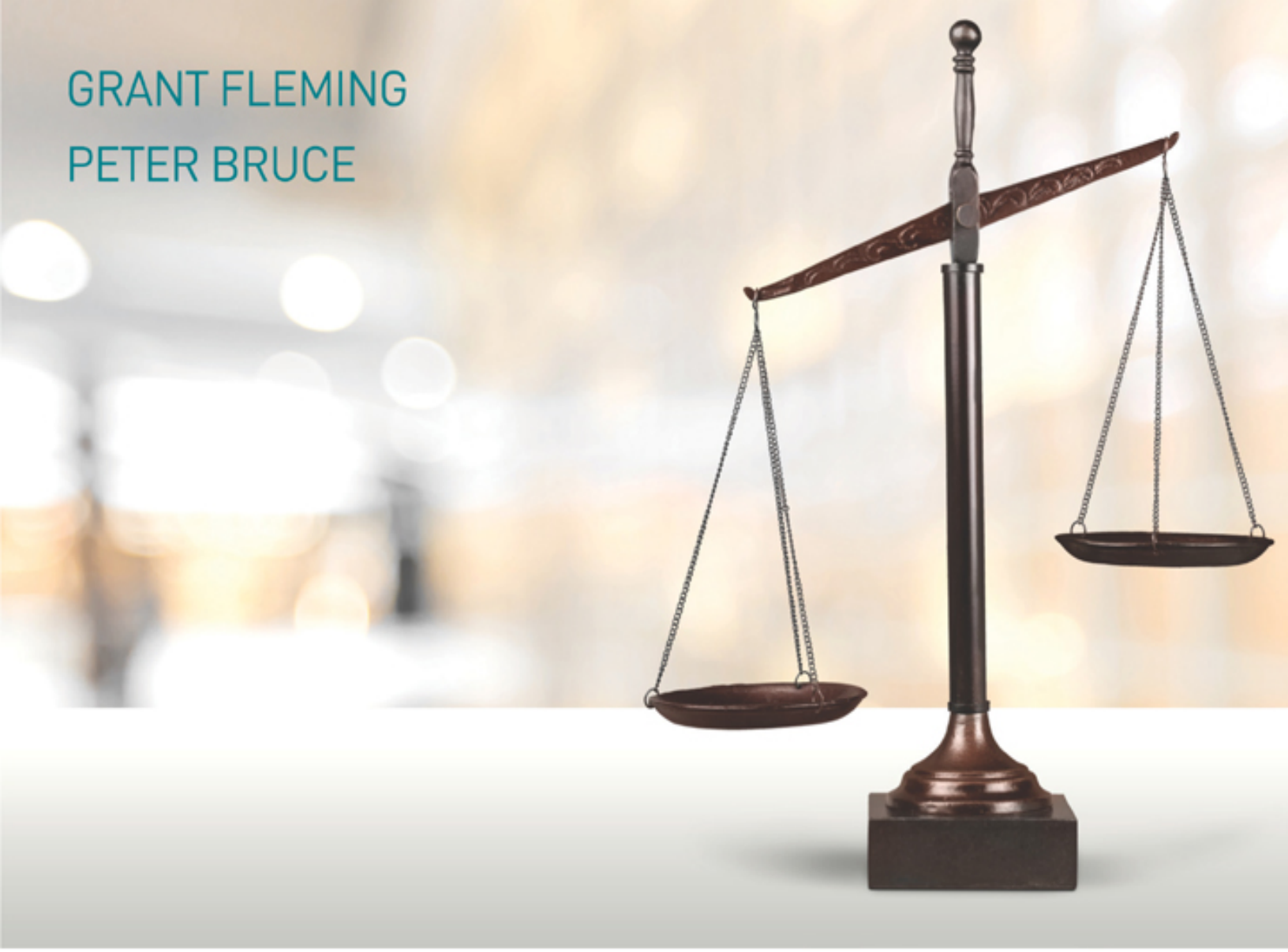
GRANT FLEMING
PETER BRUCE



RESPONSIBLE DATA SCIENCE

Table of Contents

List of Tables

List of Illustrations

Responsible Data Science

Transparency and Fairness in Algorithms

Grant Fleming

Peter Bruce

WILEY

Introduction

In this book, we will review some of the harmful ways artificial intelligence has been used and provide a framework to facilitate the responsible practice of data science. While we will touch upon mitigating legal risks, in this book we will focus primarily on the modeling process itself, especially on how factors overlooked by current modeling practices lead to unintended harms once the model is deployed in a real-world context.

