CHAPTER 4
Balancing AI Bias

Michael Mattioli[1]

Writers began worrying about artificial intelligence ("AI") hundreds of years before the technology existed. Ancient Jewish folklore contains tales of golems – human-like creatures made of mud and brought to life through magical incantations. Many golem stories end in disaster when the creature unthinkingly carries out its creator's instructions too literally. One legend tells of a golem which, ordered to carry buckets of water to a house, dutifully did so until the building flooded. These fictional creatures have reappeared in countless forms over the centuries, from Frankenstein's monster, to the dancing brooms in *The Sorcerer's Apprentice*, to the robot named HAL in *2001: A Space Odyssey*. Golem stories caution us about the power of unintended consequences and the dangers of hubris. They also teach us that if we are not careful, the things we bring into the world will mirror our worst traits.

Today, the golem is real. AI software brought to life through the language of code can mimic many forms of human decision-making with startling acuity. Because it is highly versatile, AI is swiftly permeating countless corners of society, and it has been embraced by institutions that have power over our lives. Many employers are using AI systems to evaluate job candidates, for instance. Police forces are using AI to make predictions about where crimes are likely to occur. Judges are using AI tools to advise them on criminal sentencing decisions. The technology is also becoming integral in the fields of healthcare, insurance, automotive technologies and finance.

Academics and journalists have documented a dark side of AI, however: it has the power to amplify human biases, aggravating longstanding sociological discrimination. AI bias is an urgent problem, and experts in law and public policy should take the time to thoroughly understand its causes. Although this problem is widely discussed in the media, there are few technical explanations of AI geared toward lawyers. In press accounts, the problem is often vaguely attributed to biased

1. Special thanks to Mary Christie for research assistance in connection with the literature review.

algorithms, biased data, or trite adages like "garbage-in, garbage-out".[2] In reality, the problem is nuanced, its causes are varied, and solutions are more elusive than one might guess.

This chapter is a primer on AI bias geared toward readers with background knowledge of law and public policy. This chapter also calls attention to a hopeful possibility that has curiously gone overlooked in the recent surge of negative media commentary: AI could *mitigate* the effects of human biases in many of the settings where it has raised concerns.

A. EVERYTHING EASY IS HARD; EVERYTHING HARD IS EASY

In a 1950 paper, the mathematician and computer scientist Alan Turing foretold the future with startling clarity.[3] In an age when the word "computer" referred to a job title and decades still before engineers at Bell Labs developed the first silicon transistor, Turing was exploring the theoretical limits of machine intelligence. He wanted to know whether a computer might someday be able to think like a human. Although this topic was (and remains) in the realm of science fiction, Turing brilliantly framed it in practical terms: instead of asking whether a human-level AI was *possible*, he searched for evidence that it was not.

To aid this search, Turing proposed his famous "imitation game": a thought experiment involving a human interrogator who must type questions to a subject in a different room. The hypothetical interrogator must guess whether the subject is a human or a computer by relying only on the subject's written responses. This setup allowed Turing to consider, and summarily dispense with, barriers that might prevent a sufficiently powerful computer from fooling the human interrogator. Turing concluded that there exists no engineering barrier that should prevent a computer from perfectly imitating a human. It was just up to programmers to teach computers how to mimic human responses.

Turing assumed that to behave like a human, an AI would need to reason like one. This was a sensible assumption. Deductive and inferential reasoning is how people have investigated, explained and discussed the world for thousands of years. It is how we predict the future and make sense of the past. Consider a criminal defence attorney who presents a jury with cellphone location data. The data indicate that her client, a larceny suspect, was in a different city than the victim on the night in question. The attorney then argues that her client cannot be guilty of theft because, for a robbery to occur, the thief must be in the same place as the victim. This example shows two kinds of reasoning relevant to understanding AI: inference and deduction. The factual conclusion that the suspect was in a different city is an inference, which expresses a probability; the legal conclusion that the suspect cannot be guilty if he was in a different city is a deduction that follows from the law.

---

2. As *Wired* magazine wrote, "flaming garbage in, ad flaming garbage out". *Forbes*: "Whatever goes in, goes out." < https://www.forbes.com/sites/nicolemartin1/2018/12/13/are-ai-hiring-programs-eliminating-bias-or-making-it-worse/#3453d40622b8 > .
3. Anthony Beavers, 'Alan Turing: Mathematical Mechanist', in S. Barry Cooper and Jan van Leeuwen, *Alan Turing: His Work and Impact* (Elsevier, Waltham, 2013) pp. 481–485.

Why is it so difficult to teach a computer to reason this way? Turing, like his contemporaries, assumed programmers could accomplish the task by teaching a computer some common sense about the world.[4]

> "[T]he [programing] would be largely occupied with definitions and propositions. The propositions would have various kinds of status, e.g., well-established facts, conjectures, mathematically proved theorems, statements given by an authority, expressions having the logical form of proposition but not belief-value."[5]

Well into the 1980s, much of AI research was premised on such a "rule-based" model. Ultimately, though, the task was just too cumbersome.[6] Researchers discovered that concepts even a child can grasp require unimaginably vast and complicated webs of knowledge. Programming the ideas, rules, and information that an adult draws upon effortlessly in conversation would require nothing short of building a model of all of reality as humans perceive it. As the late Carl Sagan once wrote, "If you wish to make an apple pie from scratch, you must first create the universe."[7]

Let us return to our example: imagine that we wish to teach a computer how to predict the outcome of our criminal trial. Assume that the defence attorney convinced the jury that the suspect was in a different city than the victim. (We will say the suspect was in Paris and the victim was in Milan.) We want our computer to predict the likely outcome of the case from this information alone. To do that, our computer would need the ability to parse the English language into abstract concepts with which it could work. For instance, the machine would need to understand that the word "theft" refers to one person illegally taking another person's property. The machine would also need to have concepts of "taking" and "property" that it could apply to the facts. The system would need to know that the word "jewellery" refers to a kind of property, and that, to take jewellery, a thief must usually be in the same place as the jewellery. That is just the beginning, however. Our machine would need to know that Paris and Milan are two different cities; that cities are designations given to geographic areas; that two geographic areas cannot occupy the same region of space; that a person cannot be in two places at the same time; and so on. Even if we could program all of these common sense rules and facts beforehand, our machine would be unprepared to examine the facts of any new case with slightly different events or legal issues. What is easy for humans is profoundly tricky for computers.

