested, and it outputs an assessment of the risk of the individual to recidivate (reoffend). How exactly the algorithm operates is unknown outside of Equivant, as the workings of the system are considered a trade secret. Loomis's appeal of the sentencing decision was based on concerns about COMPAS. The appeal was taken up and rejected by the Wisconsin Supreme Court. Loomis appealed this decision to the US Supreme Court which declined to hear the case, effectively letting the Wisconsin decision stand. The details of that decision will be discussed in section 5 of this essay.

Algorithms, including machine learning (ML) and artificial intelligence (AI) systems, are now ubiquitous in contemporary societies, and they increasingly perform complex decision-making tasks once reserved for human beings. In addition to being used in penal systems, ML influences

© The Author(s) 2020. Published by Canadian Journal of Philosophy.

https://doi.org/10.1017/can.2020.27 Published online by Cambridge University Press

decisions about whether or not to approve mortgage loan applications, how to evaluate the performance of teachers, which new employees to recruit, and many other decisions that have life-changing impacts (e.g., O'Neil 2016).

The use of AI for such tasks is sometimes justified by the supposed neutrality of algorithms. On this view, algorithms are "stabilizers of trust, practical and symbolic assurances that their evaluations are fair and accurate, and free from subjectivity, error, or subjective influence" (Gillespie 2014, 179). The neutrality thesis is untenable, as I will show in this essay, but it is perhaps plausible to think that algorithms, while not completely neutral, could be less biased than the human beings who would otherwise perform these tasks. On this account, for example, algorithms that assess risk of recidivism might not be value neutral, but they could be less influenced by racial and ethnic biases than human judges.

Recent scholarship in philosophy of science and technology, science and technology studies (STS), and related fields has shown that scientific and technological decision making—including decisions about the validation of hypotheses—is laden with values, including values of a social, political, and/or ethical character. This paper draws upon this work—especially work on inductive and epistemic risk, as well as on interdisciplinary research on ML—to argue that ML systems are value laden in ways similar to human decision making because the development and design of ML systems requires human decisions that involve tradeoffs that reflect values. Given this, the important questions surrounding ML systems are not whether they are value laden (they are) but *how* they are value laden—whose interests they serve, whose values they reflect, and how they might by designed and implemented to effect positive change. In making this argument, I aim both to elucidate the role of values throughout the development and design of ML systems and contribute to the philosophical literature on the role of values in science and technology by illustrating and extending the framework of epistemic risk.

The structure of the paper is as follows. After providing some background on ML systems in section 2, I will employ the concept of epistemic risk in order to examine the role of value judgments in the development and design of ML systems generally (section 3) and in recidivism-prediction algorithms more specifically (section 4). Algorithms are used in policing and in penal systems in many parts of the world, including England and Canada, but they are used widely, and perhaps most controversially, in the United States (e.g., Kehl, Guo, and Kessler 2017). The United States has one of the highest incarceration rates in the world; a disproportionate number of incarcerated are people of color, and racial and ethnic discrimination is pervasive throughout the policing and penal systems (Alexander 2012).[1] Many, including the White House Office of Science and Technology Policy, have proposed that AI systems could help lower incarceration rates and reduce racial and ethnic bias (e.g., Shueh 2016). To explore these possibilities, I examine how recidivism-prediction algorithms are developed and designed, and I highlight the steps at which researchers must make judgments that involve tradeoffs that reflect values. More specifically, I argue that developers must navigate epistemic risk that reflects values at (at least) the following stages: (1) problem identification and framing, (2) data decisions and model competencies, (3) algorithm design: accuracy and explainability, (4) algorithm design: conceptions of fairness, (5) algorithm design: choices of outputs, and (6) deployment decisions about transparency and opacity. I build upon this discussion in section 5 to propose three recommendations for the use of ML in penal systems.

2. Algorithms and machine learning

An algorithm, as I use the term, is a well-defined series of procedures that transforms an input into an output. A machine-learning (ML) algorithm, in contrast to a traditional algorithm, is one that "learns for itself" in a bottom-up manner on the basis of data. For example, ML-based prediction algorithms identify predictor variables and their relative weights on the basis of training data—as

[1] See, for example, data from the World Prison Brief, https://www.prisonstudies.org.

https://doi.org/10.1017/can.2020.27 Published online by Cambridge University Press

opposed to, say, decisions made in advance by designers. "Machine learning works to uncover patterns in data, to build and refine representative mathematical models of data that can be used to make predictions and/or describe data to gain knowledge and insight" (Singh et al. 2016, 1–2).

There are many different types of ML algorithms, including decision trees, random forests, and deep neural networks, and there are different ways in which these systems learn. Supervised learning involves the training of an algorithm on the basis of data that is labeled by human beings (e.g., images of cats and other animals that are labeled as "cats" and "not cats," respectively). Unsupervised learning involves looking for structure in data sets that are unlabeled (such as how different images might cluster together). Reinforcement learning involves a system performing an operation (e.g., playing a video game), obtaining feedback on the basis of that operation (such as a "reward" for a scored point), and then using that feedback to revise itself. ML capabilities have advanced rapidly in recent years, allowing algorithms to assist or even replace human decision making in areas as diverse as entertainment, finance, education, medicine, and law.

There are several factors that have led to the explosion of interest in, and development of, ML algorithms. One is the emergence of enhanced abilities to collect massive amounts of data, leading to the so-called "big data" revolution. According to some estimates, 90 percent of the world's data that has ever been collected has been collected in the past two years (Petrov 2019). The enhanced capacities to collect vast amounts of data, in turn, incentivized organizations to develop new and useful ML systems to process and use these data, and they created the conditions that allowed for this, in the form of large data sets could be used to train and enhance ML systems. Increases in computational power and abilities to transfer and store data have increased the speeds with which ML algorithms can be trained. Finally, there have been advances in machine learning techniques themselves, including the emergence of deep-learning methods.

Deep-learning methods involve the use of artificial neural networks, inspired by systems of neurons in the brain, that achieve learning through multiple levels of representation (LeCun, Bengio, and Hinton 2015). They consist of an input layer, one or many "hidden" layers, and an output layer. The lines that connect the nodes of the input layer with the nodes of the first hidden layer represent weights given to each input data. The nodes of the hidden layers represent an "activation function" that takes as inputs the values of the previous nodes; the outputs of the first hidden layer are input into the next hidden layer, and so on, until the output layer is reached and a final output is given. In the case of supervised learning, the system learns by computing the distance between the desired outputs (for example, the true values according to a test set of data) and the predicted outputs and then adjusting its parameters accordingly. This iterative process can be run millions of times, with each run resulting in an adjustment of the system's parameters (and, thus, a change in the algorithm). This process continues until designers determine that the parameters have been optimized sufficiently for the purposes for which the system will be used; at this point, the training stage has ended, and the algorithm is fixed.

3. ML systems, tradeoffs, and values

Recent work in the philosophy of science and technology, as well as other fields, makes clear that scientific and technological decision making are ineliminably laden with values including, in many cases, values of a social, ethical, or political nature. ML systems are not different in this respect. Philosophers have developed a number of arguments and frameworks for conceptualizing the role of values in science and technology (e.g., Biddle 2013; Brown 2013, 2020; Douglas 2000, 2009; Elliott 2011; Havstad 2020; Longino 1990, 2002; Okruhlik 1994; Rudner 1953; Solomon 2001). In the context of ML, a helpful way to frame discussions of values is in terms of epistemic risk, which is defined broadly as the risk of error or failure that can arise anywhere in knowledge-productive practices (e.g., Biddle and Kukla 2017). It is impossible to achieve all epistemic goals simultaneously. Individuals or organizations that are engaged in the pursuit of knowledge must make decisions about which epistemic goals they wish to prioritize and which they may de-emphasize. Decisions to

https://doi.org/10.1017/can.2020.27 Published online by Cambridge University Press

prioritize some epistemic goals over others—or, to put it another way, decisions to avoid some epistemic failings at the expense of others—involve tradeoffs that reflect value judgments. That these value judgments can be ethical or political in nature is evident by the fact that different epistemic failings can impact different stakeholders in different ways; because of this, to make a judgment that some mistakes or failures are more tolerable than others is to make a judgment about whom one is more or less willing to impact and how one is willing to impact them.

