

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

© 2025 Dan Hendrycks

Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

The Open Access version of this book, available at www.taylorfrancis.com, has been made available under a Creative Commons [Attribution-Non Commercial-No Derivatives (CC-BY-NC-ND)] 4.0 license.

Any third party material in this book is not included in the OA Creative Commons license, unless indicated otherwise in a credit line to the material. Please direct any permissions enquiries to the original rightholder.

Trademark notice: Product or corporate names may be trademarks or registered trademarks and are used only for identification and explanation without intent to infringe.

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 (ebl)
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 (ebl)

DOI: [10.1201/9781003530336](https://doi.org/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

Introduction	xv
--------------	----

SECTION I AI and Societal-Scale Risks

CHAPTER	1 ■ Overview of Catastrophic AI Risks	3
1.1	INTRODUCTION	3
1.2	MALICIOUS USE	6
1.2.1	Bioterrorism	7
1.2.2	Unleashing AI Agents	9
1.2.3	Persuasive AIs	10
1.2.4	Concentration of Power	11
1.3	AI RACE	14
1.3.1	Military AI Arms Race	14
1.3.2	Corporate AI Race	20
1.3.3	Evolutionary Pressures	23
1.4	ORGANIZATIONAL RISKS	28
1.4.1	Accidents Are Hard to Avoid	30
1.4.2	Organizational Factors can Reduce the Chances of Catastrophe	32
1.5	ROGUE AIs	37
1.5.1	Proxy Gaming	38
1.5.2	Goal Drift	41
1.5.3	Power-Seeking	43
1.5.4	Deception	45
1.6	DISCUSSION OF CONNECTIONS BETWEEN RISKS	48
1.7	CONCLUSION	49
1.8	LITERATURE	50
1.8.1	Recommended Reading	50
CHAPTER	2 ■ Artificial Intelligence Fundamentals	51
2.1	INTRODUCTION	51
2.2	ARTIFICIAL INTELLIGENCE & MACHINE LEARNING	52
2.2.1	Artificial Intelligence	53

2.2.2	Types of AI	58
2.2.3	Machine Learning	64
2.2.4	Types of Machine Learning	75
2.3	DEEP LEARNING	79
2.3.1	Model Building Blocks	82
2.3.2	Training and Inference	93
2.3.3	History and Timeline of Key Architectures	98
2.3.4	Applications	101
2.4	SCALING LAWS	102
2.4.1	Scaling Laws in DL	104
2.5	SPEED OF AI DEVELOPMENT	107
2.6	CONCLUSION	110
2.6.1	Summary	110
2.7	LITERATURE	112
2.7.1	Recommended Resources	113

SECTION II Safety

CHAPTER	3 ■ Single-Agent Safety	117
3.1	INTRODUCTION	117
3.2	MONITORING	118
3.2.1	ML Systems Are Opaque	118
3.2.2	Motivations for Transparency Research	121
3.2.3	Approaches to Transparency	122
3.2.4	Emergent Capabilities	127
3.2.5	Emergent Goal-Directed Behavior	129
3.2.6	Tail Risk: Emergent Goals	132
3.2.7	Evaluations and Anomaly Detection	134
3.3	ROBUSTNESS	135
3.3.1	Proxies in ML	136
3.3.2	Proxy Gaming	136
3.3.3	Adversarial Examples	143
3.3.4	Trojan Attacks and Other Security Threats	147
3.3.5	Tail Risk: AI Evaluator Gaming	148
3.4	ALIGNMENT	150
3.4.1	Deception	151
3.4.2	Deceptive Evaluation Gaming	154
3.4.3	Tail Risk: Deceptive Alignment and Treacherous Turns	156
3.4.4	Power	157
3.4.5	People Could Enlist AIs for Power Seeking	160
3.4.6	Power Seeking Can Be Instrumentally Rational	160

