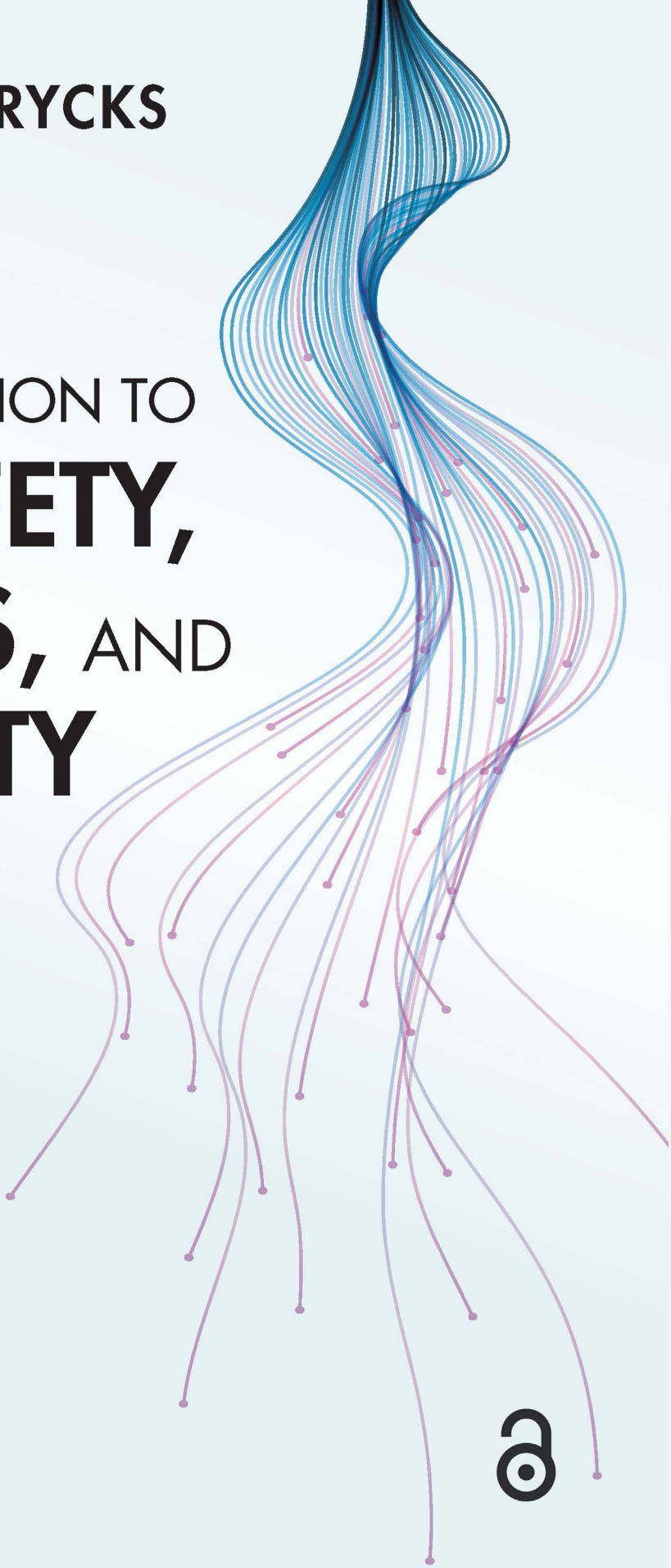
DAN HENDRYCKS

INTRODUCTION TO AI SAFETY, ETHICS, AND SOCIETY



CRC Press
Taylor & Francis Group

"This book is an important resource for anyone interested in understanding and mitigating the risks associated with increasingly powerful AI systems. It provides not only an accessible introduction to the technical challenges in making AI safer, but also a clear-eyed account of the coordination problems we will need to solve on a societal level to ensure AI is developed and deployed safely."

Yoshua Bengio, Professor of Computer Science, University of Montreal and Turing Award Winner.

"A must-read for anyone seeking to understand the full complexities of AI risk."

David Krueger, Assistant Professor, Department of Engineering, University of Cambridge

"The most comprehensive exposition for the case that AI raises catastrophic risks and what to do about them. Even if you disagree with some of Hendrycks' arguments, this book is still very much worth reading, if only for the unique coverage of both the technical and social aspects of the field."

Boaz Barak, Gordon McKay Professor of Computer Science, Harvard University

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Introduction to AI Safety, Ethics, and Society

As AI technology is rapidly progressing in capability and being adopted more widely across society, it is more important than ever to understand the potential risks AI may pose and how AI can be developed and deployed safely. ***Introduction to AI Safety, Ethics, and Society*** offers a comprehensive and accessible guide to this topic.

