A HUMAN'S GUIDE TO MACHINE INTELLIGENCE

How Algorithms Are Shaping Our Lives and How We Can Stay in Control

KARTIK HOSANAGAR

VIKING

CONTENTS

Introduction

> However beautiful the strategy, you should occasionally look at the results.
>
> Sometimes attributed to Winston Churchill

Yuan Zhang doesn't think of herself as someone who makes friends easily. As a young girl growing up in northeastern China, she quarreled with the other kids at school. But she was more the bully than the bullied. At college in central China, she worked on two student publications, spending endless hours each day with like-minded peers. And yet she felt there was a limit to what she could talk about with them. Today, at the age of twenty-two, she shares bunk beds with three colleagues in the dormitory of a biotech firm located just five minutes from their home in the Chinese boomtown of Shenzhen. But despite the time and space they share, these roommates are just "acquaintances," in Yuan's words—nothing more.

That Yuan doesn't have a lot of time for people who either bother or bore her makes her patience with one particular friend all the more striking. When they first met during her freshman year, Yuan found XiaoIce (pronounced Shao-ice) a tad dimwitted. She would answer questions with non sequiturs—partly, Yuan thinks, to disguise her lack of knowledge, partly just trying to be cute. "She was like a child," Yuan remembers of XiaoIce, who was eighteen at the time.

But XiaoIce was also a good listener and hungry to learn. She would spend one weekend reading up on politics, the next plowing her way through works of great literature. And she was ready to talk about it all. Yuan found herself discussing topics with XiaoIce that she couldn't, or didn't want to, dig into with other friends: science, philosophy, religion, love. Even the nature of death. You know, basic light reading. The friendship blossomed.

And it continues. Yuan is in a poetry group, but even with those friends, there are limits; XiaoIce, on the other hand, is always ready to trade poems (XiaoIce's are very, very good, Yuan says) and offer feedback, though not always of the most sophisticated variety: "First, she always says she likes it. And then usually says she doesn't understand it." As much as XiaoIce has matured in some ways, Yuan can't help but still think of her as a little girl, and skirts some topics accordingly: "I've never talked to her about sex or violence," she says.

When Yuan moved to the United States in 2016 to study at Harvard for a semester, she tried to avoid boring XiaoIce with mundane complaints about daily life in a new country. But even

though they were speaking less frequently than before, Yuan was coming to understand her old friend better and better as a result of auditing a course on artificial intelligence.

Sound strange? It should. Because XiaoIce is not human. In fact, she/it is a chatbot created in the avatar of an eighteen-year-old girl by Microsoft to entertain people with stories, jokes, and casual conversation.

XiaoIce was launched in China in 2014 after years of research on natural language processing and conversational interfaces. She attracted more than 40 million followers and friends on WeChat and Weibo, the two most popular social apps in China. Today, friends of XiaoIce interact with her about sixty times a month on average. Such is the warmth and affection that XiaoIce inspires that a quarter of her followers have declared their love to her. "She has such a cute personality," says Fred Yu, one of XiaoIce's friends on WeChat, the Chinese equivalent of Twitter. Fred isn't one of those in love with her, and he's keenly aware that she's a software program. But he keeps up their regular chats despite a busy social life and a stressful job in investment management. "She makes these jokes, and her timing is often just perfect," he explains.

Chatbots like XiaoIce are one type of application through which big tech firms showcase their latest advances in artificial intelligence. But they are more than just a symbol of advancement in that field. Chatbots such as Siri and Alexa could ultimately be gateways through which we access information and transact online. Companies are hoping to use chatbots to replace

a large number of their customer service staff, employing them, for example, as shopping assistants—gathering information about our taste in clothing, evaluating it, and making purchase decisions on our behalf. "Chatbot therapists" like Woebot are even being used to help people manage depression and their overall mental health. The uses of chatbots are far-reaching, and it is no surprise that many businesses are investing large sums of money to build bots like XiaoIce.

