Understand, Manage, and Prevent Algorithmic Bias

A Guide for Business Users and Data Scientists

Tobias Baer

Apress®

UNDERSTAND, MANAGE, AND PREVENT ALGORITHMIC BIAS

A GUIDE FOR BUSINESS USERS AND DATA SCIENTISTS

Tobias Baer

Apress®

Understand, Manage, and Prevent Algorithmic Bias: A Guide for Business Users and Data Scientists

Tobias Baer
Kaufbeuren, Germany

ISBN-13 (pbk): 978-1-4842-4884-3 ISBN-13 (electronic): 978-1-4842-4885-0
https://doi.org/10.1007/978-1-4842-4885-0

Managing Director, Apress Media LLC: Welmoed Spahr
Acquisitions Editor: Shiva Ramachandran
Development Editor: Laura Berendson
Coordinating Editor: Rita Fernando

Cover designed by eStudioCalamar

Distributed to the book trade worldwide by Springer Science+Business Media New York, 233 Spring Street, 6th Floor, New York, NY 10013. Phone 1-800-SPRINGER, fax (201) 348-4505, e-mail orders-ny@springer-sbm.com, or visit www.springeronline.com. Apress Media, LLC is a California LLC and the sole member (owner) is Springer Science + Business Media Finance Inc (SSBM Finance Inc). SSBM Finance Inc is a **Delaware** corporation.

For information on translations, please e-mail rights@apress.com, or visit www.apress.com/rights-permissions.

Apress titles may be purchased in bulk for academic, corporate, or promotional use. eBook versions and licenses are also available for most titles. For more information, reference our Print and eBook Bulk Sales web page at www.apress.com/bulk-sales.

Any source code or other supplementary material referenced by the author in this book is available to readers on GitHub via the book's product page, located at www.apress.com/9781484248843. For more detailed information, please visit www.apress.com/source-code.

Printed on acid-free paper

For the love algorithm in my partner's mind—
I still haven't figured out what bias caused him
to choose me but I think it's the best mistake
he has made in his life!

Contents

About the Author vii
Acknowledgments ix
Preface xi

Part I: An Introduction to Biases and Algorithms 1
Chapter 1: Introduction 3
Chapter 2: Bias in Human Decision-Making 9
Chapter 3: How Algorithms Debias Decisions 21
Chapter 4: The Model Development Process 29
Chapter 5: Machine Learning in a Nutshell 41

Part II: Where Does Algorithmic Bias Come From? . . . 51
Chapter 6: How Real-World Biases Are Mirrored by Algorithms 53
Chapter 7: Data Scientists' Biases 59
Chapter 8: How Data Can Introduce Biases 69
Chapter 9: The Stability Bias of Algorithms 79
Chapter 10: Biases Introduced by the Algorithm Itself 87
Chapter 11: Algorithmic Biases and Social Media 95

Part III: What to Do About Algorithmic Bias from a User Perspective 107
Chapter 12: Options for Decision-Making 109
Chapter 13: Assessing the Risk of Algorithmic Bias 117
Chapter 14: How to Use Algorithms Safely 123
Chapter 15: How to Detect Algorithmic Biases 129
Chapter 16: Managerial Strategies for Correcting Algorithmic Bias . . 161
Chapter 17: How to Generate Unbiased Data 167

Part IV: What to Do About Algorithmic Bias from a Data Scientist's Perspective 173

Chapter 18: The Data Scientist's Role in Overcoming Algorithmic Bias 175

Chapter 19: An X-Ray Exam of Your Data 193

Chapter 20: When to Use Machine Learning. 209

Chapter 21: How to Marry Machine Learning with Traditional Methods. 215

Chapter 22: How to Prevent Bias in Self-Improving Models 223

Chapter 23: How to Institutionalize Debiasing 233

Index 241

About the Author

Tobias Baer is a data scientist, psychologist, and top management consultant with over 20 years of experience in risk analytics. Until June 2018, he was Master Expert and Partner at McKinsey & Co., Inc., where he built McKinsey's Risk Advanced Analytics Center of Competence in India in 2004, led the Credit Risk Advanced Analytics Service Line globally, and served clients in over 50 countries on topics such as the development of analytical decision models for credit underwriting, insurance pricing, and tax enforcement, as well as debiasing decisions. Tobias has been pursuing a research agenda around analytics and decision making both at McKinsey (e.g., on debiasing judgmental decisions and on leveraging machine learning to develop highly transparent predictive models) and at University of Cambridge, UK (e.g., the effect of mental fatigue on decision bias).