---

4. . Cleverly, he argued that it would only be necessary to program an AI with the level of knowledge that a child possesses – like a child, the machine would then be able to develop an adult's intelligence through learning.
5. Alan Turing, 'Computing Machinery and Intelligence', 59(236) *Mind* (Oct., 1950).
6. A recent analysis of scholarship: "The biggest shift we found was a transition away from knowledge-based systems by the early 2000s. These computer programs draw on the assumption that all human knowledge can be reduced to rules. In their place, researchers turned to machine learning—the parent category of algorithms that includes deep learning." Karen Hao, 'We Analyzed 16,625 Papers to Figure out Where AI is Headed Next', *Technology Review*, 25 January 2019, < https://www.technologyreview.com/s/612768/we-analyzed-16625-papers-to-figure-out-where-ai-is-headed-next/ > .
7. "But as those projects unfolded, researchers hit a major problem: there were simply too many rules that needed to be encoded for a system to do anything useful. This jacked up costs and significantly slowed ongoing efforts." *Ibid*.

Modern-day AI researchers have drawn inspiration from the opposing truth: computers can excel at tasks that are difficult for humans. In place of rules, knowledge, and human-like deductions, modern AI systems imitate some aspects of human thought by spotting statistical patterns in vast droves of data. For example, a recently developed AI has correctly predicted the outcome of Supreme Court decisions with greater accuracy than human experts who closely follow the Court's decisions. Computer programmers did not provide this system with any knowledge of the law or facts about the world. Instead, they translated case decisions into data that a machine could work with, including case features and outcomes. Then, they trained their AI to recognize relevant patterns in the data.[8] The system can predict the outcome of a case it has never seen by spotting parallels to other court decisions it already has on hand. This statistical approach to AI, which is called "machine learning", is explained in greater depth in the following section.

For all its power, machine learning has a downside: unmoored from ethics and common sense, these systems can behave in ways that are inscrutable, unpredictable and unfair. As Professor Deidre Mulligan commented in a recent magazine interview, "the most powerful algorithms being used today 'haven't been optimized for any definition of fairness; they have been optimized to perform a task'".[9] An early example of this problem occurred in March 2016 when Microsoft Corporation introduced a chatbot by the name of "Tay" to the world. Taking a page from Turing's imitation game, Microsoft designed Tay to provide written responses to internet users' comments and questions. At its core, Tay was a machine learning algorithm that engineers had trained on large sets of written conversations culled from public sources, such as Twitter. The system also absorbed and incorporated the messages that internet users typed to it. It took just a few days for users to inundate Tay with racist and misogynist comments, which the AI dutifully incorporated into its personality and parroted back to other users. For Microsoft, the incident was an embarrassing public relations fiasco. For AI researchers, Tay illustrated a familiar problem with machine learning: like children, the machines can inherit the messiness of human thought.

There is far more at stake than rogue chatbots spouting bigotry. Our society is steadily outsourcing important decisions to AIs – from credit scoring, to criminal sentencing, to hiring and firing, to healthcare decisions, to automotive safety. Legal experts and technologists have warned that AI-based sociological discrimination in these settings could not only harm individuals, but it may gradually restructure society in damaging ways. Then again, the story is not so simple. AI has the potential to *mitigate* the impact of human bias in some of the settings where it has caused concerns. For us to make sense of this technology's potential for good and for ill, we first must understand how it works – the focus of the next section.

8. Lawyers in the near future might pit AIs against each other: Matthew Hutson, 'Artificial Intelligence Prevails at Predicting Supreme Court Decisions', *Science*, 2 May 2017, < https://www.sciencemag.org/news/2017/05/artificial-intelligence-prevails-predicting-supreme-court-decisions > .
9. Jonathan Vanian, 'Unmasking A.I.'s Bias Problem', *Fortune*, 25 June 2018, < https://fortune.com/longform/ai-bias-problem/ > .

B. MACHINE LEARNING FOR LAWYERS

The term "machine learning" is widely defined as "the field of study that gives computers the ability to learn without being explicitly programmed".[10] At a high level, the idea is easy to understand: in place of instructions, we provide a computer with examples from which it can learn. In a sense, the computer then programs itself. Imagine that we have been presented with the results of the multiplication of two numbers. We know one of the numbers, we know the result, but we do not know the second number (which is written below as "?"). Our job is to guess what it is:

$$8 \times ? = 40$$

We can easily deduce that the missing number is 5 if we divide 40 by 8. But suppose we are not allowed to use division and, to make things even trickier, let us make the numbers larger:

$$79 \times ? = 1{,}343$$

A simple (albeit tedious) way to solve this problem would be to guess. We will try a number and then raise or lower our next guess depending on the result. Our plan, or algorithm, for finding the right result will work as follows: Our first guess will be "1", and if that is too low, our second guess will be "100." To home in on the right answer, we will always guess the integer mid-point between the two numbers that we know to be too high and too low. In other words, if we know that "1" is too low and "100" is too high, our next guess will be "50." Our series of guesses look like this:

Table 1: Learning through iterative guesses

Factor we know	Factor we are guessing "?"	Product	Next guess higher or lower?
79	10	790	Guess higher
79	100	7,900	Guess lower
79	50	3,950	Guess lower
79	25	1,975	Guess lower
79	12.5	9,87.5	Guess higher
79	18.75	1,481.25	Guess lower
79	15.625	1,234.375	Guess higher
79	17.1874	1,357.8125	Guess lower

Our algorithm is slowly but steadily getting to the right result, which is "17". A similar process is at the heart of machine learning: we began with an example of a right

10. This definition is widely attributed to an eminent engineer named Arthur Samuel. The original source of this famous definition is quite difficult to identify, however. The source most frequently cited, Samuel's 1959 paper on machine learning as applied to the game of checkers, does not contain this quote. A.L. Samuel, 'Some Studies in Machine Learning Using the Game of Checkers', 3(3) *IBM Journal of Research and Development* (July 1959).

answer – here, "1,343" – and some input data – "79" – and through trial and error, we figured out the missing piece.[11] This process, known as supervised learning, is the most common form of machine learning in use today.[12]

Now let us consider a more interesting example that will help us better understand AI bias. Suppose we are financial managers at a large public library, and we have a tough budgeting task: we want to predict how much money our library needs to spend on new books for the year ahead. We have never created such a plan before. In the past, we have always purchased books in response to patron demand as the year went along. Conveniently, we have several pieces of data on hand that seem helpful: for each of the past ten years, we know the number of books that were checked out, the number of patrons who visited the library, and the amount of money we spent each year. Our assistant compiles this data into a table as in Table 2.