Three core themes will be developed through this book:

- Any AI algorithm can have a harmful, dark side: once they are applied in the real world, AI algorithms can cause any number of harms. An algorithm designed to help police catch murderers can later be appropriated by totalitarian states to persecute dissidents; an algorithm that expands the availability of financial credit for the vast majority of people may nonetheless intensify bias against minorities.
- The dark sides of AI algorithms are created or deepened by current modeling approaches. By focusing only on technical considerations like maximizing predictive performance, data scientists ignore the potential for their model to aggravate biases against certain groups, generate harmful predictions, or otherwise be used by other groups in the future for malicious purposes.
- New modeling approaches are needed if we want to use AI more responsibly. If data scientists and their users are going to continue to use AI algorithms to make consequential decisions, then they ought to do so with

consideration for a broader range of technical and societal factors than are normally considered.

New U.S. diplomats in training used to be told “not to give unintentional offense.” Our primary goal for this book is to tell you a variant of this: that there are a number of specific actionable steps that you, the reader, can begin taking to reduce the risk of causing unintentional harm with your models.

In particular, this book focuses on how to make models more transparent, interpretable, and fair. It will present illustrations and snippets of code in a way that a technically literate manager or executive can understand, without necessarily knowing any programming language.

What This Book Covers

Chapter 1, “Responsible Data Science,” provides historical background for the ethical concerns in statistics and an introduction to basic modeling methods. In Chapter 2, “Background: Modeling and the Black-Box Algorithm,” we define various types of predictive models and briefly discuss the concepts of model transparency and model interpretability. Chapter 3, “The Ways AI Goes Wrong, and the Legal Implications,” reviews the landscape of the types of ethics and fairness issues encountered in the practice of data science (e.g., legal constraints, privacy and data ownership concerns, and algorithms “gone bad”) and finishes by distinguishing interpretable models from black-box models. In Chapter 4, “The Responsible Data Science (RDS) Framework,” we discuss the desired characteristics of a Responsible Data Science framework, summarize the attempts by other groups at creating one, and combine the lessons learned from these other groups with those presented in the book up until this point to construct our own framework, the aptly named the Responsible Data Science (RDS) framework. Chapter 5, “Model Interpretability: The What and the Why,” prepares the reader for implementing the RDS framework in later chapters by doing a deeper dive into model interpretability and how it can be achieved for black-box models. We begin setting up a responsible data science project within our framework and performing initial checks on two datasets in Chapter 6, “Beginning a Responsible Data Science Project.” In Chapters 7, “Auditing a Responsible Data Science Project,” and Chapter 8, “Auditing for Neural Networks,” we delve into case studies in auditing conventional machine learning models and deep neural networks for failure scenarios, fairness, and interpretability. Finally, we conclude the book in Chapter 9, “Conclusion,” with a look to the future and a call to action.

Who Will Benefit Most from This Book

Much has been written elsewhere about the legal issues relevant to AI; thus, our primary audience is not corporate general counsels. Instead, this book is intended for the following two groups:

- Data-literate managers and executives
- Business-literate data scientists and analysts

Although the focus placed on responsibility in data science is relatively new, many people have been trained in the myriad wonderful things that AI can accomplish. They have also read in the news about the ethical lapses in some AI projects. These lapses are not surprising, because relatively few data scientists are trained in how to adequately understand and control their AI while maintaining high predictive performance in models. Hence, we aim this book at data science managers and executives and at data science practitioners.

Practitioners will learn of the ways in which their models, intended to provide benefits, can at the same time cause harm. They will learn how to leverage fairness metrics, interpretability methods, and other interventions to their model or dataset to audit those models, identifying and mitigating possible issues prior to deployment or result delivery. Through worked examples, the book guides users in structuring their models to have a greater consideration for ethical impacts, while assuring that best practices are followed and model performance is optimized. This is a key differentiator for our book, as most responsible AI frameworks do not provide specific technical recommendations for fulfilling the principles that they lay out.