One type of tradeoff that has been discussed extensively in the philosophical literature on values in science is that between Type I errors (accepting a hypothesis when it is false) and Type II errors (rejecting a hypothesis when it is true) (e.g., Douglas 2000, 2009, 2017; Elliott 2011). If we define inductive risk as the risk of wrongly accepting or rejecting a hypothesis on the basis of statistical evidence, then it is, in most cases, impossible to balance the risk of Type I and Type II errors in a value-neutral way (Wilholt 2009).[2] Decisions about which type of error one is more (or less) willing to tolerate will reflect a set of values. For example, a decision to set relatively high evidential standards before accepting the hypothesis that a drug has a serious side effect reflects value judgments about which types of mistakes one is more or less willing to make (a relative tolerance of Type II over Type I errors) (Rudner 1953). Additionally, depending on the circumstances, it could reflect judgments about whom one is more or less willing to impact and how one is willing to impact them (relative tolerance of exposing users to risk of harm versus exposing drug makers to financial risk). In addition to Type I and Type II errors, there is a third type of epistemic failing that one might consider under the umbrella of inductive risk, namely the risk of failing to generate results at all (the risk of ignorance) (Wilholt 2013). That is, in addition to accepting a hypothesis when it is false and rejecting a hypothesis when it is true, one might also perform an investigation and remain ignorant as to whether the hypothesis is likely to be true or false because the investigation did not provide enough, or the right kind of, evidence. Given that resources are limited, tradeoffs must be made between the risks of these three types of epistemic failings, and how these tradeoffs are made reflects values.

Over the course of the ML development cycle, there are additional tradeoffs and epistemic risk judgments—including many that are distinct from inductive risk judgments—that must be made, including decisions regarding problem identification and framing, training and test data, design, building, testing, deployment, and monitoring (*cf.* Morely, Floridi, and Elhalal 2019).[3] Throughout this cycle, developers must make "contingent decisions," or decisions that could be made in different ways, that impact overall performance, users, and publics.[4] In the remainder of this section, I will discuss a number of these decisions, specifically: (1) problem identification and framing, (2) data decisions and model competencies, (3) algorithm design: accuracy and explainability, (4) algorithm design: conceptions of fairness, (5) algorithm design: choices of outputs, and (6) deployment decisions about transparency and opacity. I will also address the question of whether these are "in practice" or "in principle" tradeoffs. In some of these cases, I will argue that even with unlimited resources and future research, the tradeoffs are inescapable.

3.a Problem identification and framing

The first stage in developing an ML system is to identify a problem or task to undertake, and to frame that problem by, for example, delimiting what falls inside and outside the scope of the

[2] More specifically, Wilholt shows that it is impossible to balance the risk of false positives and false negatives in a value-neutral way under the assumption that researchers aim not solely to accept true hypotheses and reject false hypotheses. Thus, for example, if researchers aim to accept significant truths where significance is a value-laden concept, then it is impossible to balance the risk of false positives and false negatives in a value-neutral way.

[3] For arguments that there are epistemic risks that are distinct from inductive risk, see Biddle (2016; 2018; 2020) and Biddle and Kukla (2017).

[4] The term *contingent decisions* is due to Brown (2020).

https://doi.org/10.1017/can.2020.27 Published online by Cambridge University Press

problem.[5] Consider, for example, the decision by Washington, DC, school systems in 2007 to attempt to improve underperforming schools by developing algorithms that evaluate teachers—an example that Cathy O'Neil discusses in the introduction of her *Weapons of Math Destruction* (2016). Given that resources are limited, the decision to develop these algorithms involves tradeoffs because there are opportunity costs involved in *not* intervening in other ways. Moreover, the decision to develop algorithms that evaluate teachers constitutes framings that involve tradeoffs. The decision suggests that the problem of underperforming schools is (to a significant extent) a problem of underperforming teachers, as opposed to (say) problems about class size or broader socioeconomic issues, and the decision to develop algorithms that evaluate teachers requires that designers operationalize the concept of "good (or bad) teacher" by specifying evaluative metrics. This decision, in turn, involves tradeoffs that reflect value judgments about which criteria are important to satisfy and which are not.

Decisions about problem identification and framing are particularly important in the context of algorithms used in the legal and penal systems, and I will discuss them in detail in section 5.

3.b Data decisions and model competencies

Data decisions are crucially important in the development of ML systems due to the fact that the ML algorithms, again, are data driven; characteristics of the model—e.g., the factors according to which it generates outputs, the relative weights of these factors, and the ways in which these factors are combined—are determined by the data on which the algorithm is trained, in addition to the algorithm chosen (to be discussed in sections 3.c–3.d). Because of this, decisions about data will impact model performance, including respects in which the model performs well and respects in which it does not. Furthermore, once an ML system is trained to a given level, it is typically evaluated according to a standard (or *benchmark*); the quality of the assessment, and the respects in which researchers will be able to assess the system, will depend upon decisions about the data set that is used for benchmarking. Decisions about data—including decisions about data inputs, data quantity, data quality, and representativeness—are some of the most significantly value-laden decisions in the ML design process.

One example that illustrates the value-laden character of training-data decisions come from natural-language processing. Bolukbasi et al. (2016) examined the word-embedding algorithm word2vec, which represents text data as vectors and is used to create word associations, for how it might reflect gender biases. They trained the software on Google News articles and found that it reproduced biased gender stereotypes in its associations; for example, it completed the analogy, "man is to computer scientist as woman is to x" with "x = homemaker." In this case, the software was biased because the data on which it was trained were biased.

A second example comes from facial recognition software. In their *Gender Shades* project, Joy Buolamwini and Timnit Gebru (2018) evaluated the performance of three commercial gender classification algorithms (IBM, Microsoft, and Face++) and found in all three significant disparities in accuracy across gender and skin color.[6] More specifically, they found that all perform better on male faces than female faces (error rate differences ranging from 8.1% to 20.6%); all perform better on lighter-skinned faces than darker-skinned faces (error rate differences from 11.8% to 19.2%), and all perform worse on darker-skinned female faces (error rates from 20.8% to 34.7%) (Buolamwini and Gebru 2018, 8). The IBM classifier had an error rate of 34.7% for dark-skinned females and 0.3% for light-skinned males (Buolamwini and Gebru 2018, 9). These are significant violations of the equalized odds criterion of fairness, to be discussed in section 3.d.

[5] For a discussion of epistemic risks in problem identification/framing in another context, see Biddle (2018).

[6] In addition to Buolamwini and Gebru (2018), see also gendershades.org for more information on the project.

https://doi.org/10.1017/can.2020.27 Published online by Cambridge University Press

This case illustrates the importance of data decisions in both training and benchmarking. Biases in the selection of training data can account to a significant degree for the disparities in accuracy; systems that are trained on data sets that disproportionately represent faces of white males will perform better on white males than on other groups. Additionally, these disparities in performance cannot be uncovered without data sets that are sufficiently diverse. Data sets that were available prior to Buolamwini and Gebru's *Gender Shades* project—the IJB-A and Adience data sets—disproportionately represented white faces and were not sufficiently diverse to be able to adequately evaluate the performance disparities in gender classification algorithms; to determine how well (or how poorly) an algorithm does on black and brown faces, one needs a data set that includes black and brown faces. To assess the software, Buolamwini and Gebru needed to construct a new, more representative, data set, the Pilot Parliaments Benchmark (PPB) (5).

One might respond to this discussion by arguing that these data decisions are not necessarily value laden; biased training data will lead to biased algorithms, and the correct response to this is to ensure that training data are unbiased. It is true that the facial data sets were initially very biased in the sense that they deviated significantly from any reasonable choice of baseline populations (the data were disproportionately white), so Buolamwini and Gebru's work made the data less biased, in this sense. But it does not follow that the data became value neutral, or even less value laden. There are innumerable ways in which data sets could be constructed, and designers have no choice but to make contingent decisions about how to do this—for example, decisions about how data should be structured, how much diversity should be reflected in the data, and what sort of diversity is important. These decisions, moreover, reflect values, including values embedded in the practical contexts in which the system will be used.

For example, in constructing a database of faces, designers must decide whether it is appropriate to include unstructured data (e.g., unlabeled images pulled from websites) or, if the data is structured, how it should be structured (e.g., whether it should include information about gender and race and, if so, how this data is to be represented). Should gender be represented in binary terms? Buolamwini and Gebru do so for practical purposes, but they acknowledge that this decision "does not adequately capture the complexities of gender or address transgender identities" (6). They do not include race or ethnicity as categories because "race and ethnic labels are unstable"; phenotypic features vary within racial and ethnic categories, and these categories vary across time and geographical location (4). Instead, they represent skin color according to a dermatologist-approved Fitzpatrick Skin Type classification, which divides skin color into six skin types (I–IV, with I being the lightest). Despite their decision to include skin color rather than race as a category, they acknowledge that it is perfectly acceptable to use race as a category in other contexts; they are "suitable for assessing potential algorithmic discrimination in some forms of data (e.g., those used to predict criminal recidivism rates") (4). There is no value neutral way to construct a data set; judgments must be made that reflect goals, values, or interests.