Table 2: Library data

Year	Number of books checked out	Number of patrons who entered the library	Amount spent
2019			?
2018	18,123	175,123	$8,000
2017	13,563	177,734	$8,463
2016	15,124	180,102	$7,640
2015	11,735	180,878	$9,123
2014	17,563	182,012	$9,320
2013	17,998	185,102	$10,200
2012	21,021	190,201	$9,920
2011	20,001	189,203	$9,515
2010	19,232	181,020	$11,000

Based on this table, we believe that the data we have about the past has something useful to tell us about the future. The gate-count and the checkouts probably do not carry equal predictive weight, however. After all, a patron entering the library will not necessarily check out a book.

Since we would like to have a computer do our work for us, we need to think about how to write our assumptions down in a way that a machine can understand – a simple equation. This is where data scientists earn their keep: there are many possible ways to express the foregoing problem as a formula. In this case, an expert would select an equation that reflects an essential fact about the data in the table: when either gate count or checkouts go up, the amount of money spent goes up too. Mathematicians call this a linear relationship, and they can express it in the form of a multivariate linear equation like the one below.[13] We will also reflect our assumption

11. A closely related technique called unsupervised learning in which machine finds patterns in data that do not have any examples of right answers.
12. Other forms include unsupervised learning and reinforcement learning.
13. Assuming a parametric method.

that checkouts and gate counts might not carry equal weights by multiplying each number by some unknown values, "weight1" and "weight2."

$$(checkouts \times weight1) + (gatecount \times weight2) = money\ needed\ for\ next\ year$$

If we can figure out the correct values for "weight1" and "weight2", we should be able to feed the data from any prior year in our table and receive the corresponding amount of money spent. Plugging in the numbers from 2013, for instance, should yield the money spent on new books in 2014:

$$(17{,}998 \times weight1) + (185{,}102 \times weight2) = \$9{,}312$$

But how can we learn what the right weights are? In essence, we do so by telling our computer to guess and guess again until it finds the correct values. (The way our computer does this is beyond the scope of this chapter, but it involves some clever mathematics that saves the computer time.) After thousands or potentially millions of combinations of numbers, the computer discovers that the following values for the weights accord with the annual expenses in the table:

$$(checkouts \times 0.25) + (gatecount \times 0.026) = money\ needed\ for\ next\ year$$

We will need to take some additional steps before we can trust our algorithm. For instance, we will want to make sure that it will give helpful results for numbers other than those we trained it on. Statisticians call this problem "overfitting." To do this, we might try putting in gate counts and checkouts from years that we did not have in our table – perhaps from the decade prior. Assuming our evaluation step goes well, we can feel satisfied that we have the correct equation. If we use this equation with data from the past year, we can finally get our prediction for next year's spending:

$$(19{,}203 \times 0.25) + (200{,}020 \times 0.026) = \$10{,}001.27$$

There are two important things to take away from this exercise: first, machine learning prioritizes optimization over knowledge; secondly, it requires many subjective assumptions. Consider the fact that we have taught a computer how to complete a task by providing it with examples instead of rules. We gave our machine no instructions that are special to libraries, patrons, or books. This is why the designer of a machine-learning algorithm could technically appear to have "solved" a problem without having a deep or nuanced understanding of it. If fed different inputs, the same algorithm (or a similar one) could help predict the value of a home or the likelihood of a criminal committing repeat offences. A slightly different machine learning system could identify photos or cats or even predict the appropriate response to a written question. A second important fact is that we, as the algorithm's designers, don't understand why the particular weights our system settled on work the best. Our process gives us no insight into why the numbers "0.25" and "0.026" worked best – they just did.

Consider also the many assumptions we made along the way. First, we assumed that it would be possible to predict our future budgetary needs. Secondly, we selected two pieces of information that we believed could be useful for that purpose. Thirdly, we assumed that the data was accurate and complete. Fourthly, we decided that we wanted a numerical prediction of next year's budget as opposed

to, say, a binary result such as "spend more than last year". Fifthly, we selected a specific kind of equation – a multivariate linear equation – to express our prediction. The point is that this process is riddled with judgment calls. Figure 1 summarizes our entire process.

Figure 1: Common steps in designing a machine learning system that relies upon supervised learning

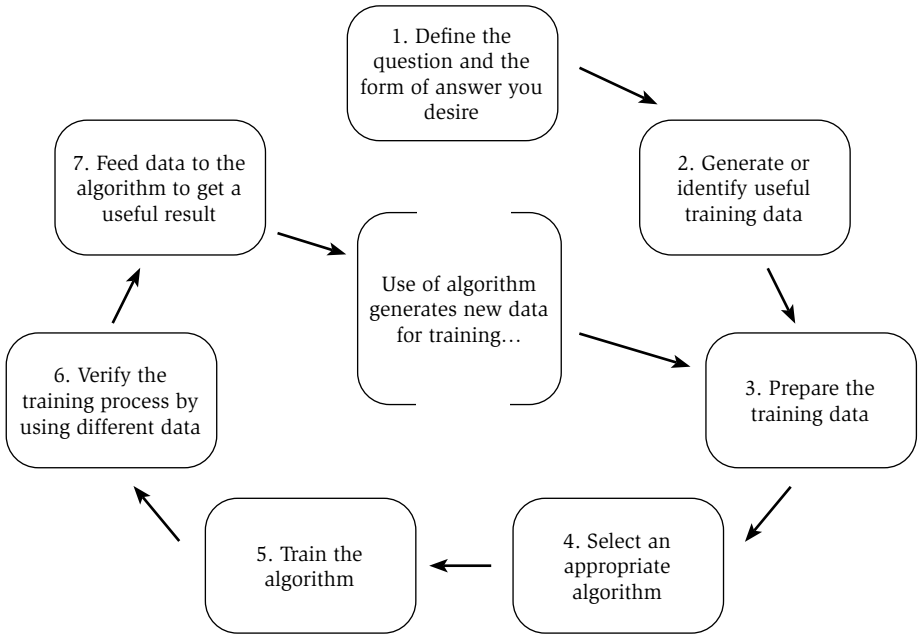

Figure 1 is highly simplified. The process of developing and training a machine-learning system is not typically as linear or sequential as it appears. As Paul Ohm and David Lehr explained in a recent publication, "much machine learning dances back and forth across our steps instead of proceeding through them linearly".[14] Despite its limitations, however, Figure 1 serves as a helpful reference as we explore the sources of AI bias.