Managers of data science teams, and managers with any responsibilities in the analytics realm, can use this book to stay alert for the ways in which analytical models can run afoul of ethical practices, and even the law. More importantly, they will learn the language and concepts to engage their analytics teams in the solutions and mitigation steps that we propose. While some code and technical discussion is provided, following it in detail is by no means needed. The overall presentation in the book is at a level that provides managers who are at least somewhat familiar with analytics the ability and tools to instill responsible best practices for data science in their organizations.

Finally, a word to individual data scientists. You may think that your project has no implications in the ethical realm. The real-world context for deployment may seem innocuous, the modeling task may seem harmless, and the content of this book may not seem relevant to your project. Though the ideas and techniques presented in this book are primarily discussed in the context of ethically fraught models, they are still useful as the basis for best practices in other modeling contexts. After all, there is a great degree of overlap between traditional best practices for modeling and best practices for responsible data science. Doing data science more responsibly, in the manner that we lay out in this book, improves understanding of the relationships between a model and its real-world deployment context, improves transparency and accountability through better guidelines for documentation, and reduces the risk of unanticipated biases creeping into models by providing workflows for model auditing. Plus, who knows when that innocuous-sounding project may later turn out to have a dark side?

Looking Ahead in This Book

The responsible practice of data science covers a lot of ground in different dimensions.

- **Formal legal and regulatory requirements**: Clearly, any company or individual developing or implementing data science solutions will want to stay on the right side of the law. The most famous attempt to regulate AI is the GDPR; it runs over 80 pages and is quite detailed. It was developed to meet the demands of a specific point in time, but there is no guarantee that it will be a useful guide in the future. Things change rapidly in the field of AI, and the GDPR is like a boulder placed in the path of a stream—sooner or later, the stream will find ways around the obstacle. There are already a number of publications on this topic, and our audience is not the corporate general counsel but rather the manager and the data science practitioner. So, while this book will touch on key laws in this area, such as the GDPR, it will not do so in great depth.
- **Bad actors**: In many cases, the pernicious use of AI is neither inadvertent nor the result of lack of understanding—it is intentional. Deep learning has been put to malicious use by cyber hacks who can digest and analyze multilayered defense mechanisms to determine quickly where weaknesses lie. When those who are responsible for data science development and implementation have malevolent intentions, a lecture on responsibility and a course on ethics will not have much impact. This book will note countermeasures that can have some effect, but dealing with bad actors, like dealing with regulators, is not the primary focus of this book.
- **AI out of control:** In many cases, those deploying AI are responsible parties, obeying the law, and yet their AI has in some sense “escaped their full control” after

deployment. Perhaps it has morphed into something that was not initially intended, or perhaps it has triggered effects and reactions that were unanticipated. Maybe not all decision-makers in the organization that designed the AI, or affiliated stakeholders throughout the project, fully understood or appreciated from the beginning all of the ways that their AI project would operate in a real-world context. The disconnect between the goals of the model and the realities of the real-world context might make it so that even a perfectly accurate model can cause a great deal of harm. This overarching issue is the main focus of the book: how executives, managers, and practitioners can follow best practices in ethical data science—in particular, how they can better understand, explain, and gain control over their AI implementations.

Special Features

DEFINITION **Throughout the book, we'll explain the meanings of terms that may be new or nonstandard.**

NOTE **Inline boxes are used to expand further on some aspect of the topic without interrupting the flow of the narrative.**

> **Small general discussions that deserve special emphasis or have relevance beyond the immediately surrounding content are called out in general sidebar notes.**

Code Repository

Code referred to in the text of each of the chapters, plus updates and expanded code for generating additional results, can be found in the repositories at `www.wiley.com/go/responsibledatascience` and `github.com/Gflemin/responsibledatascience`. Unless otherwise noted in the text, the code to reproduce the results within each of the chapters can be found by navigating to the appropriately named chapter subfolders at either of the links (e.g., the code for Chapter 6 can be found in the `responsibledatascience/ch6` subfolder.) The README file within the head of the code repository folder provides instructions for setting up your software environment, and the README files within each of the chapter subfolders provide additional information about the code for that chapter.