Tobias holds a PhD in Finance from University of Frankfurt, an MPhil in Psychology from University of Cambridge, an MA in Economics from UWM, and has done undergraduate studies in Business Administration and Law at University of Giessen. He started publishing as a teenager, writing about programming tricks for the Commodore C64 home computer in a German software magazine, and now blogs regularly on his LinkedIn page, www.linkedin.com/in/tobiasbaer/.

Acknowledgments

First of all, I want to thank my publisher, Shiva Ramachandran, who deserves sole credit for coming up with the brilliant idea of writing this book, and my editor, Rita Fernando, who not only unleashed a writing beast in myself through her relentless encouragement but also kept the red ink away from my quirky humor. She's to blame for everything—I had expected a lot more adult supervision!

I also want to thank Professor (now emeritus) Paul Shaman from the Statistics Department of The Wharton School of University of Pennsylvania with whom I had the privilege to spend two precious months in 1999 as a Visiting Scholar. He opened my eyes to the difference between running a script to estimate a model and understanding data—much of my critical attitude towards data originates in his teachings.

I finally would like to thank Clemens Baader, who graciously read the draft manuscript and was always a fabulous sounding board for my ideas.

Preface

Why did I write this book? Much has been written about algorithmic biases already; disturbing examples of algorithmic biases abound. Much less has been written about the actual causes of algorithmic biases, however, and very little seems to be known about how to solve the problem and either prevent algorithmic biases altogether or manage them in a way that prevents harm. This is what this book is about.

This book is practical. It suggests solutions that you can start implementing tomorrow. Some of the actions may take some time to reach completion or fruition—but this book is not about fancy theory. There are step-by-step guides and check-lists, as well as countless real-life examples to illustrate my points. Most importantly, though, this book encourages critical thinking by suggesting which specific questions to ask.

The more I discovered about algorithmic biases in my own modeling and consulting work, the more I realized that it is much more than a technical issue. Yes, statistics accounts for both some of the root causes of algorithmic biases and some of the solutions. However, the issue is deeply rooted in human psychology, and we cannot address algorithmic biases without understanding human biases and how the biases of users, data scientists, and society at large create and proliferate decision biases.

Therefore I do not jump right into technical solutions but take the time to explain where algorithmic biases come from—and what this means for fighting them.

And the (non-technical) users of algorithms—such as business managers and public servants—have a lot more power to fight and prevent algorithmic biases than they might believe. This book wants to empower everyone to better deal with algorithmic bias and join hands to prevent it.

Who This Book Is For

We live in a world where all of us are affected by algorithms and many of us use them, maybe even unaware that an algorithm is involved. Therefore I have written this book for all of us.

Data scientists are the scarce experts who develop algorithms and therefore have a big role to play in dealing with and preventing algorithmic bias, and I

therefore hope that many, maybe all of them, will read this book—and the last, most technical part of this book is even dedicated to them.

However, most people are not data scientists, and many outright hate statistics. The book therefore is written with a lay(wo)man in mind. It uses non-technical language, vivid analogies, and tries to keep the fun quotient up through excessive use of humor. Warning: You might even like statistics in the end, at least the biased image of it projected by this book…

As the issue of algorithmic bias has come to the fore, naturally also compliance officials and regulators have started to explore it and search for ways to prevent harm from it. Therefore, this book is not just for the actual developers and users of algorithms—and the business managers and public servants who decide where and how to use them—but also for compliance officials and regulators tasked with keeping decision processes in check.

And, as I will argue, many algorithmic biases are a mirror of deep-rooted societal biases. Therefore the issue of algorithmic biases is a much larger one, and I have written this book also for politicians, journalists, and philosophers who need to know that algorithms can be as much a solution for fighting societal biases as they can be a problem if they perpetuate and amplify such biases.

Last but not least, the book is for Martians and Zeta Reticulans. You'll figure out why soon!

What This Book Is Not

This book is not a statistics textbook. It will reference countless statistical techniques for the data scientists (and interested laymen) among the readers—but it will not explain them. Data scientists know most of the techniques already or at least know where to look them up.

This book is also no legal textbook. It does treat legal and ethical issues on a philosophical level—including how the European General Data Protection Regulation both recognizes and misses some core insights about algorithms—but it does not aim to catalogue all the laws somehow relevant in dealing with algorithmic bias or to give guidance on how to comply with specific legal requirements. This requires lawyers—ideally ones who have read this book, too.

