Trustworthy Machine Learning

Kush R. Varshney

ISBN 979-8-41-190395-9

Kush R. Varshney / Trustworthy Machine Learning
Chappaqua, New York, USA

Cover image by W. T. Lee, United States Geological Survey, circa 1925. The photograph shows a person standing on top of a horse, which is standing on a precarious section of a desolate rock formation. The horse must really be worthy of the person's trust.

How to cite:

- APA: Varshney, K. R. (2022). *Trustworthy Machine Learning*. Independently Published. http://www.trustworthymachinelearning.com.
- IEEE: K. R. Varshney, *Trustworthy Machine Learning*. Chappaqua, NY, USA: Independently Published, 2022.
- MLA: Varshney, Kush R. *Trustworthy Machine Learning*. Independently Published, 2022.
- Bibtex:

```
@book{Varshney2022,
        author="Kush R. Varshney",
        title="Trustworthy Machine Learning",
        publisher="Independently Published",
        address="Chappaqua, NY, USA",
        year="2022"
}
```

“For it is in giving that we receive.”

—St. Francis

Contents

Preface

Decision making in high-stakes applications, such as educational assessment, credit, employment, health care, and criminal justice, is increasingly data-driven and supported by machine learning models. Machine learning models are also enabling critical cyber-physical systems such as self-driving automobiles and robotic surgery. Recommendations of content and contacts on social media platforms are determined by machine learning systems.

Advancements in the field of machine learning over the last several years have been nothing short of amazing. Nonetheless, even as these technologies become increasingly integrated into our lives, journalists, activists, and academics uncover characteristics that erode the trustworthiness of these systems. For example, a machine learning model that supports judges in pretrial detention decisions was reported to be biased against black defendants. Similarly, a model supporting resume screening for employment at a large technology company was reported to be biased against women. Machine learning models for computer-aided diagnosis of disease from chest x-rays were shown to give importance to markers contained in the image, rather than details of the patients' anatomy. Self-driving car fatalities have occurred in unusual confluences of conditions that the underlying machine learning algorithms had not been trained on. Social media platforms have knowingly and surreptitiously promoted harmful content. In short, while each day brings a new story of a machine learning algorithm achieving superhuman performance on some task, these marvelous results are only in the *average* case. The reliability, safety, security, and transparency required for us to trust these algorithms in *all* cases remains elusive. As a result, there is growing popular will to have more fairness, robustness, interpretability, and transparency in these systems.

They say "history doesn't repeat itself, but it often rhymes." We have seen the current state of affairs many times before with technologies that were new to their age. The 2016 book *Weapons of Math Destruction* by Cathy O'Neil, catalogs numerous examples of machine learning algorithms gone amok. In the conclusion, O'Neil places her work in the tradition of Progressive Era muckrakers Upton Sinclair and Ida Tarbell. Sinclair's classic 1906 book *The Jungle* tackled the processed food industry. It helped spur the passage of the Federal Meat Inspection Act and the Pure Food and Drug Act, which together regulated that all foods must be cleanly prepared and free from adulteration.

In the 1870s, Henry J. Heinz started one of the largest food companies in the world today. At a time when food companies were adulterating their products with wood fibers and other fillers, Heinz started selling horseradish, relishes, and sauces made of natural and organic ingredients. Heinz offered these products in transparent glass containers when others were using dark containers. His company innovated processes for sanitary food preparation, and was the first to offer factory tours that were open to the public. The H. J. Heinz Company lobbied for the passage of the aforementioned Pure Food and Drug Act, which became the precursor to regulations for food labels and tamper-resistant packaging. These practices increased trust and adoption of the products. They provided Heinz a competitive advantage, but also advanced industry standards and benefited society.

And now to the rhyme. What is the current state of machine learning and how do we make it more trustworthy? What are the analogs to natural ingredients, sanitary preparation, and tamper-resistant

packages? What are machine learning's transparent containers, factory tours, and food labels? What is the role of machine learning in benefiting society?

The aim of this book is to answer these questions and present a unified perspective on trustworthy machine learning. There are several excellent books on machine learning in general from various perspectives. There are also starting to be excellent texts on individual topics of trustworthy machine learning such as fairness[1] and explainability.[2] However, to the best of my knowledge, there is no single self-contained resource that defines trustworthy machine learning and takes the reader on a tour of all the different aspects it entails.

I have tried to write the book I would like to read if I were an advanced technologist working in a high-stakes domain who does not shy away from some applied mathematics. The goal is to impart a *way of thinking* about putting together machine learning systems that regards safety, openness, and inclusion as first-class concerns. We will develop a *conceptual foundation* that will give you the confidence and a starting point to dive deeper into the topics that are covered.

> "Many people see computer scientists as builders, as engineers, but I think there's a deeper intellectual perspective that many CS people share, which sees computation as a metaphor for how we think about the world."
>
> —Suresh Venkatasubramanian, computer scientist at Brown University

We will neither go into extreme depth on any one topic nor work through software code examples, but will lay the groundwork for how to approach real-world development. To this end, each chapter contains a realistic, but fictionalized, scenario drawn from my experience that you might have already faced or will face in the future. The book contains a mix of narrative and mathematics to elucidate the increasingly sociotechnical nature of machine learning and its interactions with society. The contents rely on some prior knowledge of mathematics at an undergraduate level as well as statistics at an introductory level.[3]

> "If you want to make a difference, you have to learn how to operate within imperfect systems. Burning things down rarely works. It may allow for personal gains. But if you care about making the system work for many, you have to do it from the inside."
>
> —Nadya Bliss, computer scientist at Arizona State University

The topic of the book is intimately tied to social justice and activism, but I will primarily adopt the Henry Heinz (developer) standpoint rather than the Upton Sinclair (activist) standpoint. This choice is

---

[1]Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. URL: https://fairmlbook.org, 2020.

[2]Christoph Molnar. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. URL: https://christophm.github.io/interpretable-ml-book, 2019.

[3]A good reference for mathematical background is: Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong. *Mathematics for Machine Learning*. Cambridge, England, UK: Cambridge University Press, 2020.

not meant to disregard or diminish the essential activist perspective, but represents my perhaps naïve technological solutionist ethos and optimism for improving things from the inside. Moreover, most of the theory and methods described herein are only small pieces of the overall puzzle for making machine learning worthy of society's trust; there are procedural, systemic, and political interventions in the sociotechnical milieu that may be much more powerful.

This book stems from my decade-long professional career as a researcher working on high-stakes applications of machine learning in human resources, health care, and sustainable development as well as technical contributions to fairness, explainability, and safety in machine learning and decision theory. It draws on ideas from a large number of people I have interacted with over many years, filtered through my biases. I take responsibility for all errors, omissions, and misrepresentations. I hope you find it useful in your work and life.