Part I
Motivation for Responsible Data Science and Background Knowledge

In This Part

CHAPTER 1
Responsible Data Science

Data science is an interdisciplinary field that combines elements of statistics, computer science, and information technology to generate useful insights from the increasingly large datasets that are generated in the normal course of business. Data science helps organizations capture value from their data, reducing costs and increasing profits, and also enables completely new types of endeavors, such as powerful information search and self-driving cars. Sometimes, data science projects can go awry, when the predictions made by statistical and machine learning algorithms turn to be not just wrong, but biased and unfair in ways that cause harm. History has shown that the dual good and evil nature of statistical methods is not new, but rather a characteristic that was present from nearly the moment that they were conceived. However, by adjusting and supplementing statistical and machine learning methods and concepts, we can diagnose and reduce the harm that they may otherwise cause.

In popular and technical writing, these issues are often captured by the general term “ethical data science.” We use that term here, but we also use the more general phrase “responsible data science.” Ethics can refer in some usages to narrow “rules of the road” that pertain to a particular profession, such as real estate or accounting. Our goal here is broader than that: presenting a framework for the practice of data science that is ethical, but not in a narrow sense: it is *responsible*.

The Optum Disaster

In 2001, the healthcare company Optum launched Impact-Pro, a predictive modeling tool. Impact-Pro was an early success for predictive analytics (predating the term *data science*), and a decade later, Steven Wickstrom, an Optum VP, touted its use cases. For healthcare providers, it could "support steerage to appropriate programs" and "identify members [patients] with gaps in care, complications, and comorbidities." Optum termed these *care opportunities* in one document (i.e., opportunities for more revenue), but they are also of interest to those concerned with cost management: the correct early intervention in a health problem can cost significantly less than more drastic action later. For insurers, information on health risks for specific groups and individuals could be used to set premiums more accurately than is possible using traditional underwriting criteria.

DEFINITION DATA SCIENCE **We use the term *data science* broadly to cover the process of understanding and defining a problem, gathering and preparing data, using statistical methods to answer questions, fitting models and assessing them, and deploying models in an organizational setting. We consider artificial intelligence (AI) to be part of data science, and we also consider the "science" component of data science to be important.**

In 2019, though, a research team found that the tool was fundamentally flawed. For one important group—African Americans—the tool consistently underpredicted need for healthcare. The reason? The tool was essentially built to predict future spending on healthcare, and prior spending was a key predictor for that goal. And prior spending is a function not just of need, but also of ability to pay for and gain access to healthcare. Relative to other ethnic groups in the United States, African Americans have been (and

continue to be) less insured, are less able to access healthcare, and possess fewer financial resources for covering healthcare expenses. In Optum's data, therefore, African Americans had less prior spending and, hence, less predicted future need. As a result, African Americans were less targeted for preventive intervention and necessary follow-up healthcare than were other people with similar health profiles. Neither the model nor the data provided to it were able to account for the unanticipated and overlooked societal inequities lurking beneath.

Optum was blindsided. The company thought it had built a tool that was a winner on all fronts: improving health outcomes by being smarter about required follow-up care, and managing costs better in the bargain. Instead, it found itself the focus of widespread bad publicity and was pilloried for creating a product that exacerbated racial bias and widened the healthcare gap faced by African Americans. New York state regulators opened an investigation, and the controversy continued into 2020. At the time of writing, Optum continues to market Impact Pro.

In this case, and in many others, the original intent for using the algorithm was good: good for healthcare providers by optimizing the allocation of scarce resources, and good for patients by ensuring that patients with the greatest needs had those needs met. But good intentions plus smart *artificial intelligence* (AI) led to disaster.

DEFINITION ARTIFICIAL INTELLIGENCE **We use the term *artificial intelligence* generally, to cover both statistical and machine learning methods for prediction with structured numeric data and text, as well as image and voice recognition and synthesis. In this book, we think of AI as having underlying algorithms or models. When discussing solutions for reducing the harms of AI, changing these underlying algorithms or models will be one of the main focal points**

Interestingly, the scenario of good statistics being ill-used is not new. In fact, statistics as a field has a long history of being used for nefarious purposes or causing unintended harms.

Jekyll and Hyde

Let's begin with a look back over a century in history to a classic work of fiction that serves as a metaphor for the issues we face with data science today. In his gothic tale *The Strange Case of Dr. Jekyll and Mr. Hyde*, Robert Louis Stevenson describes two characters. Dr. Jekyll is an analytical man of science, a great asset to society, and a doer of good deeds. However, there is a repulsive, cruel side to him in the form of a separate character, Mr. Hyde, who gets “released” from time to time. The evil Mr. Hyde, in his times of release, tramples a young girl, commits murder, and more. The phrase “Jekyll and Hyde” has come to represent something that has two contradictory but inextricably linked natures—one respected and upright, the other base and evil.