The argument that there is no value neutral way to construct a data set is similar in many respects to arguments made by philosophers of science that there is no value neutral way to model phenomena. In constructing a model that is used to represent target phenomena (e.g., a subway map), modelers must make contingent decisions about which features of the target phenomena they wish to represent and which they are willing to ignore (e.g., Biddle and Winsberg 2010; Giere 1988; Intemann 2015; Kitcher 2001; Longino 2002; Potochnik 2012). These decisions depend on values, interests, and pragmatic considerations that relate to the purposes and target audience of the representation. In constructing a data set, one must make similar contingent decisions about which features of the data are important for some purposes and which are not. Again, because ML systems are data driven, the values that are reflected in data decisions impact the overall performance of the system.

3.c Algorithm design: accuracy and explainability

As noted in section 2, there are many different types of ML algorithms, and there are different ways in which these algorithms learn. In designing an ML system, researchers must make contingent

https://doi.org/10.1017/can.2020.27 Published online by Cambridge University Press

decisions about which algorithm and learning method to adopt and, in many cases, these decisions involve tradeoffs that reflect values. One frequently discussed tradeoff in these contexts is that between accuracy and explainability.

Deep-learning methods have, in many cases, produced astonishing improvements in prediction speed and accuracy. To take just one example, deep learning algorithms are now able to diagnose many cancers with comparable accuracy (or better) when compared to professional pathologists (e.g., Ehteshami Bejnordi et al. 2017; Haenssle et al. 2018). At the same time, these gains arguably come at a cost—namely, the ability to explain why the algorithm arrived at the output that it did. In the case of algorithms that involve multiple layers and hundreds of millions of parameters, it can be difficult, if not impossible, to interpret which features of the data or combinations of those features relate to salient aspects of a target situation. This raises a concern that algorithmic choice involves a tradeoff between accuracy and explainability. Moreover, if designers indeed face this tradeoff, then they must confront serious ethical concerns regarding the use of algorithms that have significant impacts upon human beings—such as algorithms that predict risk of recidivism or defaulting on a loan. If an individual is given a longer prison sentence because of a decision of an algorithm, it is at least plausible to think that there is a moral obligation to explain to that individual why the algorithm produced the result that it did. This prima facie moral obligation is at the heart of the "right to explanation" contained in the European Union's General Data Protection Regulation (GDPR) (e.g., Kaminski 2019).

Concerns about the ability to interpret or explain are common in ML research communities. For example, in outlining its Explainable Artificial Intelligence (XAI) program, the United States Defense Advanced Research Projects Agency (DARPA) writes:

> Dramatic success in machine learning has led to a torrent of Artificial Intelligence (AI) applications. Continued advances promise to produce autonomous systems that will perceive, learn, decide, and act on their own. However, the effectiveness of these systems is limited by the machine's current inability to explain their decisions and actions to human users" (Turek 2018).

The aim of the XAI program is to "produce more explainable models, while maintaining a high level of learning performance (prediction accuracy)" and to "enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners" (Turek 2018). Concerns about tradeoffs between accuracy and explainability are not restricted to DARPA but are common in research communities (e.g., Royal Society 2017). Whether the aim of achieving AI that is both accurate and explainable will be reached is yet to be determined.

Questions remain about algorithmic explainability, accuracy, and the connection between them. These include not only technical questions about how research will evolve but also philosophical questions about what it means for a system to lack explainability (e.g., Sullivan 2019). What is clear is that, in many current ML systems, accuracy comes at the cost of explainability.

3.d Algorithm design: conceptions of fairness

While there is a certain degree of controversy about whether some ML algorithms involve tradeoffs between accuracy and explainability, there is little controversy about the following: in many cases, designers *must* make tradeoffs between rates of accuracy and error *among different groups* (e.g., Chouldechova 2017; Verma and Rubin 2018). If we consider an algorithm that produces outputs that affect people in some way, we might intuitively think of the algorithm as *fair* for different groups if it works equally well, or is equally accurate, for all groups. An important lesson of recent work in ML is that there are multiple different conceptions of fairness and that, in many cases, designers necessarily make tradeoffs between them.

https://doi.org/10.1017/can.2020.27 Published online by Cambridge University Press

Consider, for example, an algorithm that predicts whether an individual will pay back a loan.[7] For this case, suppose we have access to a data set that contains information about loan applicants, such as credit history, gender, marital status, employment status, etc., and suppose that the data is labeled such that it indicates whether the applicants who received loans actually paid them back. We then use this data to train an algorithm to predict whether individuals will pay back loans, and we wish to ensure that the algorithm is fair with respect to gender. Let S be the predicted probability that an individual will pay back a loan. Let d be the predicted category for the individual, such that the individual is predicted to pay back the loan ($d = 1$) if S is above a particular threshold and is predicted not to pay back the loan, ($d = 0$) otherwise. Let G be the gender of the individual, which we define in this case in binary terms (m = male, f = female). (In discussions of fairness in examples like this, gender is typically defined in binary terms, though this decision is problematic in many respects, as discussed in section 3.b.) Finally, let Y be the actual classification result (i.e., whether an individual actually pays back the loan).

There are a number of plausible but distinct conceptions of fairness with respect to gender that we might wish the algorithm to satisfy. Verma and Rubin (2018) identify and discuss twenty distinct conceptions. In the interest of space, I will restrict my attention to two: *predictive parity* and *equalized odds.* (These two criteria receive significant attention in the context of criminal sentencing). The criterion of predictive parity can be represented as follows:

$$P[Y=1|d=1,\ G=m]=P[Y=1|d=1,\ G=f]$$

With regard to the credit scoring example, it states that the probability that an individual actually pays back a loan conditional upon their being predicted to pay it back is the same for both genders. Roughly, this criterion attempts to capture the idea that for an algorithm to be fair, it should generate true predictions at the same rate for both genders. Note that satisfaction of predictive parity also implies false prediction rates will be the same for both genders:

$$P[Y=0|d=1,\ G=m]=P[Y=0|d=1,\ G=f]$$

The criteria of equalized odds, on the other hand, attempts to capture the idea that for an algorithm to be fair, it should not be more likely to generate false predictions for one group than another. We can represent the criteria of *false-positive error-rate balance* and *false-negative error-rate balance* respectively as follows:

$$P[d=1|Y=0,\ G=m]=P[d=1|Y=0,\ G=f]$$

$$P[d=0|Y=1,\ G=m]=P[d=0,\ Y=1|G=f]$$

With regard to the credit scoring example, false-positive error-rate balance states that the probability that an individual is predicted to pay back a loan, conditional upon that individual actually failing to pay it back, is the same for both genders. False-negative error-rate balance states that the probability that an individual is predicted not to pay back a loan conditional upon that individual actually paying it back is the same for both genders. Finally, the criterion of *equalized odds* is satisfied when both false-positive error-rate balance and false-negative error-rate balance are satisfied (Verma and Rubin 2018, 4).

What is crucial for the current discussion of epistemic risk and value judgments is that many of the distinct conceptions of fairness trade off of one another. For example, under certain empirical conditions, it is mathematically impossible to satisfy both predictive parity and equalized odds

[7] This example is adapted from Verma and Rubin (2018).

https://doi.org/10.1017/can.2020.27 Published online by Cambridge University Press

(Chouldechova 2017). In the case of the credit scoring example, if males and females have different base rate probabilities of being in the actual positive class—that is, if $P(Y=1|G=m) \neq P(Y=1|G=f)$ —then it is impossible to satisfy both predictive parity and equalized odds (Verma and Rubin 2018, 4). In these situations, designers face tradeoffs between the two types of fairness; predictive parity can only be achieved at the cost of equalized odds and vice versa. The choice of how the algorithm should be tuned—to achieve either predictive parity or equalized odds—reflects value judgments about the types of mistakes, or the types of unfairness, that one is more or less willing to tolerate. The fact that ML designers must make tradeoffs between different types of fairness is particularly important in the context of criminal sentencing. I will discuss this issue in more detail in section 4.e.

In section 3.b, we saw that decisions about data involve unavoidable tradeoffs. In this section, we see that even if designers were not necessarily confronted with tradeoffs in data decisions, they would under some conditions still have to make value-laden decisions regarding the types of fairness that they are more or less willing to violate.

3.e Algorithm design: outputs

Designers must make contingent decisions about what ML systems should output and the conditions under which particular outputs should be generated. Consider again the algorithm discussed in section 3.d, which classified individuals into those predicted to repay a loan ($d=1$) and those predicted not to repay a loan ($d=0$). This output was generated by calculating the predicted probability S (relative to a population) of repaying a loan and choosing a threshold such that if an individual has a predicted probability above that threshold, the system predicts a good credit score. In this case, designers made contingent decisions that the possible system outputs would be binary ($d=0$ or $d=1$) and that the probability threshold would be what it is.

Other options were available but not chosen. For example, designers could have decided that the algorithm output would simply be a probability relative to a population. That is, the output might have been that a certain percentage of people in a population who have traits similar to the individual in question pay back (or default) on their loans. In this case, the system would call attention to the fact that it is not making a prediction about an individual but rather relating facts about a population of people who share traits with the individual in question.