Finally, this book is no silver bullet. Fighting biases is hard. In a sense, biases are a form of conformity—conformity with "the way it is," what your boss says, what the data says, what your lazy mind says (because you always have done it this way). There is zero chance that you will get any benefit from reading this book if you don't change some of the things you are doing. I invite you to keep thinking about what the insights from this book mean to you and what you can do differently because of what you have learned. In fact, I would

love to hear it—why don't you leave a comment on my blog at www.linkedin.com/in/tobiasbaer/? And if you want to prevent all your good ideas slipping away and going back to your old ways once you have read this book and found a good space for it on your bookshelf where it can collect plenty of allergenic dust, maybe you even want to put a reminder in your calendar right now!

How This Book Is Organized

The book has four parts. The first part is an introduction—it covers the psychology of human biases as well as how algorithms are used and developed. The chapters of the first part explain the terminology and frameworks I will keep referring to in the remainder of the book.

The second part introduces six distinct sources of algorithms. The understanding of these sources is the basis for managing and preventing algorithmic bias and therefore will be referenced throughout the remainder of the book.

The third part discusses how users of algorithms (broadly defined as anyone who is not a data scientist) can deal with algorithmic bias and what powerful possibilities they have to prevent it.

And the fourth and final part provides comprehensive, practical guidance to data scientists for preventing algorithmic biases through specific techniques for development and implementation. This part of the book is therefore the most technical one—but I still wrote it in a way that everyone from an undergraduate student to a seasoned Head of Analytics can follow and find some valuable insights.

PART

I

An Introduction to Biases and Algorithms

CHAPTER

1

Introduction

What is a bias? A widely cited source[1] defines it as follows:

> *Inclination or prejudice for or against one person or group, especially in a way considered to be unfair.*

Biases are double-edged swords. As you will see in the next chapter, biases typically are not a character flaw or rare aberration but rather the necessary cost of enabling the human mind to make thousands of decisions every day in a seemingly effortless, ultra-fast manner. Have you ever marveled how you were able to escape a fast-moving object, such as a car about to crash into you, in a split-second? Neuroscientists and psychologists have started to unravel the mysteries of the mind and have found that the brain can achieve this speed only by taking numerous shortcuts.

A shortcut means that the mind will jump to a conclusion (e.g., deem a dish inedible or a stranger dangerous) without giving all facts due consideration. In other words, the mind uses prejudice in order to gain speed.

The use of prejudice in decision-making therefore is unfair insofar as it (willfully) disregards certain facts that may advocate a different decision. For example, if your partner once ate a bouillabaisse fish soup and became terribly sick afterwards, he or she is bound to never eat bouillabaisse again, and may refuse to even try the beautiful bouillabaisse you just cooked, blissfully ignoring the fact that you graduated with distinction from cooking school and bought the best and freshest ingredients available in the country.

[1]David Marshall, "Recognizing your unconscious bias," *Business Matters*, www.bmmagazine.co.uk/in-business/recognising-unconscious-bias/, October 22, 2013.

T. Baer, *Understand, Manage, and Prevent Algorithmic Bias,*
https://doi.org/10.1007/978-1-4842-4885-0_1

Algorithms are mathematical equations or other logical rules to solve a specific problem—for example, to decide on a binary question (yes/no) or to estimate an unknown number. Just like the brain making decisions in split-seconds, algorithms promise to give an answer instantaneously (in most cases, the score value of the algorithm's equation can be calculated in a fraction of a second), and they are also a shortcut because they consider only a limited number of factors in a predetermined fashion.

On one level, algorithms are a way for machines to emulate or replace human decision-makers. For example, a bank that needs to approve thousands of loan applications every month may turn to an algorithm applied by a computer instead of human credit officers to underwrite these loans; this often is motivated by an algorithm being both faster and cheaper than a human being.

On another level, however, algorithms also can be a way to reduce or even eliminate bias. Statisticians have developed techniques to develop algorithms specifically under the constraint of being unbiased—for example, the ordinary least squares (OLS) regression is a statistical technique defined as BLUE, the best linear unbiased estimate. Sadly, I had to write that algorithms "*can*" reduce or eliminate bias—algorithms also can be as biased or even worse than human decision-making. Several chapters of this book are dedicated to explaining the many ways an algorithm can be biased.

In the context of algorithms, however, the definition of *bias* should be more specific. Problems solved by algorithms have at least theoretically a correct answer. For example, if I estimate the number of hairs on the head of a well-known president, nobody may ever have counted them, but anyone with unlimited time and access to the president could verify my estimate of 107,817 hairs.