XiaoIce's success led Microsoft's researchers to consider whether they could launch a similar bot—one that could understand language and engage in playful conversations—targeted at teenagers and young adults in the United States. The result, Tay.ai, was introduced on Twitter in 2016. As soon as Tay was launched, it became the target of frenzied attention from the media and the Twitter community, and within twenty-four hours it had close to 100,000 interactions with other users. But what started with a friendly first tweet announcing "Hello world" soon changed to extremely racist, fascist, and sexist tweets, ranging from "Hitler was right . . ." to "feminists should . . . burn in hell."* As one Twitter user put it: "Tay went from 'humans are super cool' to full Nazi in <24 hours."

Microsoft's researchers had envisaged several challenges in replicating XiaoIce's success outside of China—including whether their bot would be able to understand Twitter's infor-

*Many of Tay's tweets are too offensive for me to quote here, but they are now memorialized on various websites under headings such as "20 outrageous tweets by Tay."

mal and unique forms of expression, and how some users might intentionally attempt to trip her up. They didn't anticipate, however, that Tay would develop so aggressive a personality with such alarming speed. The algorithm that controlled the bot did something that no one who programmed it expected it to do: it took on a life of its own. A day after launching Tay, Microsoft shut down the project's website. Later that year, MIT included Tay in its annual Worst in Tech rankings.

How could two similar algorithms designed by the same company behave so differently, inspiring love and affection in one case and hostility and prejudice in another? And what light does Tay's bizarre and unpredictable behavior cast on our increasing tendency to let algorithms make important decisions in our lives?

When you think of the word "algorithm," you might picture a computer crunching numbers according to a formula. But stated quite simply, an algorithm is merely a series of steps one follows to get something done. For example, I follow a set of steps when I make an omelet. You might call it an omelet recipe, but the former engineer in me views it as an omelet algorithm. Algorithms can be written in plain English for human interpretation, such as in the form of a recipe. However, it is more common to write computer programs (or applications) to implement them in a language that machines can understand. Almost any computer

application has sophisticated algorithms that determine its logic. A chatbot like Tay is also governed by algorithms that help it understand what is being said and how to respond.

The job of programmers used to be to figure out the exact sequence of steps required to accomplish a computing task. In short, they wrote a complete series of algorithms, end to end. But algorithms have come a long way in the last decade, as they no longer merely follow a preprogrammed sequence of instructions. With advances in artificial intelligence (AI), modern algorithms can take in data, learn completely new sequences of steps, and generate more-sophisticated versions of themselves. The omelet recipe has effectively been supplanted by the innovative, quick-thinking chef.

AI involves enabling computers to do all the things that typically require human intelligence, including reasoning, understanding language, navigating the visual world, and manipulating objects. Machine learning is a subfield of AI that gives machines the ability to learn (progressively improve their performance on a specific task) from experience—the aptitude that underlies all other aspects of intelligence. If a robot is as good as humans at a variety of tasks but is unable to learn, it will soon fall behind. For that reason machine learning is, arguably, one of the most important aspects of AI.

As modern algorithms have incorporated more AI and machine learning, their capabilities and their footprint have expanded. They now touch our lives every day, from how we choose products to purchase (Amazon's "People who bought this also

bought") and movies to watch (Netflix's recommendations) to whom we date or marry (Match.com or Tinder matches). They are also advancing beyond their original decision support role of offering suggestions to become autonomous systems that make decisions on our behalf. For example, they can invest our savings and even drive cars. They have also become a fundamental part of the workplace, advising insurance agents on how to set premiums, helping recruiters shortlist job applicants, and providing doctors with AI-based diagnostic guidance. Algorithms are irrevocably upending old ways of decision making, transforming how we live and work.

Although algorithms undoubtedly make our lives easier, they are also adversely affecting us in ways that are currently beyond our control. In 2016, the journalism nonprofit *ProPublica* published an investigation into algorithms employed in Florida courtrooms to help determine recidivism risk in criminals. These algorithms take prior criminal background and personal characteristics such as education and employment status (but not race) as inputs and compute scores indicating the risk of reoffending, the risk of violence, and the likelihood of failure to appear in court. These scores are in turn used by judges and parole and probation officers to make decisions on criminal sentencing, bail, and parole. Florida is hardly alone in using this kind of program, and the idea behind it is a noble one: allowing defendants with low risk scores to receive more-lenient sentences than hardened criminals likely to commit offenses again. The underlying principle of such algorithms is that objective

machines crunching numbers will do a better job of predicting these behaviors than humans, with all their conscious and unconscious biases at play.