The dual nature of humanity—good and evil combined in the same package—is a universal theme in literature. As humans carry their intelligence into the artificial realm, this duality has come with it.

Artificial intelligence has taken on this Jekyll and Hyde character trait. The enormous benefits brought by AI are evident: it has been a major force powering economic growth over the last several decades. Most aspects of life and industry now incorporate AI approaches in some way. Here are just a few examples:

- When you apply for a loan or a credit card, it is an algorithm that judges whether the application should be approved. This speeds the process, lowers the cost of providing credit, and, by making the process more scientific, standardizes decisions and expands access to credit among the truly creditworthy.
- When you use Facebook, Instagram, Twitter, or other social media services, the ads you see are optimized by an algorithm to be those most likely to get you to respond. This “microtargeting” makes them more relevant to you and, more importantly, makes it possible to provide these social media services at no charge to the user.
- Criminals are often caught on camera at or near the scene of a crime, and facial recognition and identification algorithms make it much more likely that they will be identified and caught.

In each of these cases we can point to a related “Mr. Hyde” that lurks in the background.

- Loan approval algorithms, it turns out, are prone to “redlining” just as humans are, blocking whole neighborhoods from credit, rather than making decisions on the basis of individual characteristics. Moreover, unlike humans, algorithms, if they are not transparent, are resistant to moral suasion and are hard to correct.

- The economic efficiencies wrought by microtargeting of ads is offset by the unease many people feel about being “surveilled.” What's more, algorithmic curation of content feeds, seeking to maximize user engagement, drives users towards content that is provocative, inflammatory, and often fabricated. Even without actively provoking, these same recommender algorithms that underpin social media companies also enable political extremists to coalesce and take action.

Computer image recognition algorithms that have been so helpful to law enforcement facilitate dramatic erosions of privacy: one company has scraped the Web and built a database of billions of tagged face images, allowing individuals to upload images of people and find out who they are. When these facial recognition approaches are deployed by law enforcement, the harm resulting from erroneous identifications is magnified, especially for darker skinned individuals who are more likely to be falsely identified by these approaches. Sometimes, the negative Mr. Hyde aspect is only weakly counterbalanced by a good Mr. Jekyll. The science of image and voice synthesis has introduced the world to destructive “deep fakes”: fabricated videos of people (usually political figures or celebrities) saying things they never said. Individuals or organizations bent on sowing discord or disinformation, or inciting violence have already used deep fakes for these aims. The plus side of the technology is comparatively minimal: better avatars for video games and production efficiencies for Hollywood, which needn't hire so many actors. The public has been highly exposed to these failures (possibly more than to the successes) through public controversies and popular science journalism and books. The good and evil sides to AI are now widely recognized, but this is not the first time that statistics has gone over to the “dark side.” Indeed, some of the most foundational

breakthroughs in statistical methodology were motivated by goals we now recognize as morally reprehensible.

Eugenics

Turn back the clock to 1886, the very year *The Strange Case of Dr. Jekyll and Mr. Hyde* was published. This was also the year that the famous British statistician Francis Galton published his article “Regression Towards Mediocrity in Hereditary Stature,” referring to the tendency of very tall and very short parents to have children closer to average height. This phenomenon gave us the phrase “regression to the mean.”

Galton, Pearson, and Fisher

Galton, in addition to his seminal work on regression, also made contributions in correlation and survey methods. His half-cousin was Charles Darwin, and Galton was much taken with Darwin's *The Origin of Species*. Galton thought that, with the help of statistical methods, the evolution of humans could be guided in a positive and useful way. He coined the term *eugenics*, focused much of his research and scientific publications on eugenics, and became the Honorary President of the British Eugenics Society.

Karl Pearson, who contributed to statistics the correlation coefficient, principal components, the (increasingly maligned) p-value, and much more, was a protégée of Galton who assumed the Galton Chair of Eugenics at the University of London. Pearson saw the ideal society as:

> an organized whole, kept up to a high pitch of internal efficiency by insuring that its numbers are substantially recruited from the better stocks, and kept up to a high pitch of external efficiency by contest, chiefly by way of war with inferior races.