In most situations (including presidential hair), the correct answer cannot be known at least *a priori* (i.e., at the time the algorithm is applied). Algorithms therefore often are a way to make predictions. Through predictions, algorithms help to reduce and to manage uncertainty. For example, if I apply for a loan, the bank doesn't know (yet) whether I will pay back the loan, but if an algorithm tells the bank that the probability of me defaulting on the loan is 5%, the bank can decide whether it will make any profit on me if it gives me the loan at a 5.99% interest rate by comparing the expected loss with the interest charged and other costs incurred by the bank. This illustrates a typical way algorithms are used: algorithms estimate probabilities of specific events (e.g., a customer defaulting on a loan, a car being damaged in an accident, or a person dying by the end of the term of a life insurance contract), and these probabilities allow a business underwriting risks to make an approve/reject decision based on an objective expected risk-adjusted return criterion.

Algorithms are deployed in situations with imperfect information (e.g., the bank's credit rating algorithm doesn't know about the gambling debt I incurred

last night, nor does it know if my company will fire me next month). Algorithms therefore *will* make mistakes; however, they are supposed to be correct *on average*. A **bias** is present if the average of all predictions systematically deviates from the correct answer. For example, if the bank's algorithm assigns a 5% probability of default to 10,000 different customers, one would expect that 500 of the 10,000 will default (500/10,000 = 5%). If you investigate the situation and find that in reality 10% of customers default but every time an applicant has a German passport, the algorithm cuts the true estimate by half, the algorithm is biased—in this case, in favor of Germans. (Is it a coincidence that this algorithm was created by a German guy?)

Systematic errors in predictions—whether made by humans or by algorithms—can have serious implications for businesses, and sadly they happen all the time. For example, one study of mega infrastructure projects—analyzing 258 projects in 20 different countries—found cost overruns in almost 9 out of 10 of them, indicative of a systematic underestimation of true cost.[2] During the global financial crisis, banks such as Northern Rock, Lehman Brothers, and Washington Mutual went under because they had systematically underestimated credit, market, and liquidity risks.

Sometimes human bias is to blame. For example, one US bank had an economic capital model (a sophisticated model quantifying those "unexpected losses" of a given portfolio that can cause a bank run or bankruptcy) that prior to the global financial crisis hinted at the out-sized risks looming in home equity loans by estimating unexpected losses many times larger than expected losses; tragically, management dismissed those estimates because they were used to seeing unexpected losses much closer to expected losses and therefore deemed the model to be faulty.

At other times, however, algorithms themselves are flawed. For example, an Asian bank bought a scoring model for consumer credit cards that looked at the card's utilization ratio as one of the predictors of default. The algorithms believed that customers with a low utilization (e.g., using just 10% of the credit limit) were safer than customers with a high utilization; for safe customers, the algorithm increased the limit. However, this created a circular reference: in the moment the algorithm increased the credit limit, the utilization (calculated by dividing the current outstanding balance by the credit limit) dropped, causing the algorithm to further increase the limit (so if the outstanding was 10 and the limit was 100, utilization was 10%; if the system increased the limit by 25% from 100 to 125, utilization dropped to 8% (= 10/125), triggering another increase of the limit, and so on). This happened until credit limits reached stratospheric levels that were totally beyond the customers' means to repay the bank. When more and more customers

[2]B. Flyvbjerg, M.S. Holm, and S. Buhl, "Underestimating costs in public works projects: Error or lie?," *Journal of the American Planning Association*, 68(3), 279-295, 2002.

started to actually use their very large credit limits, unsurprisingly many defaulted, and the bank almost went bankrupt after having written off more than a billion USD in bad debt.

Algorithmic bias comes in all kinds of shapes and colors. In 2016, ProPublica published a research report showing that COMPAS, an algorithm used by US authorities to estimate the probability of a criminal to re-offend, is racially biased against blacks.[3] MIT reported on natural language processing algorithms being sexist by associating homemakers with women and programmers with men.[4] And research conducted in 2014 showed that setting the user's profile to female in Google's Ad Settings can lead to less high-paying job offers appearing in ads.[5] As more and more decisions are made by algorithms—affecting consumers, companies, employees, governments, the environment, even pets and inanimate objects—the dangers and impact of algorithmic bias is growing day by day. However, this is not by necessity—bias is merely a side-effect of an algorithm's working and therefore a by-product of conscious and unconscious choices made by the creators and users of algorithms. These choices can be revisited and changed in order to reduce or even eliminate algorithmic bias.

This book is about algorithmic bias. First of all, we want to understand better what it is—where it comes from and how it can wreak havoc with important decisions. Second, we want to control its damage by exploring how you can manage algorithmic bias—be it as a user or as a regulator. And third, we want to explore ways for data scientists to prevent algorithmic bias.