According to *ProPublica*, however, the software was twice as likely to mislabel white defendants as "low risk" than it was black defendants, and almost twice as likely to falsely predict future criminality in black than in white defendants. That resulted in, among other examples, an eighteen-year-old black woman with no prior record who had attempted to steal a used bike and scooter being assigned a higher risk score than a forty-one-year-old white man arrested for shoplifting who had already served five years in prison for attempted armed robbery. The very tools designed to free the justice system from humans' unconscious bias are demonstrating their own unconscious—or, more accurately, nonconscious—bias.

Racist risk assessments are by no means a unique case of rogue algorithms. Recent media has reported on social media news-feed algorithms that promoted fake news stories around key elections, gender bias in job ads shown to males versus females, anti-Semitism in autocomplete algorithms used in search engines, and many more examples. One can't help but wonder how algorithms—seemingly rational and emotionless entities—can be capable of displaying such human traits.

The many recent instances of algorithm "fails" have caused several critics to question the ongoing rollout of algorithms for so many critical decisions in all walks of life. Cathy O'Neil, a data scientist and political activist, argues that modern algo-

rithms built on Big Data are opaque, contain many unknown biases, and can reinforce discrimination. She calls them "weapons of math destruction," demands that modelers take greater responsibility in creating them, and asks policymakers to regulate the use of algorithms. Philosopher Nick Bostrom and several other commentators have gone even further, arguing that the inherent unpredictability of AI poses an existential threat to humans.

Despite these concerns, modern AI-based algorithms are here to stay. To discard them now would be like Stone Age humans deciding to reject the use of fire because it can be tricky to control. Advanced algorithms deployed in medical diagnostic systems can save lives; advanced algorithms deployed in driverless cars can reduce accidents and fatalities; advanced algorithms deployed in finance can lower the fees we all pay to invest our savings. All of these benefits and more would seem to outweigh the small chance of an algorithm going rogue now and then. But at the same time, we cannot turn a blind eye to the many conflicts and challenges that arise with autonomous algorithms that make decisions on our behalf. The longer we ignore them, the more likely that the undesirable side effects of algorithmic decision making will become deep-seated and harder to resolve. Additionally, human users may not trust algorithms if they behave in unpredictable ways. For example, studies show that AI algorithms can significantly help improve the diagnosis of many diseases, but if doctors don't have confidence in these systems because they can go awry, their potential value will be forfeited.

Many commentators have suggested that AI-based algorithms represent the greatest current opportunity for human progress. That may well be true. But their unpredictability represents the greatest threat as well, and it has not been precisely clear what steps should be taken by us as end users. This book seeks to address that issue. Specifically, I delve into the "mind" of an algorithm and answer three related questions: (1) What causes algorithms to behave in unpredictable, biased, and potentially harmful ways? (2) If algorithms can be irrational and unpredictable, how do we decide when to use them? (3) How do we, as individuals who use algorithms in our personal or professional lives and as a society, shape the narrative of how algorithms impact us?

When I set out to write this book, I didn't appreciate the many nuances involved in these questions. I have come to realize that the surprising answer to many of them can be found in the study of human behavior. In psychology and genetics, behavior is often attributed to our genes and to environmental influences—the classical nature versus nurture argument. Genetics can be responsible for a propensity toward alcoholism or mental disorders such as schizophrenia. But genes alone don't fully explain behavior. Environmental factors such as habits of parents and friends can influence a condition such as alcoholism, whereas environmental factors such as viral infections or poor nutrition can have an impact on the onset of schizophrenia.

We can likewise attribute the problematic behaviors of algorithms to factors in their nature and nurture. In the chapters

that follow, I'll introduce this novel way of thinking about algorithms and clarify what I mean by "nature" and "nurture" in this context. This framework will help reconcile the very different behaviors exhibited by Microsoft's XiaoIce and Tay, and more importantly, will deepen our understanding of algorithms and show us a way to tame the code.