These decisions about algorithmic outputs reflect value judgments at multiple points. In the case of the choice of a threshold, the decision reflects value judgments concerning the relative costs of false positives versus false negatives (or value judgments about how to manage inductive risk). The costs of these mistakes will, in many cases, be borne by different stakeholders, and so choosing which type of mistake one is more or less willing to tolerate amounts to choosing which stakeholders one is more or less interested in prioritizing.

Taking a step back, the decision to generate a binary output in the first place—as opposed to generating an output about a probability relative to a population—reflects value judgments about who is best equipped, or most appropriately placed, to make particular inductive inferences. In creating an algorithm that generates a binary output, the designers are deciding that they—rather than the users of the algorithm—will make the inference about how an individual should be classified. The users do not need to draw an inference from data; they are simply given an answer and, moreover, they do not know how this answer is generated unless they investigate the internal workings of the system. In this case, the designers are making an evaluative judgment about users—either that they are not well equipped to make inferences from data, or that they do not wish to, or both—as well as an evaluative judgment about themselves, namely, that they are well equipped to make these inferences (and that they are appropriately placed to make the value judgments that these inferences reflect). On the other hand, the decision to refrain from generating a binary output and, instead, to output a probability relative to a population, involves a different set of evaluative judgments—either that the designers are not in an appropriate position to make the value-laden decision to set a particular threshold or that the users are better placed to make such determinations.

https://doi.org/10.1017/can.2020.27 Published online by Cambridge University Press

3.f Algorithm deployment: transparency and opacity

The previously discussed decisions regarding algorithmic technique, fairness criteria, and output selection are all decisions of algorithm design. The issue of transparency has received much attention in discussions of AI and ML and, while it is implicated in design decisions, I will discuss it primarily in the context of deployment, especially given concerns over the secrecy of algorithms such as COMPAS.[8]

What transparency means is contested, and an exhaustive treatment of the issue is beyond the scope of this paper.[9] As a starting point, consider the following "principle" of transparency for autonomous/intelligence systems (A/IS): "The basis of a particular A/IS decision should always be discoverable" (IEEE 2019, 11). This is one of the General Principles identified by the Institute for Electrical and Electrons Engineers (IEEE) in their ethics document, *Ethically Aligned Design.*

An examination of this principle reveals that transparency is a concept that is complex and contestable. What does it mean for a "basis" to be "discoverable"? Must the system be explainable? In the case of a predictive algorithm, must designers be able to identify the predictor variables that the system uses to generate predictions (as opposed to the attributes of the training data)? If this is the case, then transparency requires the attainment of particular technical benchmarks. Or perhaps transparency requires only that information about training data and performance metrics be made available?

Additionally, to satisfy this principle of transparency, for whom must a basis be discoverable? For the designers of the system (or other technical experts)? Or also to some users of the system? Or to everyone? And what is required in order for a system to be discoverable by a given stakeholder? Must that stakeholder actually understand the basis? What level of understanding is required, and what is the best way to communicate information so that the desired level of understanding can be reached? In some cases, overloading someone with irrelevant information can actually impede understanding; how much of what information, presented in which way, is most effective at eliciting understanding? Furthermore, how do legal considerations, such as intellectual property protections, intersect with questions about for whom a basis should be discoverable?

From these questions, we can glean (at least) three important lessons about transparency in ML. First, transparency is a stakeholder dependent concept; what is required in order to be transparent will, in general, differ from stakeholder to stakeholder. Second, questions about transparency are, at bottom, questions about communication that intersect with questions about technical design, law, and ethics (among others). What is appropriate and/or obligatory to communicate, and for what purposes? With whom is it appropriate and/or obligatory to communicate? For any given stakeholder, what is the most effective way to communicate? Third, judgments about transparency involve tradeoffs that reflect values.

To illustrate this third point, consider that to demand that one party communicate some information to another is to place a number of burdens on that party. It might place upon them the burden to acquire the information, in case they do not already possess it (e.g., information about how to explain the decision that a system makes). It places upon them burdens in terms of resources; it takes time, energy, and in some cases, money to communicate, and there are opportunity costs involved in this. It might also expose parties to significant financial risks in that communication of some information might involve the risk that trade secrets will be lost. Of course, it is in many cases appropriate and even obligatory that parties shoulder these burdens from an ethical and/or legal perspective. But there are tradeoffs involved, and these tradeoffs reflect values. And these are just some of the tradeoffs required in decisions about communicating with one other party. In many cases, it will be important to communicate with multiple stakeholders. Given that resources are

[8] For a discussion of the limitations of "algorithmic transparency," or transparency about the technical workings of algorithms, see Desai and Kroll (2018).

[9] See Elliott (2020) for a helpful discussion of complexities of the concept of transparency.

https://doi.org/10.1017/can.2020.27 Published online by Cambridge University Press

limited, communicating extensively with one stakeholder will take away resources that could be used to communicate with others.

Because judgments about transparency involve tradeoffs that reflect values, we should be cautious about claims to the effect that designers should be "as transparent as possible," or even that there is some "optimal" level of transparency (*cf.* Elliott and Resnik 2014; IEEE 2019). Additionally, because judgments about transparency reflect values, we should be cautious about appealing to transparency as a solution to the problem of the value-ladeness of science. Appeals to transparency cannot resolve questions about values in science because judgments about transparency (how much is required, in which respects, in which ways, for whom, and under which circumstances) are themselves significantly value laden.

4. Value judgments and recidivism prediction algorithms

Prediction algorithms, including ML algorithms, are increasingly being used in penal systems to influence a wide range of decisions, including decisions about whether an accused individual should be released or incarcerated pretrial, whether a convicted offender should be incarcerated or put on probation, the length of a prison sentence, and the types and strengths of interventions made while in the system. There is considerable variance from country to country and state to state in the types of algorithms used and the ways in which they are used. In the past, most algorithms were non-ML logistic regression models, but a variety of ML systems are increasingly being developed and used. Some algorithms, such as COMPAS, are developed by private, for-profit corporations, while others, including the Canadian Violence Risk Appraisal Guide (VRAG), are developed by public–private entities (e.g., Eaglin 2017, 69–71; Kehl, Guo, and Kessler 2017, 15–17). The different ownership models tend to correspond to differing levels of transparency; some are relatively opaque and covered by trade secrets (e.g., COMPAS), while others are more open. Different algorithms perform different functions; some, for example, only predict risk of recidivism, while others also identify ways to target interventions.

The increasing use of prediction algorithms in penal systems serves as an important case study on how value-laden decisions have significant—and in some cases, disparate—impacts on human lives. More specifically, I will focus on decisions that involve tradeoffs at the following stages: problem identification and framing, collection and use of data for training and testing, algorithm design, and decisions about transparency and opacity in the deployment of systems.[10]

4.a Problem identification and framing: Why prediction? What recidivism?

As noted in the introduction to this paper, the development and deployment of ML-based tools in legal and penal systems have been motivated in part by concerns about bias and racial discrimination; proponents of these systems argue that ML tools have the potential to reduce bias, discrimination, and incarceration rates. It is important to note that the prioritization of the development of recidivism-prediction tools involves a particular framing of the problem of bias and discrimination—namely, that predicting more accurately the behavior of those who have been arrested is an important and effective strategy for reducing bias and discrimination. As I will discuss in section 5, there are other ways of framing the problem. For example, one might frame the problem in terms of identifying bias in policing and judicial decisions rather than predicting the behavior of those who have been arrested. In any case, the decision to develop recidivism-prediction tools in the first place involves a particular problem-framing that involves significant tradeoffs.

Once one has decided to develop a recidivism-prediction tool, one must then operationalize the concept of recidivism. There are a variety of ways that this might be done, and different designers

[10] The discussion in this section is indebted to, and builds on, that of Eaglin (2017).

https://doi.org/10.1017/can.2020.27 Published online by Cambridge University Press

have done it differently. Possible operationalizations include: re-arrest and conviction; re-arrest and being formally charged with a crime; re-arrest whether or not one is charged; and violation of probation (e.g., failing to meet with a parole officer, failing a drug test, failing to pay a debt to the court system, etc.). Additionally, operationalizing recidivism should specify what types of crime count as triggering events, and designers must decide how far into the future the tool should predict. Equivant, for example, uses any arrest as a triggering event (misdemeanor or felony), and its COMPAS algorithm assesses risk over a period of two years after the initial intake assessment (Northpointe 2012). VRAG, alternatively, uses "any new criminal charge for a violent offense" and it assesses risk over a period of five years (Eaglin 2017, 76–77; Harris et al. 2015, 122).