C. SOURCES OF AI BIAS

In the recent outpouring of commentary on the dangers of AI, the term "bias" has become synonymous with unfairness. The word has several distinct definitions,

14. David Lehr (FNa1) and Paul Ohm, 'Playing with the Data: What Legal Scholars Should Learn About Machine Learning', 51 *U.C. Davis L. Rev.* (2017) pp. 653, 669.

though.[15] In science, bias can refer to a set of data that does not accurately describe an underlying phenomenon because it is incomplete or skewed – e.g., sampling bias. In psychology, a cognitive bias is a systematic error in reasoning, such as the tendency to underestimate how long a particular task will take. The word can refer to prejudice, leading to unfair treatment based on race, ethnicity, class, gender, or some other characteristic. By and large, this last problem concerns legal scholars the most. As the following discussion explains, different kinds of bias can lead to this result. Vexingly, however, the problem does not require *any* form of biased data, biased algorithms, or biased thinking in order to occur.

I. The Questions That We Ask

The very decision to use machine learning in the first place can lead to unfair results. In our library budgeting example, consider our assumption that next year's spending was, in some sense, predictable. That is, we assumed that we could anticipate this particular fact about the future by relying on data from the past. Not everything about the future is knowable. Had our assumption been wrong, our algorithm would not have told us so – instead, like an unthinking golem, it would have dutifully given us year after year of poor estimates. Ideally, we could catch this weakness during the validation stage – Figure 1, Step 6. But had we missed the problem, we might rely on poor estimates for years before realizing that the algorithm was not improving. Had this algorithm been tasked with a job more critical to human life than a library acquisition budget – a criminal sentencing recommendation for instance – the mistake would not only be unfair and harmful, but it would be difficult or impossible to detect after the fact. A judge might blindly rely on the recommendation – an ostensibly value-neutral decision that yields a biased result.

II. The Answers That We Ask For

Consider the type of answer that we designed our system to deliver: a precise dollar value. Suppose instead that we had designed our algorithm to characterize a prior year's spending as either "too aggressive" or "too cautious" for the year ahead. Such a system might have been helpful for offering general advice about adjusting spending, but it also could create problems. Imagine that every year, the library's director evaluates her employees. As finance managers, our performance is rated partially on how sound our budgetary decisions have been. If the labels we have told our AI to use – "too aggressive" and "too cautious" – are incorporated into our annual review, our director might wrongly conclude that we have been making poor budgeting decisions.

In a recent essay, Solon Barocas and Andrew Selbst lucidly explained how this sort of problem could lead to discriminatory hiring practices. Focusing on labels, they wrote, "While different choices for the […] class labels can seem more or less reasonable, valid concerns with discrimination enter at this stage because the different

15. Cassie Kozyrkov, 'What is AI Bias?' Jan. 24, 2019 < https://towardsdatascience.com/what-is-ai-bias-6606a3bcb814 > .

choices may have a greater or lesser adverse impact on protected classes. For example […] hiring decisions made on the basis of predicted tenure are much more likely to have a disparate impact on certain protected classes than hiring decisions that turn on some estimate of worker productivity."[16] Once again, no discernible bias or "garbage-in" is necessary for AIs to lead to facilitate bias. The information that AIs report to us, the ways they report this information, and the ways that we use it can serve as the foundations for unfair treatment.

III. The Amount and Sufficiency of Data We Include

Failing to include certain types of data can also lead to skewed results. In our library example, we selected two pieces of information as inputs: the number of checkouts and the library's total gate-count in the prior year. Suppose, though, that our data on checkouts related only to physical volumes and not e-books. Suppose also that e-books are highly popular among our library's patrons, and that nearly 80% of all science fiction books checked out are in an e-book format. Our algorithm would not only under-allocate funds for collection development overall, but this under-allocation will have a disproportionate impact on science fiction readers.[17] Once again, no deliberate human bias is required for our system to yield an unfair result.

Legal scholars have recently described how similar mistakes can lead to more severe harms. In a 2016 publication, Mikella Hurley and Julius Adebayo explored how the data for credit scoring software that lenders rely upon can lead to unfair decisions and discrimination, sometimes leading to the systematic denial of credit access to specific groups.[18] They reported on a man in his late 20s, for example, with a sterling credit history whose credit score quickly dropped when a lender noticed that he shopped at certain stores frequented by individuals with poor credit. The algorithm that determined the scores heavily weighted shopping habits and, in this case, overlooked the fact that the man's history of paying his bills was excellent – a case of credit scoring by affiliation. The authors argued that current laws are insufficient to respond to challenges like this one.

Even if we had included all of the relevant types of data, we might just not have enough of it to provide a complete picture. In a 2018 article, I examined this problem of under-representative datasets as it affects cancer research.[19] Over the past ten years, several nonprofit and professional organizations in the United States have attempted to aggregate cancer treatment data to train machine-learning tools to provide treatment recommendations to doctors. Some experts estimate that better treatment decisions could dramatically reduce the mortality rate from various cancers. When I interviewed experts involved with these efforts, however, I learned that hospitals, pharmaceutical firms and academic researchers were reluctant to share

16. Solon Barocas and Andrew D. Selbst, 'Big Data's Disparate Impact', 104 *Cal. L. Rev.* (2016) pp. 671, 680.
17. A related problem, mentioned earlier, is training an algorithm on data that is not probative – it can lead to overfitting.
18. Mikella Hurley and Julius Adebayo, 'Credit Scoring in the Era of Big Data', 18 *Yale J.L. & Tech.* (2016) p. 148.
19. Michael Mattioli, 'The Data-Pooling Problem', 32 *Berkeley Technology Law Journal* (2018) p. 179.

such data. The reason, in large part, stemmed from concerns over competition and public perceptions. Interview subjects told me that hospitals are concerned that sharing patient treatment records could reveal a pattern of poor treatment outcomes that could lead patients to visit competing care providers. Pharmaceutical firms similarly expressed concerns that sharing clinical trial data could provide their competitors with hints about future research plans. Academic researchers, some of whom possess a wealth of useful data to share, explained that they needed to guard their data so they could write important publications and advance professionally.[20] The upshot, generally, is that many institutions with valuable data are unable to see beyond the population of patients or customers they serve. An AI trained only on such data may yield inaccurate or unfair results for groups and individuals that are under-represented.