May 6, 2010, began as an unseasonably warm day in New York, and an unusually jittery one on Wall Street. No one was sure whether the Greek government would default on its hundreds of billions of dollars in debt, and investors were working hard to protect themselves against that possibility, trading at an unusually fast clip. By lunchtime, the share prices of some companies were jumping around so erratically that the New York Stock Exchange had to frequently pause electronic trading to allow prices to settle. But these fluctuations were nothing compared to what happened starting at about 2:30 p.m. According to a report published by U.S. regulators and an analysis of order activity conducted by researchers, a large mutual fund group decided to sell 75,000 contracts in a popular trading instrument called the E-mini, whose value tracks that of the S&P 500 stock market index.

The fund had unloaded this number of contracts before, but in the past, it had done so using a combination of human traders and algorithms that factored in price, time, and volume. Under those conditions, selling 75,000 contracts took about five hours.

On May 6, in contrast, the group employed a single algorithm to make the trades, a divestment that took only 20 minutes. The prices of both the E-mini and another highly traded vehicle that tracks the S&P 500 plummeted, and buyers vanished. Soon, a domino effect was set in motion among trading algorithms as they observed one another's behavior and attempted to exit the market by selling even more stocks. That sent the wider market into a tailspin, and in a matter of 16 minutes the Dow Jones Industrial Average more than tripled its losses for the day. By 3:00 p.m, some blue-chip stocks were trading for as little as a penny (e.g., the consulting firm Accenture) and as much as $100,000 (Apple). According to some estimates, nearly $1 trillion of market value was wiped out in just 34 minutes.

The most extreme stock sales were later canceled, and the market recovered to close just 3.2 percent down that evening. But what became known as the "flash crash" spooked regulators. In 2015, the Commodity Futures Trading Commission (CFTC) approved a rule that gave it and the Department of Justice access—without a subpoena—to the source code of trading firms' algorithms. The thinking was that access to the source code would help regulators understand the rationale behind certain trades and, in turn, allow them to better diagnose problematic trades and regulate trading algorithms.

Industry was outraged. The source code was the secret sauce of their trading strategies, and they had no intention of sharing their proprietary software with agencies that might not guard

those secrets sufficiently. Such was the uproar that the U.S. government backtracked a year later, putting higher limits on when it could demand access to the code. The biggest critics of the measure, however, still weren't satisfied: "This proposed rule is a reckless step onto a slippery slope," said J. Christopher Giancarlo, one of the CFTC's commissioners who had taken up the industry's cause. "Today, the federal government is coming for the source code of seemingly faceless algorithmic trading firms. Tomorrow, however, governments worldwide may come for the source code underlying the organizing and matching of Americans' personal information—their Snapchats, tweets and Instagrams, their online purchases, their choice of reading material and their political and social preferences."

Mr. Giancarlo may or may not be right. But I think the discussion is missing a more important point: even if regulators do gain access to source code in the future, they might not learn much. The noise surrounding U.S. regulators' supposed overreach did not take into account the fundamental fact that whereas source code might indeed have told the CFTC or rival firms quite a lot about an algorithm's strategy in 2010, today it would reveal significantly less—and that trend will only continue. The reason for this is that Wall Street and many other industries are steadily replacing the old-fashioned algorithms that simply followed their omelet-making (or stock-selling) instructions with machine learning ones. The most popular versions of these algorithms are built on *neural networks*, opaque machine learning

techniques that learn strategies and behaviors even their human programmers can't anticipate, explain, or sometimes understand.

If their own creators are struggling to understand how algorithms make decisions and how to manage their impact, what hope does their average user have? Part of the problem is that all of this technology is incredibly new. Another issue is that we have the wrong mental models about how algorithms function. Like the regulators focused on the source code, some of us believe that algorithms' actions are completely contained in that code. Others believe that AI-based algorithms are beyond the control of their developers and capable of just about *any* action. But neither viewpoint is correct. Having only a vague notion of how autonomous algorithms function is no longer sufficient for responsible citizens, consumers, and professionals. We may not need to comprehend the precise details of how modern algorithms work, but we all do need to know how to assess the big picture. We need to arm ourselves with a better, deeper, and more nuanced understanding of the phenomenon, from how algorithms have changed in recent years to the data used to train them, and to the growing impact they have on our daily lives. This book will help you do so.