The first part, Chapters 2-5, introduces the topic. I will start with a quick review of psychology and human decision biases as algorithmic biases mirror them in more ways than easily meets the eye (Chapter 2) and discuss how algorithms can help to remove such biases from decisions (Chapter 3). Keeping in mind that many readers of this book are laymen and not data scientists, I'll then review how the sausage is made—i.e., how algorithms are developed (Chapter 4) and demystify what is behind machine learning (Chapter 5).

The second part of the book, Chapters 6-11, explores where algorithmic biases come from. Chapter 6 examines how real-world biases can be mirrored by algorithms (rather than rectified). Chapter 7 turns to the persona of the data scientist and how the data scientist's own (human) biases can cause algorithmic biases. Chapter 8 dives deeper into the role of data, and Chapter 9 reviews how the very nature of algorithms introduces so-called stability

[3] J. Larson, S. Mattu, L. Kirchner, and J. Angwin, "How we analyzed the COMPAS recidivism algorithm," *ProPublica*, 9, 2016.
[4] W. Knight, "How to Fix Silicon Valley's Sexist Algorithms," *MIT Technology Review*, November 23, 2016.
[5] A. Datta, M.C. Tschantz, and A. Datta, "Automated experiments on ad privacy settings," *Proceedings on Privacy Enhancing Technologies*, 92-112., 2015.

biases. Chapter 10 looks at new biases arising from statistical artifacts, and Chapter 11 deep-dives into social media where human behavior and algorithmic bias can reinforce each other in a particularly diabolical manner.

The third part of the book, Chapters 12-17, approaches algorithmic bias from a user's perspective. It sets out with a brief discussion of whether or not to actually use an algorithm (Chapter 12) and how to assess the severity of the risk of algorithmic bias for a particular decision problem (Chapter 13). Chapter 14 gives an overview of techniques to protect yourself from algorithmic bias. Chapter 15 more specifically describes techniques for diagnosing algorithmic bias, and Chapter 16 discusses managerial strategies for overcoming a bias ingrained in an algorithm (if not real life). Chapter 17 discusses how users of algorithms can make a critical contribution to the debiasing of algorithms by producing unbiased data.

The fourth part of the book, Chapters 18-23, addresses data scientists developing algorithms. Chapter 18 provides an overview of the various ways data scientists can guard against algorithmic bias. Chapter 19 deep-dives into specific techniques to identify biased data. Chapter 20 discusses how to choose between machine learning and other statistical techniques in developing an algorithm in order to minimize algorithmic bias, and Chapter 21 builds on this by proposing hybrid approaches combining the best of both worlds. Chapter 22 discusses how to adapt the debiasing techniques introduced by this book for the case of self-improving machine learning models that require validation "on the fly." And Chapter 23 takes the perspective of a large organization developing numerous algorithms and describes how to embed the best practices for preventing algorithmic bias in a robust model development and deployment process at the institutional level.

CHAPTER

2

Bias in Human Decision-Making

As you will see in the following chapters, algorithmic biases originate in or mirror human cognitive biases in many ways. The best way to start understanding algorithmic biases is therefore to understand human biases. And while colloquially "bias" is often deemed to be a bad thing that considerate, well-meaning people would eschew, it actually is central to the way the human brain works. The reason is that nature needs to solve for three competing objectives simultaneously: accuracy, speed, and (energy) efficiency.

Accuracy is an obvious objective. If you are out hunting for prey but a poorly functioning cognitive system makes you see an animal in every second tree trunk or rock you encounter, you obviously would struggle to hunt down anything edible.

Speed, by contrast, is often overlooked. Survival in the wild often is a matter of milliseconds. If a tiger appears in your field of vision, it takes at least 200 milliseconds until your frontal lobe—the place of logical thinking—recognizes that you are staring at a tiger. At that time, the tiger very well may be leaping at you, and soon after you'll have ended your life as the tiger's breakfast. Our survival as a species may well have hinged on the fact

T. Baer, *Understand, Manage, and Prevent Algorithmic Bias,*
https://doi.org/10.1007/978-1-4842-4885-0_2

that nature managed to bring down the time for the flight-or-fight reflex to kick in to 30-40 milliseconds—a mere 160 milliseconds difference between extinction and by some accounts becoming the crown of the creation! As John Coates describes in great detail in his book *The Hour Between Dog and Wolf*,[1] nature had to go through a mindboggling array of tweaks and tricks to accomplish this. A key aspect of the solution: if in doubt, assume you're seeing a tiger. As you will see, biases are therefore a critical item in nature's toolbox to accelerate decisions.