IV. Data Provenance and Pedigree

An understanding of data's provenance and pedigree is vital to getting fair results. Continuing with our library example, suppose that our data about the number of patrons who had visited the library (the "gate-count") was inaccurate. Perhaps one of the machines installed in a building entrance was broken, leading us to underestimate the number of patrons last year. As a result, our algorithm recommends a budget that is far too low in light of demand in the next year. Meanwhile, the library decided to spend funds that should have gone toward books on new computer terminals. As Harry Surden has written, "In general, machine learning algorithms are only as good as the data that they are given to analyze."[21]

But the problem of data quality is subtler than whether data is merely good or bad. Data that is useful for answering one question can lead to unfair or harmful results when applied to a different type of problem. In a 2014 publication, I interviewed a data scientist who had been hired by a hospital network based in the United States. His employer had asked him to compile cancer patient treatment records from several hospitals into a single database that could be used by researchers. As he delved into the data, the researcher found that doctors at a particular hospital had been routinely neglecting to note the biological sex of particular patients. When he inquired about this missing data, the data scientist was told by a manager that the patients in question had identified themselves as transgender and that their doctors were unsure whether they should indicate "male" or "female" on the patients' health records. The data scientist was dismayed over the hospital's apparent disregard for this class of patients. He also told his supervisor that the data would be useless without at least these patients' biological sex identified. The supervisor instructed the researcher just to make guesses made on factors such as the types of cancer involved – prostate cancer would suggest a biologically male patient, for example – and other features, such as height and weight. The database almost surely contained many inaccuracies.

20. *Ibid*. Similar patterns seem present in other industries. In the autonomous vehicle industry, for instance, vehicle training data is closely guarded by manufacturers despite the fact that widespread data-sharing could lead to smarter cars. Michael Mattioli, 'Autonomy in the Age of Autonomous Vehicles', *Boston University Journal of Science and Technology Law* (2018).
21. Harry Surden, 'Machine Learning and Law', 89 *Wash. L. Rev*. (2014) pp. 87, 106.

Sometimes these mistakes would be harmless. For instance, if the database was used to train an algorithm to predict the rate of cancer overall in the United States next year, there might be no problem. On the other hand, the data offer a skewed picture of the rate of cancer among women in a particular zip code. An AI trained on this data could possess a distorted view of specific patient groups, which could, in turn, lead to poor medical decisions.

Sometimes, data is manipulated not to correct errors, but instead to preserve anonymity. In the US, for instance, the Health Insurance Portability and Accountability Act ("HIPAA") requires healthcare providers to remove specific personality-identifying information from health records before sharing them with third parties, such as researchers. Such information includes names, zip codes, treatment dates, and the like. To comply with the law without rendering the dataset useless, data scientists often mask and suppress particular information. For instance, it is possible to preserve the fact that a patient sought treatment for three months by altering the patient's exact treatment dates by a specified offset. Other techniques are far more sophisticated. A machine learning algorithm trained on such data might be able to make useful predictions about future patients, but could be blind to a problem that occurred at the hospital during a specific data range – a pattern of negligent behaviour by a single doctor, for instance. In settings where data contains subjective judgments, context is king.

The above problems persevere when institutions that correct errors or anonymize useful data fail to carefully document and disclose the details of those steps. As explained in a later section of this chapter, laws that seek to encourage technological disclosures do not address this metadata disclosure problem.

V. The Algorithms We Select

We assumed that a multivariate linear equation would be the most appropriate for our exercise. We did this because we guessed that there was a linear relationship between the gate count, checkouts, and spending, but we could have been wrong. If we were wrong, perhaps we should have used a different kind of machine learning technique better suited for non-linear relationships. A technique called logistic regression, for instance, is well suited to tackling classification problems, such as whether a criminal convict is likely to be a repeat offender. Another machine learning technique called a neural network is excellent at highly complex problems like image recognition and game playing. There are also "unsupervised learning" processes for tackling problems that have no "correct" answer, such as finding unknown patterns in large sets of data.

VI. AI Bias is Hard to Detect or Predict

The opacity of the machine learning process aggravates the problems discussed thus far in two ways: first, it makes it difficult for anyone to detect biased decision-making; secondly, it obfuscates from those subjected to an algorithm's decisions any clues about how to tailor their behaviour.

As the discussion in the next section shows, the legal framework channels AI inventions toward secrecy in a number of ways. But the problem is more profound:

recall that machine learning involves replacing explicit instructions with a process of learning from samples. A consequence of this is that often, we cannot explain the "rules" that an AI has taught itself. Our library budgeting algorithms, for instance, identified two "weights" that did not truly mean anything to us. We just understood that they worked. If we wanted to understand how the system arrived at those two results, we would need to step back through every permutation of "weight 1" and "weight 2" that our hypothetical computer considered. Doing so would entail analysing thousands or millions of combinations. That is just an example with two missing weights. In real-world AI applications, the number of variables, weights, inputs, and training data samples can number in the thousands or millions, making any retrospective analysis entirely impracticable. Often, even the engineers who design machine-learning systems cannot explain precisely how they work.

Nicholson Price explained this issue in a recent article as follows: "In many cases, [machine-learning systems] generate algorithms that are unavoidably opaque. These algorithms typically cannot identify the reasons for the patterns they find, due to the iterative process by which the algorithms are developed […] And even when patterns discovered by an algorithm can be stated, those patterns are typically far too complex to be of much use in understanding underlying mechanisms."[22] Ultimately, we not only risk unfair treatment, but neither those who run the machinery nor those subject to it have the power to detect that unfairness.

Relatedly, it is important to appreciate that AI bias is not always immediately apparent. An algorithm may produce fair and useful results until a particular set of inputs causes it to yield a result that humans could not expect. In essence, everything works fine until it does not. The opacity of the algorithms can make these failures very difficult to identify and prevent, however. As Zeynep Tufekci said in a 2017 TED Talk, "We don't really understand what the system learned – that's its power."[23]

VII. Feedback Loops Compound the Problem

AI bias can also be aggravated through feedback loops. When people rely on AIs to make decisions, those decisions can, in turn, generate new data that is tinted with the prior decision the AI has made. In *Weapons of Math Destruction*, Cathy O'Neal explained this phenomenon through a worrisome example: "[W]hen poor people and immigrants qualify for a loan, their substandard language skills might drive up their fees. If they then have trouble paying those higher fees, this might validate that they were a high risk to begin with and might further lower their credit scores. It's a vicious feedback loop, and paying bills on time plays only a bit part."[24] O'Neal goes on to lucidly explain how a similar process can affect other corners of daily life. For instance, when hiring decisions are influenced by algorithms that take into account the credit scores of job applicants, applicants who have low credit scores for reasons outside their control – medical bills, say – are less able to get hired, which in turn

22. Roger Allan Ford and W. Nicholson Price II, 'Privacy and Accountability in Black-Box Medicine', 23 *Mich. Telecomm. & Tech. L. Rev.* (2016) p. 1.
23. < https://www.ted.com/talks/zeynep_tufekci_we_re_building_a_dystopia_just_to_make_people_click_on_ads > .
24. Cathy O'Neil, *Weapons of Math Destruction* (Crown, 2016), p. 189.

causes their credit scores to fall further. As O'Neal puts it, "poisonous assumptions are camouflaged by math and go largely untested and unquestioned."[25] Legal scholars whose work is discussed later in this chapter, have documented how feedback loops can reinforce bias in many contexts.