This book explores a range of ways in which societies could fail to harness AI safely in coming years, such as malicious use, accidental failures, erosion of safety standards due to competition between AI developers or nation-states, and potential loss of control over autonomous systems. Grounded in the latest technical advances, this book offers a timely perspective on the challenges involved in making current AI systems safer. Ensuring that AI systems are safe is not just a problem for researchers in machine learning – it is a societal challenge that cuts across traditional disciplinary boundaries. Integrating insights from safety engineering, economics, and other relevant fields, this book provides readers with fundamental concepts to understand and manage AI risks more effectively.

This is an invaluable resource for upper-level undergraduate and postgraduate students taking courses relating to AI safety & alignment, AI ethics, AI policy, and the societal impacts of AI, as well as anyone trying to better navigate the rapidly evolving landscape of AI safety.

Dr. Dan Hendrycks is a machine learning researcher and Director of the Center for AI Safety (CAIS). He has conducted pioneering research in AI such as developing the GELU activation function, used in several state-of-the art neural networks such as GPT, and creating MMLU, one of the leading benchmarks used to assess large language models. His research has been covered by the BBC, New York Times, and Washington Post.

His work currently focuses on improving the safety of AI systems and mitigating risks from AI. He has advised the UK government on AI safety and has been invited to give talks on this topic at OpenAI, Google, and Stanford, among other institutions. He has written on AI risks for the Wall Street Journal and TIME Magazine. Dan Hendrycks holds a Ph.D. in Machine Learning from UC Berkeley.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Introduction to AI Safety, Ethics, and Society

Dan Hendrycks

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

First edition published 2025
by CRC Press
2385 Executive Center Drive, Suite 320, Boca Raton, FL 33431

and by CRC Press
4 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

CRC Press is an imprint of Taylor & Francis Group, LLC

Library of Congress Cataloging-in-Publication Data
Names: Hendrycks, Dan, author.
Title: Introduction to AI safety, ethics, and society / Dan Hendrycks.
Other titles: Introduction to artificial intelligence safety, ethics, and society
Description: First edition. | Boca Raton : CRC Press, 2025. | Includes bibliographical references and index.
Identifiers: LCCN 2024031863 (print) | LCCN 2024031864 (ebook) | ISBN 9781032869926 (hbk) | ISBN 9781032917221 (pbk) | ISBN 9781003530336 (ebk)
Subjects: LCSH: Artificial intelligence--Moral and ethical aspects. | Artificial intelligence--Social aspects. | Risk management.
Classification: LCC Q334.7 .H46 2025 (print) | LCC Q334.7 (ebook) | DDC 174/.90063--dc23/eng/20240928
LC record available at https://lccn.loc.gov/2024031863
LC ebook record available at https://lccn.loc.gov/2024031864

ISBN: 978-1-032-86992-6 (hbk)
ISBN: 978-1-032-91722-1 (pbk)
ISBN: 978-1-003-53033-6 (ebk)

DOI: 10.1201/9781003530336

Publisher's note: This book has been prepared from camera-ready copy provided by the authors.

Typeset in Latin Modern
by KnowledgeWorks Global Ltd.

Access the Support Material: https://www.routledge.com/9781032798028

Contents

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Introduction

Artificial intelligence (AI) is rapidly embedding itself within militaries, economies, and societies, reshaping their very foundations. Given the depth and breadth of its consequences, it has never been more pressing to understand how to ensure that AI systems are safe, ethical, and have a positive societal impact.

This book aims to provide a comprehensive approach to understanding AI risk. Our primary goals include consolidating fragmented knowledge on AI risk, increasing the precision of core ideas, and reducing barriers to entry by making content simpler and more comprehensible. The book has been designed to be accessible to readers from diverse backgrounds. You do not need to have studied AI, philosophy, or other such topics. The content is skimmable and somewhat modular, so that you can choose which chapters to read. We introduce mathematical formulas in a few places to specify claims more precisely, but readers should be able to understand the main points without these.

AI risk is multidisciplinary. Most people think about problems in AI risk in terms of largely implicit conceptual models, which significantly affect how they approach these challenges. We aim to replace these implicit models with explicit, time-tested models. A full understanding of the risks posed by AI requires knowledge in several disparate academic disciplines, which have so far not been combined in a single text. This book was written to fill that gap and adequately equip readers to analyze AI risk, and moves beyond the confines of machine learning to provide a holistic understanding of AI risk. We draw on well-established ideas and frameworks from the fields of engineering, economics, biology, complex systems, philosophy, and other disciplines that can provide insights into AI risks and how to manage them. Our aim is to equip readers with a solid understanding of the technical, ethical, and governance challenges that we will need to meet in order to harness advanced AI in a beneficial way.