One of the tradeoffs involved in operationalizing the concept of recidivism is between data quantity and data quality. The decision to define recidivism in terms of re-arrest whether or not an individual is formally charged has the benefit of allowing for relatively higher quantities of data to be collected more quickly and conveniently. In this case, for example, developers obtain data about "recidivists" without needing to wait for potentially lengthy court proceedings to conclude. This is beneficial for developers, especially those constructing ML algorithms that require large amounts of training data. At the same time, defining recidivism in this way risks categorizing individuals who did not recommit crimes as recidivists. Given that individuals from historically disadvantaged groups are arrested at disproportionately high levels, the decision to define recidivism in terms of re-arrest will likely result in the systematic overestimation of risk in these populations. The decision to define recidivism in terms of re-arrest and conviction could result in higher data quality at the expense of data quantity.

More broadly, defining recidivism involves tradeoffs between who is wrongly counted and not counted. One consequence of all of the operationalizations stated above is that they leave out individuals who commit crimes but are never arrested. Individuals who commit the types of crimes that are less stringently enforced (e.g., white-collar crimes) will be more likely to fall outside the scope of the concept. On the other hand, some definitions of recidivism have the implication that individuals who never commit a crime at all can count as recidivists. In the case of COMPAS, individuals who are arrested in some states (and therefore entered into those penal systems) must undergo an initial intake assessment; if they are arrested again within two years, they will count as recidivists, regardless of whether or not they ever actually committed a crime. Again, individuals who are more likely to fall into this group are those who are historically disadvantaged and more heavily policed.

4.b Data decisions: base-line populations and feature choice

Once the concept of recidivism has been defined, researchers can collect data that fits the concept definition. This creates a baseline population that "creates the world within which statistical models generate predictions" (Eaglin 2017, 73). That is, it provides the baseline information that can be used to identify factors that correlate to reoffence. Given that there are many different possible populations that could be used as a baseline, designers must make judgment calls about which data to collect. For example, in order to predict the risk of reoffending, researchers must operationalize the notion of having offended in the first place, and different researchers do this in different ways. Some collect information about individuals who have been charged with a crime, others on individuals who have been charged and convicted of a crime. Designers must also make decisions about which crimes to include (all crimes? only violent crimes? only felony crimes?) as well as decisions about which geographical areas to use as sites for data collection.

Designers must also make decisions that will impact the composition of the baseline population in terms of race, gender, and socioeconomic background. In the case of penal systems that are populated disproportionately by historically marginalized groups, if researchers collect information about individuals who are already in these systems (as all of the developers of these tools do), then the baseline data will reflect the discriminatory practices that have led to these injustices. In this

https://doi.org/10.1017/can.2020.27 Published online by Cambridge University Press

case, the baseline population data will itself be racially and ethnically biased. Should researchers curate the data in an attempt to reduce this bias? If so, how? Whatever decisions are made with regard to the composition of baseline population data will be difficult and value laden.

The next step of the tool-development process involves identifying and selecting factors to measure that potentially correlate with recidivism. Factors commonly identified include criminal history, age at first arrest, current age, gender, socioeconomic status, current employment status, treatment for substance abuse, criminal companions/associations, antisocial behavior, and family criminality, among others (e.g., Eaglin 2017, 78–80; Kehl, Guo, and Kessler 2017, 7–9). As in the previous steps, decisions about which factors to measure can be made in different ways, and different designers make them in different ways. In the case of non-ML systems, designers might identify factors from the criminological literature and then collect data and build tools on this basis; alternatively, they might build tools on the basis of factors that are present in readily available data sets. In the case of ML systems, factors will be identified, weighted, and combined in ways determined by the training data. Whether or not the system will be sufficiently explainable to allow researchers or users to understand which factors or combinations of factors are relevant (i.e., which are the predictor variables), and to what extent, will depend on the choice of algorithmic technique (discussed in section 3.c).

Regardless of what type of algorithm is chosen, the decision to include some factors can involve important tradeoffs that reflect values. Criminological research suggests that factors such as socioeconomic status and family criminality are predictors of recidivism (e.g., Eaglin 2017, 79). Because of this, considerations of predictive accuracy would suggest that they be included in recidivism risk assessments. Most algorithms, including COMPAS, include them. At the same time, it is difficult to square the inclusion of these factors with the ideal of equality under the law. Including socioeconomic status in risk assessment tools, for example, implies that individuals who are economically disadvantaged will, all else being equal, be assessed a higher risk of recidivism than individuals who are wealthy. To the extent that higher risks of recidivism will lead to longer sentences, individuals who are poor will be given longer sentences than those who are rich. This, in turn, tends to create a feedback loop in which high-risk predictions become self-fulfilling prophecies that can have long-term, even multigenerational, effects. Because of this, the inclusion of these factors would seem to violate the political ideal of equality under the law. In the US legal system, this is potentially a violation of the Equal Protection clause of the Fourteenth Amendment to the Constitution, though it has not (yet) been challenged on these grounds (e.g., Kehl, Guo, and Kessler 2017; Starr 2015). In any case, however one judges this legal question, if it is granted that socioeconomic status (or some combination of factors related to socioeconomic status) is indeed a predictor of recidivism, then the judgment of whether it should be included in risk assessment tools involves an inescapable tradeoff that reflects values.

4.c Output selection

In the next step, designers must decide what the system should output. In the case of tools that translate a quantitative risk score into a qualitative risk output, designers must decide cutoffs for the different categories of risk. As in the previous steps, different designers do this in different ways. For example, the COMPAS algorithm generates a probability of re-arrest relative to a population, and it translates this probability into a score of 1 to 10; it then outputs a qualitative assessment of "high," "medium," or "low" risk, where scores of 1 to 4 are labeled "low," 5 to 7 are "medium," and 8 to 10 are "high." As discussed in section 3.e, these decisions reflect value judgments at multiple points. These include interpretations about what it means for someone to be "high," "medium," or "low" risk, as well as judgments about relative tolerances of different types of errors that are reflected by choices of cutoffs for these risk categories. Additionally, they include judgments about who is most appropriately placed to make these decisions—e.g., designers of technical systems or judges who are elected or appointed by elected officials.

https://doi.org/10.1017/can.2020.27 Published online by Cambridge University Press

It is also important to note that, especially in the case of recidivism-prediction algorithms, there is a danger that judges might see an output of "high" or "low" risk and fail to appreciate the myriad ways in which these outputs are structured by values that operate within the system. Judges who are unfamiliar with the process of tool design (which is probably most of them) might not appreciate the ways in which decisions about how to operationalize the concept of recidivism (for example) could impact the likelihood of different types of mistakes for different populations and, instead, simply assume that the outputs are generated by neutral processes.

4.d Choice of fairness criterion

Since 2016, a heated debate has arisen about racial bias in recidivism-prediction algorithms and how these might exacerbate existing inequalities. The debate can be traced back to a story in *ProPublica* entitled "Machine Bias," which charged that Northpointe's (now Equivant's) COMPAS algorithm was biased against African Americans (Angwin et al. 2016). COMPAS does not input data about race, but it does input myriad other information that correlates with race and is, according to some, a proxy for race (e.g., Harcourt 2015). For the story, *ProPublica* obtained COMPAS scores from the Broward County Sheriff's Office (in Florida) over a two-year period, 2013–2014. Broward County uses COMPAS scores to determine whether or not to release a defendant awaiting trial. Because of this, *ProPublica* focused its analysis on those who received a COMPAS score in the pretrial phase (Larson et al. 2016). COMPAS, again, assesses the risk of recidivism (any arrest) over a two-year period after the initial intake assessment; using the Broward County data, *ProPublica* compared the predictions made by COMPAS with actual outcomes —i.e., whether or not the individuals were actually re-arrested. They distinguished those who were classified as higher risk to recidivate (which they defined as receiving a COMPAS score of 5 or above) with those who were classified as low risk, and they examined the data to determine if there were systematic differences in accuracy and error rates between black and white populations.

They found that, while COMPAS correctly predicted recidivism at approximately equal rates for black and white populations (63 percent and 59 percent, respectively), there were significant disparities in the types of errors found among the two groups. Black people were significantly more likely to be classified mistakenly harshly, whereas white people were significantly more likely to be classified mistakenly leniently. More specifically, 45 percent of black defendants who did not recidivate were classified as medium or high risk compared to 23 percent of white defendants. On the other hand, 48 percent of white defendants who did recidivate were classified as low risk compared to 28 percent of black defendants (Angwin et al. 2016; Larson et al. 2016). This is a significant departure from the equalized odds criterion of fairness—and it is a departure that recreates and reaffirms existing racial inequalities.

Following the publication of the *ProPublica* story, Northpointe responded and denied the charges that COMPAS is racially biased (Dieterich, Mendoza, and Brennan 2016). Their response is multifaceted, but perhaps their most significant argument is that COMPAS is not racially biased because it meets the fairness criterion of predictive parity, which states (again) that the rate at which a tool generates true predictions should be the same for different groups.