Efficiency is the least known aspect of nature's approach to thinking and decision-making. Chances are that you grew up believing that logical, conscious thinking is all your brain does. If you only knew! Most thinking is actually done subconsciously. Even what feels like conscious thinking often is a back-and-forth between conscious and subconscious thinking. For example, imagine you want to go out for dinner tonight. Which restaurant would you choose? Please pause here and actually do make a choice! Ready? Have you made your choice? OK. Was it a conscious or subconscious choice? You probably looked at a couple of options and then consciously made a choice. However, how did that short list of options you considered come about? Did you create a spreadsheet to meticulously go through the dozens or thousands of restaurants that exist in your city, assess them based on carefully chosen criteria, and then make a decision? Or did you magically think of a rather short selection of restaurants? That's an example of your subconscious giving a hand to your conscious thinking—it made the job of deciding on a dinner place a lot easier by reducing the choices to a rather short list.

The reason why nature is so obsessed with efficiency is that your logical, conscious thinking is terribly inefficient. The average brain accounts for less than 2% of a person's weight, yet it consumes 20% of the body's energy.[2] That means 20% of the food you obtain and digest goes to powering your brain alone! That's a lot of energy for such a small part of the body. And most of that energy is consumed by the logical thinking you engage in (as opposed to almost effortless subconscious pattern recognition). Just as modern planes and ships have all kinds of technological methods to reduce energy consumption, Mother Nature also embedded all kind of mechanisms into the brain to minimize energy consumption by logical thinking (lest you need to eat 20 steaks per day). Not surprisingly, it introduced all kind of biases through this.

If you collect all the various biases described across the psychological literature, you will find over 100 of them.[3] Many of them are specific realizations of more fundamental principles of how the brain works, however, and therefore several authors have brought down the literature to 4–5 major types of biases.

---

[1]John Coates, *The Hour Between Dog and Wolf*, New York: The Penguin Press, 2012.
[2]Daniel Drubach, *The Brain Explained*. New Jersey: Prentice-Hall, 2000.
[3]Buster Benson, "Cognitive Bias Cheat Sheet," https://betterhumans.coach.me/cognitive-bias-cheat-sheet-55a472476b18, September 1, 2016.

I personally like the framework developed by Dan Lovallo and my former colleague Olivier Sibony:[4] they distinguish action-oriented, stability, pattern-recognition, interest, and social biases. I will loosely follow that framework when in the following I discuss some of the most important biases required for an understanding of algorithmic bias.

Action-Oriented Biases

Action-oriented biases reflect nature's insight that speed is often king. Who do you think is more likely to survive in the wild, the careful planner who will compose a 20-page risk assessment and think through at least five different response options before deciding whether fight or flight would be a better response to the tiger that just appeared five meters in front of him, or the dare-devil that in a split-second decides to fight the tiger?

A couple of biases illustrate the nature of action-oriented biases. To begin with, biases such as the von Restorff effect (focus on the one item that stands out from the other items in front of us) and the bizarreness effect (focus on the item that is most different from what we expect to see) draw our attention to the yellow fur among all those bushes and trees around us; overoptimism and overconfidence then douse the self-doubt that might cause deadly procrastination.

The bizarreness effect can bias our cognition like outliers and leverage points can have an outsized effect in estimating the coefficients of an algorithm. This is because of the availability bias—if we recall one particular data point more easily than other data points (e.g., because it stood out from most other data points), we overestimate the representativeness of the easy-to-remember data point. This can explain why, say, a single incident of a foreigner conducting a spectacular crime can severely bias our perception of people with that foreigner's nationality, causing out-of-proportion hostility and aggression against them.

Overconfidence deserves our special attention because it also goes a long way to explain why not enough is done about biases in general and algorithmic biases in particular. Many researchers have demonstrated overconfidence by asking people how they compare themselves to others.[5] For example, 70% of high school seniors surveyed believed that they have "above average" leadership skills but only 2% believed they were "below average" (where by definition, roughly 50% each should be below and above average, respectively). On

[4]D. Lovallo and O. Sibony, "The case for behavioral strategy," *McKinsey Quarterly*, 2(1), 30-43, 2010.

[5]The examples here are taken from and further referenced in D. Dunning, C. Heath, and J.M. Suls, "Flawed self-assessment: Implications for health, education, and the workplace," *Psychological science in the public interest*, 5(3), 69-106, 2010.

their ability to get along with others, 60% even believed to be in the top 10% and 25% in the top 1%. Similar results have been found for technical skills such as driving and software programming. Overoptimism is essentially the same bias but applied to the assessment of outcomes and events, such as whether a large construction project will be able to remain within its cost budget.