In order to understand the challenges of AI safety, it is important to consider the broader context within which AI systems are being developed and applied. The decisions of and interplay between AI developers, policy-makers, militaries, and other actors will play an important role in shaping this context. Since AI influences many different spheres, we have deliberately selected time-tested, formal frameworks to provide multiple lenses for thinking about AI, relevant actors, and AI's impacts. The frameworks and concepts we use are highly general and are useful for reasoning about various forms of intelligence, ranging from individual human beings to corporations, states, and AI systems. While some sections of the book focus more directly on AI

risks that have already been identified and discussed today, others set out a systematic introduction to ideas from game theory, complex systems, international relations, and more. We hope that providing these flexible conceptual tools will help readers to adapt robustly to the ever-changing landscape of AI risks.

This book does not aim to be the definitive guide on all AI risks. Research on AI risk is still new and rapidly evolving, making it infeasible to comprehensively cover every risk and its potential solutions in a single book, particularly if we wish to ensure that the content is clear and digestible. We have chosen to introduce concepts and frameworks that we find productive for thinking about a wide range of AI risks. Nonetheless, we have had to make choices about what to include and omit. Many present harms, such as harmful malfunctions, misinformation, privacy breaches, reduced social connection, and environmental damage, are already well-addressed by others [1, 2]. Given the rapid development of AI in recent years, we focus on novel risks posed by advanced systems: risks that pose serious, large-scale, and sometimes irreversible threats that our societies are currently unprepared to face.

Even if we limit ourselves to focusing on the potential for AI to pose catastrophic risks, it is easy to become disoriented given the broad scope of the problem. Our hope is that this book provides a starting point for others to build their own picture of these risks and opportunities, and our potential responses to them.

The book's content falls into three sections: AI and Societal-Scale Risks, Safety, and Ethics and Society. In the AI and Societal-Scale Risks section, we outline major categories of AI risks and introduce some key features of modern AI systems. In the Safety section, we discuss how to make individual AI systems more safe. However, if we can make them safe, how should we direct them? To answer this, we turn to the Ethics and Society section and discuss how to make AI systems that promote our most important values. In this section, we also explore the numerous challenges that emerge when trying to coordinate between multiple AI systems, multiple AI developers, or multiple nation-states with competing interests.

The AI and Societal-Scale Risks section starts with an informal overview of AI risks, which summarises many of the key concerns discussed in this book. We outline some scenarios where AI systems could cause catastrophic outcomes. We split risks across four categories: malicious use, AI arms race dynamics, organizational risks, and rogue AIs. These categories can be loosely mapped onto the risks discussed in more depth in the Governance, Collective Action Problems, Safety Engineering, and Single-Agent Safety chapters, respectively. However, this mapping is imperfect as many of the risks and frameworks discussed in the book are more general and cut across scenarios. Nonetheless, we hope that the scenarios in this first chapter give readers a concrete picture of the risks that we explore in this book. The next chapter, Artificial Intelligence Fundamentals, aims to provide an accessible and non-mathematical explanation of current AI systems, helping to familiarise readers with key terms and concepts in machine learning, DL, scaling laws, and so on. This provides the necessary foundations for the discussion of the safety of individual AI systems in the next section.

The Safety section gives an overview of core challenges in safely building advanced AI systems. It draws on insights from both machine learning research and general theories of safety engineering and complex systems, which provide a powerful lens for understanding these issues. In Single-Agent Safety, we explore challenges in making individual AI systems safer, such as bias, transparency, and emergence. In Safety Engineering, we discuss principles for creating safer organizations and how these may apply to those developing and deploying AI. The need for a robust safety culture at organizations developing AI is crucial, so organizations do not prioritize profit at the expense of safety. Next, in Complex Systems, we show that analyzing AIs as complex systems helps us to better understand the difficulty of predicting how they will respond to external pressures or controlling the goals that may emerge in such systems. More generally, this chapter provides us with a useful vocabulary for discussing diverse systems of interest.

The Ethics and Society section focuses on how to instill beneficial objectives and constraints in AI systems and how to enable effective collaboration between stakeholders to mitigate risks. In Beneficial AI and Machine Ethics, we introduce the challenge of giving AI systems objectives that will reliably lead to beneficial outcomes for society, and discuss various proposals along with the challenges they face. In Collective Action Problems, we utilize game theory to illustrate the many ways in which multiple agents (such as individual humans, companies, nation-states, or AIs) can fail to secure good outcomes and come into conflict. We also consider the evolutionary dynamics shaping AI development and how these drive AI risks. These frameworks help us to understand the challenges of managing competitive pressures between AI developers, militaries, or AI systems themselves. Finally, in the Governance chapter, we discuss strategic variables such as how widely access to powerful AI systems is distributed. We introduce a variety of potential paths for managing AI risks, including corporate governance, national regulation, and international coordination.

The website for this book (www.aisafetybook.com) includes a range of additional content. It contains further educational resources such as videos, slides, quizzes, and discussion questions. For readers interested in contributing to mitigating risks from AI, it offers some brief suggestions and links to other resources on this topic. A range of appendices can also be found on the website with further material that could not be included in the book itself.

Dan Hendrycks
Center for AI Safety