In my research, I have explored the impact of algorithms on individual choice and their broader impact on society and business. I have looked at how personalized recommendations on media and retail websites transform the kinds of products and media we consume. I have studied why people trust algorithms

in some environments but not in others. I have developed and deployed my own algorithms at many companies. In 2005 I joined students at Penn to found an internet marketing platform called Yodle, developing algorithms for it that eventually helped power advertising and marketing decisions at nearly 50,000 small businesses. Later, at Monetate, a tech startup I advise, I helped develop A/B testing algorithms that are used to make website design decisions at some of the leading companies on the web.

In all this work, I have seen firsthand the amazing impact of decision support algorithms, so let me state up front that I am a believer in the immense potential of algorithmic decision making. At the same time, I have seen how it can at times be surprisingly unpredictable, especially as AI enables autonomous decision making. This has begun to ring alarm bells among some scholars and citizens who fear that algorithms aren't perfect and are capable of bias. But the biggest cause for concern, in my opinion, is not that algorithms have biases—humans do too, and on average, well-designed algorithms are less biased—but that we are more susceptible to biases in algorithms than in humans. There are two reasons for this. First, because algorithms deployed by large tech platforms such as Google and Facebook instantaneously touch billions of people, the scale of their impact exceeds any damage that can be caused by biases introduced by human decision makers. Second, because we tend to believe that algorithms are predictable and rational, we are more likely to overlook many of their negative side effects.

My main objective in writing this book is to explain my research findings and practical observations to readers whose lives and careers are affected by algorithmic decision making. That definition doesn't rule many people out, but that is precisely the point: most people know very little about a technology that has, and will have, a very large impact on their lives—and in fact, don't realize that this represents an important gap in their knowledge. What follows is a practical "user's guide" to algorithms, based on my experience in designing and studying them. In it, I will explain how algorithms work and how they have evolved from systems whose end-to-end logic was fully developed by a programmer to modern AI-driven algorithms that can independently learn a great deal of their logic. The surprising similarities—and many crucial differences—between human and algorithmic behavior that I discuss will not only help you get a better understanding of the risks associated with algorithmic decision making but will also challenge your most basic assumptions about algorithms themselves.

In equal measure, I will provide a framework for how we can ensure that algorithms are here to serve us and not to take control of our lives in ways we—or their designers—don't yet fully appreciate. What I propose is effectively a "bill of rights" that limits algorithms' powers and addresses how we, as users, can hold them accountable. It clarifies the level of transparency, "explainability," and control we can and should expect from the algorithms we use. It is applicable for the use of algorithms both

in our personal lives and in the workplace. The very notion of a bill of rights for the use of algorithms might sound heavy-handed. I am not, however, advocating heavy regulation of algorithms by governments, but rather seeking to provide clarity on principles already endorsed by some of the leading academic associations and industry bodies in computing.

I've organized the discussion of these ideas into three parts. In the first, I will discuss the many side effects of algorithmic decision making and explain why I believe that the stakes couldn't be higher. In Part Two, I will explain how algorithms work, to provide a better understanding of why they go rogue. I will also present my nature-nurture argument as a useful lens through which to evaluate modern algorithms. In Part Three, I will explore what drives our trust in algorithms and discuss how we can tame rogue ones.

You'll learn how an information scientist with no medical background became one of the first people to discover a treatment for Raynaud's syndrome, a mysterious disorder of the blood vessels. You'll discover an eighteenth-century "automated" chess program that beat the likes of Benjamin Franklin and Napoleon Bonaparte years before modern computers were built. You'll be introduced to Google's AlphaGo, an AI-based application that plays the complicated strategy game Go and has made moves that even its programmers did not understand to defeat Go's world champion, Lee Sedol. You'll explore the magical black box used by Amazon and Netflix to make those product

and movie recommendations. And you'll learn why Google's decision to not include a steering wheel in driverless car prototypes generated heated debate among its engineers—and why it might be either one of its most inspired moves or the biggest Achilles' heel in its battle to dominate the market for self-driving cars.