What does this mean for fighting bias? Even if people accept the fact that others may be biased, they overestimate their own ability to withstand biases when judging—and as a result resist efforts to debias their own decisions. With most people succumbing to overoptimism, we can easily have a situation where most people accept that biases exist but still the majority refuses to do anything about it.

Another fascinating aspect of the research of overoptimism: it has been found in Western culture but not in the Far East.[6] This illustrates that both individual personality and the overall culture of a country (or company/organization) will have an impact on the way we make decisions and thus on biases. A bias we observe in one context may not occur in another—but other biases might arise instead.

■ **Note** An excellent demonstration of overconfidence is the fact that I observe that because of overconfidence, most people fail to take action to debias their decisions—but I write a book on debiasing algorithms anyhow, somehow believing that against all odds I will be able to overcome human bias among my readers and compel them to implement my suggestions. However, I also know that *you*, my dear reader, *are* different from the average reader and a lot more prone to actually take actions than others; therefore, let me just point out that in order to be consistent with your well-deserved positive self-image, you should make an action plan today of how you will apply the insights and recommendations from this book in your daily work and actively resist the tempting belief that you are immune to bias, lest you fail to meet the high expectations of both of us in our own respective skills.☺

Stability Biases

Stability biases are a way for nature to be efficient. Imagine you find yourself the sole visitor at the matinee showing of an art movie—you therefore could choose literally any of the 200 seats. What would you do: jump up every 30 seconds to try out a different one, or pretty much settle into one seat, at

---

[6]It is a general limitation of social psychology that most empirical research is done within the context of Western culture, with a substantial portion of the research carried out even more narrowly with North American college students. The few studies testing Western theories in Asian cultures such as Japan or China regularly find important cultural differences.

most changing it once or twice to maybe gain more legroom or escape the cold breeze of an obnoxious air conditioning? From nature's perspective, every time you just think about changing your seat, you have already burned mental fuel, and if you actually get up to change a seat, your muscles consume costly energy, let alone that you might be missing the best scene of the movie. A number of biases try to prevent waste of mental and physical resources by "gluing" you to the status quo.

Examples for these biases include the status quo bias and loss aversion. You like the seat you are sitting on better than other seats simply because it is the status quo—and you hate the idea of losing it. This is a specific manifestation of loss aversion that is dubbed the *endowment effect*; it has been shown in experiments involving university coffee mugs and pens that once an object is in your possession (i.e., you are "endowed" with the object), the minimum price at which you are willing to sell might be roughly *double* the maximum price you would be willing to pay for the item.[7]

While economists consider such a situation irrational and abnormal, from nature's perspective it appears perfectly reasonable—nature wants you to either take a rest or do more productive things than trading petty items at negligible personal gain! At times, however, this status quo bias overshoots. For example, corporate decisions in annual budgeting exhibit a very strong status quo bias, with one analysis reporting a 90% correlation in budget allocations year after year (of individual departments or units). While this might have avoided an acrimonious debate of taking away budget from some units, this stability comes at enormous economic cost: companies with more dynamic budget allocation grow twice as fast as those ceding to the status quos bias.[8]

Another important stability bias is the anchoring effect. Econometricians studying time series models often are surprised at how well the so-called naïve model works[9]—for many time series, this period's value is an excellent predictor of the next period's value, and many complex time series models barely outperform this naïve model. Nature must have taken notice because when humans make an estimate, they often root it heavily in whatever initial value they have and make only minor adjustments if new information arises over time. At times, this bias leads seriously astray, however—namely if the initial value is seriously wrong or simply random. A popular demonstration of the anchoring effect involves asking participants to write down the last two digits of their social security or telephone number before estimating the price

[7]D. Kahneman, J.L Knetsch, and R.H. Thaler, "Anomalies: The endowment effect, loss aversion, and status quo bias," *Journal of Economic Perspectives*, 5(1), 193-206, 1991.
[8]T. Baer, S. Heiligtag, and H. Samandari, *The business logic in debiasing*, McKinsey & Co, 2017.
[9]https://blogs.sas.com/content/forecasting/2014/04/30/a-naive-forecast-is-not-necessarily-bad/

of an item, such as a bottle of wine or a box of chocolates. Even though there is obviously absolutely no relationship with these numbers and the price of the item, those writing down high numbers consistently estimate prices 60 to 120 percent higher than those with low numbers.[10]