D. HOW THE LEGAL FRAMEWORK CONTRIBUTES TO THE PROBLEM

The preceding discussion explained how AI bias stems from human decisions, and how feedback loops, algorithm opacity, and unpredictability aggravate the problem. These technical problems are perpetuated in part by a legal framework that channels data, metadata, databases and algorithms toward secrecy.

The legal framework pertaining to data and algorithms does not actively encourage disclosure. Patent law's application procedure is designed to promote technological disclosures, for instance, but data itself is a generally unpatentable subject matter. The patentability of machine learning algorithms, meanwhile, is uncertain and context-specific. In part, this is a result of recent Supreme Court jurisprudence that cast doubt on whether algorithms qualify as patentable subject matter.[26] There are other barriers to the patentability of algorithms, though: as explained earlier, the machine-learning training process can yield algorithms that work in ways that not even their designers can understand or explain. In the US, this could hinder their ability to meet patent law's written-description, enablement and definiteness requirements.[27] In the end, patent protection seems to be available mostly around the margins. A method of gathering data, or of cleaning or anonymizing data, or for training an algorithm, for instance, might be eligible.

Copyright protection for individual data, meanwhile, is relatively insubstantial: it doesn't extend to individual pieces of data because facts are generally unprotectable. Datasets, meanwhile, can be copyrighted if they reflect originality in the selection or arrangement of data. Such originality may be present where professional judgment has gone into the selection of training data. As a practical matter, though, the issue is rarely litigated because trade secrecy, contracts and encryption make data copying difficult. As explained earlier, metadata necessary to determine data's provenance and pedigree often go undocumented altogether, or may accompany data and rely on similar protections against copying. Moreover, any copyright that attaches to data selection and arrangement will likely be relatively thin.

Meanwhile, the law actively encourages secrecy in several ways. First, trade secret protection could pertain to nearly any subject matter related to an AI – from training data to algorithms to processes for preparing data for analysis. Such secrecy can be bolstered by contracts that limit the power of data-subjects such as customers and patients to asserts complains based on privacy or tort. Regulatory action designed to protect consumer privacy or regulate competition may serve as a backstop, however. Table 3 illustrates this channelling toward secrecy.

---

25. *Ibid.*, p. 13.
26. See, e.g., *Alice Corp. v. CLS Bank International*, 573 U.S. 208 (2014) (holding unpatentable algorithms cannot be made patentable simply by implementing them on computers).
27. For a deeper discussion, see W. Nicholson Price II, 'Big Data, Patents, and the Future of Medicine', 37 *Cardozo L. Rev*. (2016) p. 1401.

Table 3: How the law influences data use

	Data	Metadata	Datasets	Algorithms
Copyright	Unavailable	Weak: Possibly available but thin	Weak: Possibly available but thin	Possibly available but uncertain: Source is code potentially copyrightable, but protection does not extend to functional elements or ideas. Moreover, human authorship may be difficult to establish due to the training process.
Patent	Generally unavailable: Possibly available for methods of generating or collecting data.	Generally unavailable: Possibly available for methods of preparing data – e.g., cleaning, anonymizing, etc.	Generally unavailable: Datasets are not generally a patentable form of subject matter.	Possibly available but uncertain: Potential problems with subject matter, enablement, and disclosure requirements
Trade Secrecy	Trade secret protection generally available (at the state and federal level), provided that other threshold requirements are met.			
Privacy Laws & Regulations	Various privacy and consumer protection laws at the state and federal levels may limit the types of data that institutions can collect, whether and how this data may be shared, and how the results of such analyses may be used.			
Contracts	Corporations and other institutions that collect consumer data widely use end-user-license agreements and similar form-contracts to establish permissive uses. These contracts may provide the data collector with permissions that privacy laws (and tort law) would otherwise forbid. Separately, data-collectors and brokers rely on contracts to establish the boundaries of permissible uses.			

It's easy to appreciate how this framework might perpetuate some of the forms of AI bias outlined earlier in this chapter. First, secrecy can make it difficult to assemble a large enough dataset to get good results – i.e., the problem of under-representation. Secondly, secrecy can make metadata unclear, leading to misuse of data. Thirdly, secrecy can contribute to the problems associated with AI opacity, making it difficult for researchers to interrogate how an algorithm arrived at a particular result.

E. POLICY GOALS AFFECTED BY BIAS

Having examined AI bias and how the legal framework may contribute to it, we turn in this section to its impact on public policy goals. The following is a non-exhaustive summary of recent legal scholarship on point. At the time of writing, there is an

outpouring of legal scholarship on point. Rather than report an exhaustive catalog of this expanding body of work, the following discussion aims to provide readers with a sampling of the many topics that are being explored.

An article that merits special credit in this summary is David Lehr and Paul Ohm's 'Playing with the Data: What Legal Scholars Should Learn about Machine Learning'. This piece offers an excellent deep-dive into the process of machine learning, focusing on the distinct workflow the authors identify as "playing with the data".[28] Lehr and Ohm contend that most scholarship has focused on the workflow they identify as "the running model", thereby neglecting the possibilities and pitfalls arising from the playing workflow. (The present book chapter's high level discussion of machine learning was inspired by the authors' focus on algorithm design and development.) The authors identify the potential costs and benefits machine learning poses for our legal system, including discrimination, reason-giving, and due process and inaccuracy. Referring to machine-learning systems as "inscrutable black boxes", Lehr and Ohm pay special attention to harms in criminal justice and procedure. Striking an optimistic note, they point out that the human involvement required in machine learning could provide opportunities for policy prescriptions.