Research on algorithmic fairness has identified what is at the heart of this debate: under some empirical conditions, it is mathematically impossible to achieve both predictive parity and equalized odds; under these conditions, one is achieved at the expense of the other (Chouldechova 2017). More specifically, if black and white populations have different base rates of recidivism, then it is impossible for any tool to meet both fairness criteria for these populations. In many countries (including the United States), this empirical condition is met—black populations do have higher base rates of recidivism; in these areas, designers face unavoidable tradeoffs between the two fairness criteria.

For the case of COMPAS, we can represent the predictive parity criterion as follows:

$$P[Y=1|s>S_{HR}, R=w]=P[Y=1|s>S_{HR}, R=b]$$

https://doi.org/10.1017/can.2020.27 Published online by Cambridge University Press

This states that the probability that an individual actually recidivates conditional upon their receiving a higher risk assessment is the same for both black and white populations. For the equalized odds criteria, we can represent the criteria of false-positive error-rate balance and false negative error rate balance, respectively, as follows:

$$P[s > S_{HR} | Y = 0, R = w] = P[s > S_{HR} | Y = 0, R = b]$$

$$P[s \leq S_{HR} | Y = 1, R = w] = P[s \leq S_{HR}, Y = 1 | R = b]$$

False-positive error-rate balance states that the probability that an individual is classified as higher risk, conditional upon that individual actually not recidivating, is the same for both black and white populations. False-negative error-rate balance states that the probability that an individual is classified as lower risk conditional upon that individual actually recidivating is the same for both populations. The criterion of equalized odds, again, is satisfied when both false-positive error-rate balance and false-negative error-rate balance are satisfied.

If there are higher base rates of recidivism among black populations, then there will be proportionately higher percentages of black people who are labeled as higher risk of recidivating. Because the accuracy of the algorithm is roughly the same for both populations (it meets predictive parity), and because black people are proportionately more likely to be labeled as higher risk, they will also be more likely to be *mislabeled* as higher risk. A similar argument can be given to show that under these conditions white people will be more likely to be mislabeled as low risk (Chouldechova 2017; Corbett-Davies et al. 2016).

Depending on how COMPAS (and other recidivism-prediction algorithms) is used, it can have disparate impacts on different populations. Black populations are already incarcerated at disproportionately high levels in the United States, and they are subjected to racism throughout the policing and penal systems, which exacerbates existing inequalities and further perpetuates systems of injustice. The issue of algorithmic fairness is complex, especially given the fact that different conceptions of fairness trade off of one another. But judgments about the design of these systems have real and disparate impacts on people, and the process of developing ML systems needs to take account of this.

4.e *Transparency*

Issues of transparency, opacity, and communication are especially important in debates over ML software used in penal systems. More specifically, questions have been raised about the due process implications of using algorithms that are covered by trade secrets and, hence, hidden from the courts (and the defendants). Additionally, questions concerning explainability have been raised about the ability of anyone to know precisely which predictor variables are used by a given ML system. Both of these concerns are prominent in the *Wisconsin v. Loomis* decision, to which I will now turn.

5. *Wisconsin v. Loomis* and recommendations for the use of ML in penal systems

Thus far, I have examined the role of value judgments in the design of ML systems, including algorithms used in penal systems to predict recidivism; there are numerous places in the upstream development and design phases where value-laden decisions are made that involve epistemic risk and that have significant—and in some cases, disparate—downstream impacts on human lives. What are the implications of this argument for the assessment of these instruments? Should they be used at all? If so, under what conditions? To probe these questions, let us return to the *Wisconsin* *v. Loomis* decision. Loomis appealed the sentencing decision by arguing that the decision violated his due process rights for three reasons (State v. Loomis 2016, 757). First, he argued that it violated his right to be sentenced according to accurate information because the proprietary nature of the algorithm prohibited him from accessing its accuracy. Second, he argued that it violated his right to an individualized sentence because it assessed risk of recidivism relative to a population. Third, he argued that it improperly used gendered assessments because the algorithm takes into account the gender of the offender. Importantly, Loomis did not argue that COMPAS's use of gender violates the Equal Protection Clause; he appealed on due-process grounds, not equal-protection grounds. The issue of race did not arise in the trial as Loomis is white.

The court decided against Loomis on all grounds. It argued that it did not violate his right to be sentenced according to accurate information because COMPAS uses as inputs information provided either by the defendant (in the form of a questionnaire) or from public records, and Loomis had the ability to examine this information (State v. Loomis 2016, 761–62). The court acknowledged that the proprietary nature of the algorithm made it impossible for Loomis to know how this information was used—how these factors were combined and/or weighted—but this, in the court's view, did not amount to a due-process violation. Regarding the second charge, the court argued that the use of COMPAS did not violate Loomis's right to an individualized sentence as long as the risk assessment was merely a factor in the decision, not the determinative factor (State v. Loomis 2016, 765). Regarding the third charge, the court argued that as long as the use of gender increases the accuracy of the algorithm, then it is not improper to employ it (State v. Loomis 2016, 767).

For these reasons, the court argued that the use of COMPAS to influence sentencing decisions does not necessarily violate a defendant's due-process rights. However, the court acknowledged that the use of COMPAS could, in some circumstances, be problematic, and it took some steps to regulate the way in which it is used. In particular, the court argued that, when a judge receives a presentence investigation report that includes a COMPAS score, the score should be accompanied by the following statement of "cautions" regarding the score's accuracy:

> the proprietary nature of COMPAS has been invoked to prevent disclosure of information relating to how factors are weighed or how risk scores are to be determined; (2) risk assessment compares defendants to a national sample, but no cross-validation study for a Wisconsin population has yet been completed; (3) some studies of COMPAS risk assessment scores have raised questions about whether they disproportionately classify minority offenders as having a higher risk of recidivism; and (4) risk assessment tools must be constantly monitored and re-normed for accuracy due to changing populations and sub-populations. (State v. Loomis 2016, 763–64)

An examination of the constitutionality of this decision is beyond the scope of this paper. In the remainder of this section, however, I will draw on the arguments of sections 3–4 of this essay in order to argue that from a broader ethics and policy perspective, the court's statement of "cautions" does not go far enough. In what follows, I will briefly discuss three recommendations, ranging from more modest to more radical. Each of these, in my view, represents an improvement over the status quo, with the first representing the most modest improvement and the third the most significant.

5.a Transparency, Opacity, and Pluralism

The first recommendation operates under the assumption that courts continue to allow the use of recidivism-prediction algorithms that are covered by trade secrets and that are opaque to the courts. In this case, judges should not consider the results of only one system; if they consider one, then they should also consider the outputs of multiple additional algorithms that are developed by different entities (e.g., public–private partnerships or nonprofit organizations) and that are designed in different ways such that at least some are accessible to inspection by a broad range of stakeholders and publics. As shown in sections 3–4, value-laden decisions about data, fairness criteria, output

https://doi.org/10.1017/can.2020.27 Published online by Cambridge University Press

selection, and others can impact algorithm performance; because of this, it is prudent to avoid reliance upon one single algorithm, and it is important for ethical reasons that at least some systems are accessible to users.

By way of comparison, consider the development and use of climate models. There is a plurality of different climate models in use, and policy makers and others who must decide how to act on the basis of climate change predictions are wise to consider the range of predictions made by these different models (e.g., Parker 2006). In some cases, different value-laden decisions in model construction can lead to differences in model performance (e.g., Biddle and Winsberg 2010; Intemann 2015). If a plurality of differently designed models all tend to make the same predictions, then decision makers have stronger grounds for acting confidently on the basis of these predictions. If there are significant differences in the predictions, then decision makers should be more cautious, as the differences in predictions might indicate a lack of understanding of the relevant climate systems. In any case, the climate science and policy communities have recognized that it is unwise to make consequential decisions on the basis of the outputs of single models, especially if there are significant uncertainties in those models, as is the case with recidivism-prediction algorithms.

This recommendation raises many additional questions, including: How many different algorithms should be used? What sorts of quality-control criteria should be employed to determine which algorithms are suitable for use? What types of pluralism are important—e.g., in which respects should the algorithms be different from one another? All of these are questions that require further attention if this recommendation is to be implemented. But given the significant role that value judgments play in the design of recidivism-prediction algorithms, the recommendation serves as a useful starting point.

5.b Values and public input

The second recommendation, which is preferable to the first, represents a greater departure from the status quo. It operates under the assumption that courts continue to allow the use of recidivism-prediction algorithms but that they no longer allow these algorithms to be legally opaque to users; in this case, courts should only use algorithms that are accessible to—and have input from—diverse stakeholders and publics. After all, courts operate within the public sector, and it is reasonable to require that systems that are used in the public sector—especially systems that are significantly value-laden—be open to scrutiny by, and receive input from, diverse publics.