Pattern-Recognition Biases

Pattern-recognition biases deal with a very vexing problem for our recognition: much of our sensual perception is incomplete, and there is a lot of noise in what we perceive. Imagine the last time you talked with someone—probably it was just a few minutes ago, maybe you spoke to the train conductor or the flight attendant if you're reading this book on the go. Think of a meaty, information-rich sentence the other person said in the middle of the conversation. Very possibly a part of the sentence was actually completely drowned out by a loud noise (e.g., another person's sneeze), several syllables might have been mumbled, and you also may have missed part of the sentence because you glanced at your phone. Did you ask the person to repeat the sentence? Or did you somehow still have a good idea of what the person said? Very often it's the latter—because of an amazing ability of our brain to "fill in the gaps." Our brains excel at very educated guessing—but sometimes these guesses are systematically wrong, and this is the realm of pattern-recognition biases.

Pattern-recognition biases are particularly relevant to this book because pattern-recognition is essentially what algorithms do.

In order to solve the problem of making sense from noisy, incomplete data (be it visual or other sensual perception, or be it actual data such as a management information system report full of tables in small print), the brain needs to develop rules. Systematic errors (i.e., biases) occur if either the rules are wrong or a rule is wrongly applied.

The Texas Sharpshooter fallacy is an example of a flawed rule. Your brain sees rules (i.e., patterns) in the data where none exists. This might explain many superstitions. If for three times in a row a sales person closes a deal while wearing the red tie she got from her husband for her birthday, the brain might jump to a conclusion that it is a "lucky tie." Interestingly, the brain may not be wrong—it's possible that the color red has a psychological effect on buyers that does increase the odds of closing the deal—it's just that three closed deals is a statistically insignificant sample and way too little data to make any robust inference. This illustrates that the way nature thinks about pattern recognition is heavily driven by a "rather safe than sorry" mentality—how many times does the neighbor's dog have to bite you in order for you to conclude that you better not get anywhere close to this cute pooch? By the

[10]E. Teach, "Avoiding Decision Traps," *CFO*, June 1, 2004; Retrieved October 29, 2018.

same token, the brain is hardwired to think that even if there is only a small chance that the red tie helps, why risk a big deal by not wearing it?

Confirmation bias can be an accomplice of the Texas Sharpshooter fallacy and is nature's way of being efficient in the recognition of patterns. Confirmation bias can be seen as a "hypothesis driven" approach to collecting data points. It means that where the mind has a hypothesis (e.g., you already have a belief that buying this book was a great idea), you tend to single out new data that confirms your belief (e.g., you praise the five-star review of this book for its brilliant insights) and reject contradictory data (e.g., you label the author of the one-star review a fool—of course rightly so, may I hasten to say!). Underneath the confirmation bias seems to be nature's desire to get to a decision quickly and to reduce cognitive effort. Laboratory experiments have shown that participants are much more likely to read news articles that support their views than contradictory ones. You'll therefore encounter confirmation bias as a central foe in Chapter 11 about algorithmic bias in social media.

Confirmation bias also can shape how we process "noisy" information. Imagine the above mentioned interaction with a flight attendant or train conductor. She asked about the book you were reading and you proudly showed her the cover of this book. Just as she replied, a loud noise drowned out part of her sentence. There really is no way to tell if she said "I loved that book!" or "I loathed that book!" Except that most likely you "heard" her say that she loved the book. This is because your brain of course would have expected her to say so, and an inconclusive sound would be automatically and subconsciously replaced with the expected content.

Stereotyping is an extension of the confirmation bias and an example of a bias where a rule is applied overly rigidly. First, imagine that you are in a swanky restaurant. The waiter just brought the check to the table next to you where now a stately, senior, white man pulls out a black object from his pants. What do you think it is? You probably imagined a wallet. Now imagine a police car passing a visibly distressed woman lying on the side of the street. As the police car pulls by, the woman shouts "my purse, my purse!" and waves into the air. At this moment, the police officers become aware of a young black man nearby running towards a subway station. They immediately run after the man, shouting "Stop! Police!" and aim their guns at the man. As the man reaches the steps of the entrance of the subway station, he pulls a black object out of his pocket. What is it? If you imagined a gun (not the wallet containing the man's subway pass, which he needs to produce quickly if he doesn't want to miss his train and hence arrive late for his piano lesson), then you fell victim to stereotyping. Based on the situation's context, your brain already has some expectations of what reasonably could happen next. A person in a restaurant who just received a check is likely to pull out a wallet, credit card, or bundle of banknotes from his pocket; a person who appears to have committed a robbery is likely