Jessica Eaglin has written extensively on judges' use of machine-trained software to make sentencing recommendations.[29] After reporting that the use of such tools is widespread and steadily expanding, Eaglin zeros in on many of the sources of bias outlined earlier in this chapter. She explores, for instance, the threshold question of whether future criminal conduct can and should be predicted in the first place; the selection of what outcomes these tools are designed to predict; what factors (inputs) these tools should use to make their predictions; and how sentencing decisions trigger feedback loops that re-enforce unfair treatment.

Solon Barocas and Andrew Selbst have analysed data-mining bias and discrimination concerns in the United States through antidiscrimination law and Title VII's employment discrimination provisions.[30] Barocas and Selbst provide a look at the steps involved in solving problems with data mining and the ability of data miners to make intentional discrimination to appear accidental. Next, they review Title VII jurisprudence as applied to data mining. Finally, the authors examine the difficulties faced by reformers in addressing those issues identified in the second part of the article.

In a 2014 article, Danielle Keats Citron and Frank Pasquale examined problems associated with the use of AI to assign credit scores.[31] They argue that current systems have a disparate impact on minorities and women – problems that are perpetuated by the opacity of algorithms. One problem is that, because consumers don't know how the system works, it is difficult to argue that it works unfairly. In their words, "Secret credit scoring can undermine the public good, since opaque methods of scoring make it difficult for those who feel – and quite possibly are – wronged to press

28. Lehr and Ohm, *supra* note 12.
29. Jessica M. Eaglin, 'Predictive Analytics' Punishment Mismatch', 14 *ISJLP* (2017) p. 87; Jessica Eagling, 'Constructive Recidivism Risk', 67 *Emory Law Journal* (2017) p. 59.
30. Solon Barocas and Andrew D. Selbst, 'Big Data's Disparate Impact', 104 *Calif. L. Rev*. (2016) p. 671.
31. Danielle Keats Citron and Frank Pasquale, 'The Scored Society: Due Process for Automated Predictions', 69 *Wash. L. Rev*. (2014), p. 1.

their case."[32] They also explain that, because the inner workings of the algorithms is a mystery, "consumers cannot determine optimal credit behavior or even what to do to avoid a hit on their scores".[33]

Ryan Calo's 2017 article, 'Artificial Intelligence Policy: A Primer and Roadmap', provides an overview of AI policy.[34] The piece centres on difficult policy questions related to justice, equity, safety, and cybersecurity in light of AI's potential for bias.

In a 2016 article, Andrew Guthrie Ferguson examined how lawyers and courts are using AI-based tools to create pools of jurors, and how this has had a problematic impact on jury selection.[35] First, Ferguson analysed discrimination in the process of selecting juries. Next, he examined the concept of "bright data"—more detailed information—to respond to diversity and discrimination problems in jury pools. Ferguson also looked at how a bright data system might be implemented.

In 2017, Sharona Hoffman concluded that the Americans with Disabilities Act does not provide workers with sufficient anti-discrimination protections in the age of big data and AI.[36] As Hoffman explains, AI has given employers the power to predict who, within a group of healthy employees, is most likely to become sick in the future. Hoffman offers a two-step plan to address this problem: "(1) amend the ADA to prohibit discrimination that is based on the belief of an employer that their employee likely will develop a future mental or physical impairment; and (2) require by law employers to disclose in writing any practices apart from medical exams and direct inquiries used to seek health-related information."

In his 2017 article titled, 'Hope, Hype, and Fear: The Promise and Potential Pitfalls of Artificial Intelligence in Criminal Justice', William S. Isaac looked at the increase in the use of algorithmic decision systems as tools of objective data to overcome inequalities in order for government agencies to better serve underrepresented groups.[37] The author challenged the assumption that data can be objective in the area of criminal justice. Isaac looked at predictive policing and discussed "how machine learning algorithms are unaware and often unable to adjust for institutional biases within policing data". This leads to predictions that reflect human biases within a dataset. So, failing to understand the limitations of these predictive tools and their data may lead to further perpetuating historical discrimination, violating the civil and human rights of underrepresented groups.

In their 2017 article, 'Accountable Algorithms', Joshua A. Kroll et al., presented a technological toolkit that could be used to verify that automated decisions comply with legal fairness standards, challenging the argument that transparency will address issues of accountability of algorithms.[38] First, the authors discussed computer science

32. *Ibid.*, p. 31.
33. *Ibid.*, p. 11. Finally, the authors argue for transparency in scoring systems and call for "technological due process", subjecting the scoring systems to expert review.
34. Richard Calo, 'Artificial Intelligence Policy: A Primer and Roadmap', 51 *U.C. Davis L. Rev*. (2017) p. 399.
35. Andrew Guthrie Ferguson, 'The Big Data Jury', 91 *Notre Dame L. Rev*. (2016) p. 935.
36. Sharona Hoffman, 'Big Data and the Americans with Disabilities Act', 68 *Hastings L.J*. (2017) p. 777.
37. William S. Isaac, 'Hope, Hype, and Fear: The Promise and Potential Pitfalls of Artificial Intelligence in Criminal Justice', 15 *Ohio St. J. Crim. L*. (2018) p. 543.
38. Joshua A. Kroll et al., 'Accountable Algorithms', 165 *U. Pa. L. Rev*. (2017) p. 633.

techniques that they believe can be used "to verify that automated decisions comply with standards of legal fairness". Next, they showed how these techniques can make sure that decisions are made with procedural regularity and how their proposed approach could resolve issues with the diversity visa lottery used by the US State Department. Fundamentally, the authors sought to explore how computational techniques can make sure that the algorithms follow substantive legal and policy choices and can be used to detect and remove discrimination in algorithms, holding accountable automated decision-making processes.

In 2017, Pauline T. Kim, looked at the impact of and appropriate legal response to data analytics on workplace equality.[39] First, Kim analysed how workplace bias arises. She then considered whether AI-based tools or big data models could minimize or eliminate bias, surveying risks and potential harms that may result from reliance on algorithms. Finally, Kim considered whether responses other than anti-discrimination laws may effectively address classification bias, concluding that traditional privacy protection and market forces are not likely to be successful.

In 2016, Allan G. King and Marko J. Mrkonich, also wrote on the theme of workplace equity.[40] This article looks at the biggest risks employers should consider before adopting big data methods for selecting employees, particularly noncompliance with anti-discrimination laws, in particular Title VII and the Americans with Disabilities Act. It looks at issues like causation, correlation, disparate impact and validity of big data methods and criteria.