For a system to be accessible to the public, developers should (at a minimum) communicate with the courts sufficiently to allow relevant parties to understand the algorithms to the extent possible or desired. Information to be communicated should include not just which data are input into the system, but also what type of algorithm and learning methods are employed and how the system weights and/or combines the input variables to reach its outputs. If the system is not explainable in this sense, then this fact should be communicated to the courts.

Beyond this minimum degree of communication, states that allow or require the use of recidivism-prediction algorithms should undertake efforts to foster engagement of relevant stakeholders and publics in the tool-design and implementation process to ensure that instrument design and use reflects the values of affected communities. For example, state agencies could create—or participate in the creation of—risk assessment tools, and seek input from affected publics on decisions such as how to operationalize the concept of recidivism and what the systems should output. Democratic deliberation about the purposes of punishment, and the ways in which these purposes intersect with prediction systems, should also be encouraged. In the United States, for example, there are multiple stated goals of punishment, including backward-looking considerations (giving offenders "what they deserve") and forward-looking considerations (promoting welfare by protecting by public); different tools facilitate these goals to different degrees. For example, instruments that use socioeconomic status to predict recidivism might facilitate the goal of protecting the public, but they do little to determine what an offender deserves. The question of

https://doi.org/10.1017/can.2020.27 Published online by Cambridge University Press

the purpose of punishment is a political one and, given that different algorithmic tools are more or less suited to different purposes, governments should elicit democratic input into decisions about tool use and development.[11]

This second recommendation is consistent with part of the first—namely the prohibition on judges considering only one algorithmic output. Under the second recommendation, courts could consider the results of a plurality of different algorithms as long as none of them are legally opaque to users. The second recommendation also has the significant benefit of alleviating concerns about due-process violations. In *Wisconsin v. Loomis*, the court decided that the use of COMPAS did not necessarily constitute a due-process violation; at the same time, there was clearly some level of discomfort surrounding the opacity of the algorithm as evidenced by the fact that the first "caution" highlighted by the court concerned the fact that some algorithms are subject to trade secrets.

5.c Predictive algorithms and inequality exacerbation

The third recommendation—which deviates the most from the status quo and which, to my mind, is the most important of the three—is that courts prohibit the use of recidivism-prediction algorithms that can be shown to disadvantage groups that are already unjustly disadvantaged. From the perspectives of both ethics and political philosophy, the maxim to refrain from disadvantaging groups that are already unjustly disadvantaged is relatively uncontroversial; for example, it is consistent with, and is in many cases weaker than, principles of justice found in theories such as utilitarianism, Rawlsian liberalism, and capabilities approaches. Moreover, it is at least plausible to think that this recommendation is grounded firmly in the ideal of equality under the law—which, again, is a potential basis for challenging the use of COMPAS that has not yet been pursued.[12]

Given current conditions in the United States, the application of the principle that one should refrain from disadvantaging groups that are already unjustly disadvantaged implies that the use of any recidivism-prediction algorithm that satisfies the criteria of predictive parity should be discontinued, at least temporarily. Any such algorithm disadvantages people of color who are already unjustly disadvantaged, especially African Americans, by tending to classify them mistakenly harshly (see section 4.d). Unless and until these conditions change—for example, by lowering base rates of recidivism of black populations to the point that they are equal to those of white populations—the use of these algorithms should be halted.

None of this is to say that ML algorithms should not be used in penal systems. Courts should explore the use of ML systems—just not those that exacerbate existing and unjust inequalities (and especially not those that can be shown to do this). One such system that might be explored is an instrument that could be applied to judges, alerting them to potential biases in their own decisions, such as tendencies to give minority offenders longer sentences than white offenders. This is an example of a system that would change the power dynamics of how algorithms function in court systems; rather than assessing and controlling those who have been arrested (and who are predominantly from disadvantaged backgrounds), this system would provide a check on judges' decision making.

Given this power shift, one might expect judges to resist the use of such a system—and in some countries, they have done so emphatically. In 2019, France banned the publication of judicial

[11] Eaglin (2017) discusses a variety of ways in which governmental entities could engage communities on these issues.

[12] Schroeder (2020) distinguishes between normative standards for values in science that are grounded in ethics versus political philosophy. The normative standard discussed in this section is grounded in political philosophy. In countries that guarantee equal protection under the law, courts should not employ technological systems that systematically disadvantage groups that are already unjustly disadvantaged. This is not to say that this standard could not be grounded in ethics—I believe that it could—but my concern in this section is on the use of ML systems in courts of law, and hence the normative standard that I emphasize is political in nature.

https://doi.org/10.1017/can.2020.27 Published online by Cambridge University Press

analytics and imposed a punishment of up to five years in prison for anyone who violates the ban.[13] According to the new law, Article 33 of the Justice Reform Act, “No personally identifiable data concerning judges or court clerks may be subject to any reuse with the purpose or result of evaluating, analyzing or predicting their actual or supposed professional practices” (quoted in Tashea 2019). The French government is not opposed to using data analytics in other areas. It is embracing the use of data analytics in policing its citizens—just not its judges (e.g., Perrot 2017).

The opposition to using ML systems to evaluate judges illustrates the fact that ML systems can be instruments of power with normative force. It is clear that ML systems are value laden; the important questions surrounding their development and use, again, are how they are value laden, whose interests they serve, whose values they reflect, and how they might by designed and implemented to effect positive change.

6. Conclusion

This paper has examined the role of value judgments in the design of ML systems in general and in recidivism-prediction algorithms in particular. There are numerous places in the upstream design of ML systems that require decisions that involve tradeoffs that reflect values; in many cases, these decisions have significant—and in some cases, disparate—downstream impacts on human lives. In the design of recidivism-prediction algorithms, the tradeoffs that have been made have tended to disproportionately harm groups that are already unjustly disadvantaged—but this need not be the case. ML systems can be designed and used in ways that have significant benefits, including for marginalized groups. But to achieve this, those who design, implement, monitor, and regulate these systems must be cognizant of the tradeoffs that are made and how they can have differential impacts on different stakeholders and publics.

Acknowledgments. Early versions of this paper were presented at the University of Alberta, Université Jean Moulin Lyon, the University of Cambridge, and the 2020 Association for Practical and Professional Ethics (APPE). I am grateful to all in attendance for their feedback, as well as to Ingo Brigandt, Daniel Schiff, Andrew Schroeder, and two anonymous reviewers for their comments and suggestions. I am also grateful to the Public Interest Technology University Network for support.

Justin B. Biddle is an associate professor in the School of Public Policy at the Georgia Institute of Technology.

References

Alexander, Michelle. 2012. *The New Jim Crow: Mass Incarceration in the Age of Colorblindness*. Rev. ed. New York: The New Press.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. “Machine Bias.” *ProPublica.* May 23. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Biddle, Justin B. 2013. “State of the Field: Transient Underdetermination and Values in Science.” *Studies in History and Philosophy of Science* 44: 124–33.

Biddle, Justin B. 2016. “Inductive Risk, Epistemic Risk, and Overdiagnosis of Disease.” *Perspectives on Science* 24 (2): 192–205.

Biddle, Justin B. 2018. “‘Antiscience Zealotry’? Values, Epistemic Risk, and the GMO Debate.” *Philosophy of Science* 85: 360–79.

Biddle, Justin B. 2020. “Epistemic Risks in Cancer Screening: Implications for Ethics and Policy.” *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 79: 101200.

Biddle, Justin B., and Rebecca Kukla. 2017. “The Geography of Epistemic Risk.” In *Exploring Inductive Risk: Case Studies of Values in Science*, edited by K. Elliott and T. Richards, 215–37. Oxford: Oxford University Press.

Biddle, Justin, and Eric Winsberg. 2010. “Value Judgements and the Estimation of Uncertainty in Climate Modeling.” In *New Waves in Philosophy of Science*, edited by P.D. Magnus and J. Busch, 172–97. Basingstoke, England: Palgrave MacMillan.

Bolukbasi, Tolga, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. “Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings.” In Advances in Neural Information Processing Systems *29*, edited by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, 4349–57.

[13] Thanks to John Walsh for bringing this law to my attention.

https://doi.org/10.1017/can.2020.27 Published online by Cambridge University Press

Brown, Matthew. 2013. "Values in Science beyond Underdetermination and Inductive Risk." *Philosophy of Science* 80 (5): 829–39.

Brown, Matthew. 2020. *Science and Moral Imagination: A New Ideal for Values in Science*. Pittsburgh: University of Pittsburgh Press.

Buolamwini, Joy, and Timnit Gebru. 2018. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." *Proceedings of Machine Learning Research* 81: 1–15.

Chouldechova, Alexandra. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5 (2). https://doi.org/10.1089/big.2016.0047.

Corbett-Davies, Sam, Emma Pierson, Avi Feller, and Sharad Goel. 2016. "A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased against Blacks. It's Actually Not That Clear." *Washington Post*, October 17. https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas.

Desai, Devin, and Joshua Kroll. 2018. "Trust but Verify: A Guide to Algorithms and the Law." *Harvard Journal of Law and Technology* 31 (1).

Dieterich, William, Christina Mendoza, and Tim Brennan. 2016. "COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity." Northpointe Inc. Research Department. http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf.

Douglas, Heather. 2000. "Inductive Risk and Values in Science." *Philosophy of Science* 67 (4): 559–79.

Douglas, Heather. 2009. *Science, Policy, and the Value-Free Ideal*. Pittsburgh, PA: University of Pittsburgh Press.

Douglas, Heather. 2017. "Why Inductive Risk Requires Values in Science." In *Current Controversies in Values and Science*, edited by Kevin Elliott and Daniel Steel, 81–93. New York: Routledge.

Eaglin, Jessica. 2017. "Constructing Recidivism Risk." *Emory Law Journal* 67: 59–122.

Ehteshami Bejnordi, Babak, Mitko Veta, Paul Johannes van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen A. W. M. van der Laak *et al.* 2017." Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women with Breast Cancer." *JAMA* 318 (22): 2199–210. https://doi.org/10.1001/jama.2017.14585.

Elliott, Kevin. 2011. *Is a Little Pollution Good for You? Incorporating Societal Values in Environmental Research*. New York: Oxford University Press.

Elliott, Kevin. 2020. "A Taxonomy of Transparency in Science." *Canadian Journal of Philosophy.*

Elliott, Kevin, and David Resnik. 2014. "Science, Policy, and the Transparency of Values." *Environmental Health Perspectives* 122 (7): 647–50.

Giere, Ronald. 1988. *Explaining Science: A Cognitive Approach*. Chicago: University of Chicago Press.

Gillespie, Tarleton. 2014. "The Relevance of Algorithms." In *Media Technologies*, edited by Tarleton Gillespie, Pablo Boczkowski, and Kirsten Foot, 167–94. Cambridge, MA: MIT Press.

Haenssle, H. A., C. Fink, R. Schneiderbauer, F. Toberer, T. Buhl, A. Blum, A. Kalloo, A. Ben Hadj Hassen *et al.* 2018. "Reader Study Level-I and Level-II Groups, Man against Machine: Diagnostic Performance of a Deep Learning Convolutional Neural Network for Dermoscopic Melanoma Recognition in Comparison to 58 Dermatologists." *Annals of Oncology* 29 (8): 1836–42. https://doi.org/10.1093/annonc/mdy166.

Harcourt, Bernard. 2015. "Risk as a Proxy for Race: The Dangers of Risk Assessment." *Federal Sentencing Reporter* 27 (4): 237–43.

Harris, Grant T., Marnie E. Rice, Vernon L. Quinsey, and Catherine A. Cormier. 2015. *Violent Offenders: Appraising and Managing Risk*, 3rd ed. Washington, DC: American Psychological Association.

Havstad, Joyce. 2020. "Archaic Hominin Genetics and Amplified Inductive Risk." *Canadian Journal of Philosophy.*

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. 2019. *Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems.* 1st ed. https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html.

Intemann, Kristen. 2015. "Distinguishing between Legitimate and Illegitimate Values in Climate Modeling." *European Journal for Philosophy of Science* 5 (2): 217–32.

Kaminski, Margot E. 2019. "The Right to Explanation, Explained." *Berkeley Technology Law Journal* 34 (1). http://doi.org/10.2139/ssrn.3196985.

Kehl, Danielle, Priscilla Guo, and Samuel Kessler. 2017. *Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing.* Responsive Communities Initiative, Berkman Klein Center for Internet and Society, Harvard Law School. http://nrs.harvard.edu/urn-3:HUL.InstRepos:33746041.

Kitcher, Philip. 2001. *Science, Truth, and Democracy*. New York: Oxford University Press.

Larson, Jeff, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. "How We Analyzed the COMPAS Recidivism Algorithm." *ProPublica*, May 23. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

Lashbrook, Angela. 2018. "AI-Driven Dermatology Could Leave Dark-Skinned Patients Behind." *The Atlantic*, August 16. https://www.theatlantic.com/health/archive/2018/08/machine-learning-dermatology-skin-color/567619.

LeCun, Y., Y. Bengio, and G. Hinton. 2015. "Deep Learning." *Nature*, 521 (7553): 436–44.

Longino, Helen. 1990. *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton, NJ: Princeton University Press.

Longino, Helen. 2002. *The Fate of Knowledge*. Princeton, NJ: Princeton University Press.

https://doi.org/10.1017/can.2020.27 Published online by Cambridge University Press

Morley, Jessica, Luciano Floridi, Libby Kinsey, Anat Elhalal. 2019. "From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices." *Science and Engineering Ethics*. https://doi.org/10.1007/s11948-019-00165-5.

Northpointe. 2012. *COMPAS Risk and Need Assessment System: Selected Questions Posed by Inquiring Agencies*. http://www.northpointeinc.com/files/downloads/FAQ_Document.pdf.

Okruhlik, Kathleen. 1994. "Gender and the Biological Sciences." *Biology and Society* 20: 21–42.

O'Neil, Cathy. 2016. *Weapons of Math Destruction*. New York: Crown.

Parker, W. 2006. "Understanding Pluralism in Climate Modeling." *Foundations of Science* 11: 349–68.

Perrot, Patrick. 2017. "What about AI in Criminal Intelligence? From Predictive Policing to AI Perspectives." *European Police Science and Research Bulletin* 16: 65–76.

Petrov, Christo. 2019. "Big Data Statistics 2019." *TechJury* (blog). https://techjury.net/stats-about/big-data-statistics.

Potochnik, Angela. 2012. "Feminist Implications of Model-Based Science." *Studies in History and Philosophy of Science* 43: 383–89.

The Royal Society. 2017. *Machine Learning: The Power and Promise of Computers That Learn by Example*. https://royalsociety.org/~/media/policy/projects/machine-learning/publications/machine-learning-report.pdf.

Rudner, Richard. 1953. "The Scientist Qua Scientist Makes Value Judgments." *Philosophy of Science* 20 (1): 1–6.

Schroeder, Andrew. 2020. "Values in Science: Ethical vs. Political Approaches." *Canadian Journal of Philosophy*.

Shueh, Jason. 2016. "White House Challenges Artificial Intelligence Experts to Reduce Incarceration Rates." *Government Technology*, June 7. https://www.govtech.com/computing/White-House-Challenges-Artificial-Intelligence-Experts-to-Reduce-Incarceration-Rates.html.

Singh, Jatinder, Ian Walden, Jon Crowcroft, and Jean Bacon. 2016. "Responsibility and Machine Learning: Part of a Process." SSRN, October 27. https://ssrn.com/abstract=2860048 or http://doi.org/10.2139/ssrn.2860048.

Solomon, Miriam. 2001. *Social Empiricism*. Cambridge, MA: MIT Press.

Sullivan, Emily. 2019. "Understanding from Machine Learning Models." *British Journal for Philosophy of Science*. axz035. doi: 10.1093/bjps/axz035.

Starr, Sonja B. 2015. "The New Profiling: Why Punishing Based on Poverty and Identity Is Unconstitutional and Wrong." *Federal Sentencing Reporter* 27 (4): 229–36.

State v. Loomis. 2016. Supreme Court of Wisconsin, 881 N.W. 2d 749. https://casetext.com/case/state-v-loomis-22.

Tashea, Jason. 2019. "France Bans Publishing of Judicial Analytics and Prompts Criminal Penalty." *ABA Journal*, June 7. http://www.abajournal.com/news/article/france-bans-and-creates-criminal-penalty-for-judicial-analytics.

Turek, Matt. 2018. "Explainable Artificial Intelligence (XAI)." DARPA. https://www.darpa.mil/program/explainable-artificial-intelligence.

Verma, Sahil, and Julia Rubin. 2018. "Fairness Definitions Explained." In *Proceedings of the International Workshop on Software Fairness*. New York, NY: Association for Computing Machinery. https://doi. org/10.1145/3194770.3194776.

Wilholt, Torsten. 2009. "Bias and Values in Scientific Research." *Studies in History and Philosophy of Science* 40: 92–101.

Wilholt, Torsten. 2013. "Epistemic Trust in Science." *British Studies in Philosophy of Science* 64: 233–53.

Cite this article: Biddle, J. B. 2022. On Predicting Recidivism: Epistemic Risk, Tradeoffs, and Values in Machine Learning. *Canadian Journal of Philosophy* 52: 321–341, doi:10.1017/can.2020.27

https://doi.org/10.1017/can.2020.27 Published online by Cambridge University Press