In 2017, Karen Levy and Solon Barocas looked at how the way users interact on online platforms used for dating, transportation, employment and housing may lead to sociological discrimination.[41] The authors looked at over 50 platforms in seven areas: dating, consumer-to-consumer sales, tasks and gigs, hiring, transportation, housing, and crowdfunding and lending.

Also in 2018, Deirdre K. Mulligan and Daniel S. Griffin, looked at the controversy surrounding Google search engine results for the query, "did the holocaust happen".[42] The authors looked at mismatches fuelling public objections to results and resistance from corporations like Google to change the script producing those results and where the objections and resistance come from. The article concludes that, "the emerging soft law requirement that businesses respect and remedy human rights violations entangled in their business operations provides a normative basis for rescripting search." The final section of the article argues that "the right to truth", now recognized in human rights law, is directly affected by the search scripts used by search engine providers.

In a 2017 article, Andrew D. Selbst looked at the new practice of predictive policing, in which police departments use machine learning systems to predict criminal outcomes based on a set of input data (e.g., the locations of previous crimes

39. Pauline T. Kim, 'Data-Driven Discrimination at Work', 58 *Wm. & Mary L. Rev*. (2017) p. 857.
40. Allan G. King and Marko J. Mrkonich, 'Big Data and the Risk of Employment Discrimination', 68 *Okla. L. Rev*. (2016) p. 555.
41. Karen Levy and Solon Barocas, 'Designing against Discrimination in Online Markets', 32 *Berkeley Tech. L.J.* (2017) p. 1183.
42. Deirdre K. Mulligan and Daniel S, Griffin, 'Rescripting Search to Respect the Right to Truth', 2 *Geo. L. Tech. Rev*. (2018) p. 557.

and consumer data describing potential criminals).[43] One of the main dangers of this practice is reproducing existing discriminatory patterns. First, the author provided a technical perspective overview of predictive policing, looking at how the practice works and how it has impacted minority communities in the US. Next, the author looked at how various standing legal strategies have failed. Finally, the author introduced "algorithmic impact statements", modelled on the National Environmental Policy Act's environmental impact statements. Under this proposed regulation, police would have to consider predicted efficacy of the practice, any disparate impact that may result from the technology, and all reasonable alternatives. These statements would require police to consider problems of discrimination and bias early in the process. The author argued that these impact statements can be used outside of policing.

In a 2017 article, Cary Coglianese and David Lehr examined whether society should be alarmed by the government and its agencies using machine-learning applications.[44] The authors first summarized how modern machine-learning algorithms work and identified existing and potential future applications in the administrative state. They then considered whether the ways technology is implemented could result in systematic discrimination.

Although nearly all of the legal scholarship focusing on this subject has focused on the dangers AI poses for bias, some scholars have started to look at the potential AI holds to reduce human bias. In 2017, Sharad Goel et al., looked at the flip side of the coin: AI's potential to reduce bias. They focused on new strategies made possible by big data to oversee police and improve fairness in law enforcement.[45] The authors sought to call attention to how big data might make the police more accountable to the public and improve police practices; specifically, the authors focused on reducing racial discrimination in the context of "stop-and-search" or Terry stops. Professor Daniel L. Chen, a researcher at the University of Toulouse Faculty of Law, suggested that AI could be used to help judges make decisions that are fairer.[46] At the time of writing, a number of companies are exploring how AI can reduce bias in judicial decisions, hiring choices and in facial recognition systems used by police.[47]

43. Andrew D. Selbst, 'Disparate Impact in Big Data Policing', 52 *Ga. L. Rev*. (2017) p. 109.
44. Cary Coglianese and David Lehr, 'Regulating by Robot: Administrative Decision Making in the Machine-Learning Era', 105 *Geo. L.J*. (2017) p. 1147.
45. Sharad Goel et al., 'Combatting Police Discrimination in the Age of Big Data', 20 *New Crim. L. Rev*. (2017) p. 181.
46. Daniel L. Chen, 'Machine Learning and the Rule of Law', Working Paper available at < https://ssrn.com/abstract = 3302507 > ; < https://www.theverge.com/2019/1/17/18186674/daniel-chen-machine-learning-rule-of-law-economics-psychology-judicial-system-policy > .
47. Sean Captain, 'Can Using Artificial Intelligence Make Hiring Less Biased?', < https://www.fastcompany.com/3059773/we-tested-artificial-intelligence-platforms-to-see-if-theyre-really-less- > ; but see < https://www.forbes.com/sites/nicolemartin1/2018/12/13/are-ai-hiring-programs-eliminating-bias-or-making-it-worse/#70f87f5022b8; https://www.fastcompany.com/3052053/how-artificial-intelligence-is-finding-gender-bias-at-work; https://www.theverge.com/2019/1/30/18202335/ai-artificial-intelligence-recruiting-hiring-hr-bias-prejudice > ; < https://www.engadget.com/2019/01/27/mit-automatically-reduces-racist-biases-in-face-detection/ > .

F. CONCLUSION AND RECOMMENDATIONS

Governments and corporations are delegating critical decisions about our lives to AI. These decisions include where police should go, how criminals should be sentenced, who should be considered for a job, how credit is extended, how disease is assessed and treated, and how juries are assembled. The conceit behind AI is that machines will bring order and predictability to our society. This might be so, but the opposite is also possible. Like a modern-day golem, AI can mirror and amplify the same disordered and biased thinking that we are prone to.

The potential for harm is real and complex. The questions that we ask AI, the forms of the answers that we request, the ways we use these answers, the providence, pedigree, and completeness of the data we provide these systems, and the incredibly complex algorithms at the heart of AI systems all can lead to unfair results. The risk of bias is aggravated by the opacity of AI systems, and the severity of the harm is worsened by the prevalence of feedback loops. The risks of AI only tell half the story, though. At the time of writing, experts are exploring how AI can mitigate human bias that pervades the very same arenas that legal experts are concerned about.

This leads to the question presented at the outset of the chapter: how can society ensure that AI reduces human bias without introducing new forms of computerized discrimination? This question is as much about law and policy as it is about engineering. The existing legal framework – in particular, intellectual property laws and privacy laws and regulations – contribute to some of the sources of AI bias. In addition, new laws and policies may need to be developed to address the many individual policy interests at risk. We cannot afford to ignore the problem. Turing's dream that computers would one day behave like humans has arrived, but not in the way he imagined. As writers predicted long before AI was a reality, machines have no greater claim to objectivity or fairness than humans do. A golem, it seems, is only as good as its creator. Moving forward, we must ensure that these technologies do not come between us, our institutions, and our shared responsibility to treat one another fairly.