3.4.7	Structural Pressures Toward Power-Seeking AI	165
3.4.8	Tail Risk: Power-Seeking Behavior	167
3.4.9	Techniques to Control AI Systems	167
3.5	SYSTEMIC SAFETY	169
3.6	SAFETY AND GENERAL CAPABILITIES	171
3.7	CONCLUSION	173
3.8	LITERATURE	176
3.8.1	Recommended Reading	176
CHAPTER	4 ■ Safety Engineering	178
4.1	RISK DECOMPOSITION	179
4.1.1	Failure Modes, Hazards, and Threats	179
4.1.2	The Classic Risk Equation	180
4.1.3	Framing the Goal as Risk Reduction	181
4.1.4	Disaster Risk Equation	181
4.1.5	Elements of the Risk Equation	182
4.1.6	Applying the Disaster Risk Equation	183
4.2	NINES OF RELIABILITY	185
4.3	SAFE DESIGN PRINCIPLES	188
4.3.1	Redundancy	189
4.3.2	Separation of Duties	189
4.3.3	Principle of Least Privilege	190
4.3.4	Fail-Safes	190
4.3.5	Antifragility	191
4.3.6	Negative Feedback Mechanisms	192
4.3.7	Transparency	192
4.3.8	Defense in Depth	193
4.3.9	Review of Safe Design Principles	194
4.4	COMPONENT FAILURE ACCIDENT MODELS AND METHODS	194
4.4.1	Swiss Cheese Model	194
4.4.2	Bow Tie Model	196
4.4.3	Fault Tree Analysis Method	197
4.4.4	Limitations	199
4.5	SYSTEMIC FACTORS	205
4.5.1	Systemic Accident Models	206
4.6	DRIFT INTO FAILURE AND EXISTENTIAL RISKS	215
4.7	TAIL EVENTS AND BLACK SWANS	217
4.7.1	Introduction to Tail Events	217
4.7.2	Tail Events Can Greatly Affect the Average Risk	218
4.7.3	Tail Events Can Be Identified From Frequency Distributions	220
4.7.4	A Caricature of Tail Events	221

4.7.5	Introduction to Black Swans	224
4.7.6	Known Unknowns and Unknown Unknowns	224
4.7.7	Implications of Tail Events and Black Swans for Risk Analysis	227
4.7.8	Identifying the Risk of Tail Events or Black Swans	233
4.8	CONCLUSION	235
4.8.1	Summary	235
4.8.2	Key Takeaways	237
4.9	LITERATURE	239
4.9.1	Recommended Reading	239
CHAPTER	5 ▪ Complex Systems	240
5.1	OVERVIEW	240
5.2	INTRODUCTION TO COMPLEX SYSTEMS	241
5.2.1	The Reductionist Paradigm	241
5.2.2	The Complex Systems Paradigm	244
5.2.3	DL Systems as Complex Systems	247
5.2.4	Complexity Is Not a Dichotomy	248
5.2.5	The Hallmarks of Complex Systems	248
5.2.6	Social Systems as Complex Systems	260
5.3	COMPLEX SYSTEMS FOR AI SAFETY	265
5.3.1	General Lessons from Complex Systems	265
5.3.2	Puzzles, Problems, and Wicked Problems	269
5.3.3	Challenges With Interventionism	271
5.3.4	Systemic Issues	276
5.4	CONCLUSION	278
5.5	LITERATURE	280
5.5.1	Recommended Reading	280
SECTION III	Ethics and Society	
CHAPTER	6 ▪ Beneficial AI and Machine Ethics	283
6.1	INTRODUCTION	283
6.2	LAW	285
6.2.1	The Case for Law	286
6.2.2	The Need for Ethics	289
6.3	FAIRNESS	292
6.3.1	Bias	293
6.3.2	Sources of Bias	294
6.3.3	AI Fairness Concepts	297

6.3.4	Limitations of Fairness	299
6.3.5	Approaches to Combating Bias and Improving Fairness	300
6.4	THE ECONOMIC ENGINE	303
6.4.1	Allocative Efficiency of Free Markets	304
6.4.2	Market Failures	305
6.4.3	Inequality	309
6.4.4	Growth	313
6.4.5	Beyond Economic Models	314
6.5	WELLBEING	318
6.5.1	Wellbeing as the Net Balance of Pleasure over Pain	318
6.5.2	Wellbeing as a Collection of Objective Goods	319
6.5.3	Wellbeing as Preference Satisfaction	319
6.5.4	Applying the Theories of Wellbeing	322
6.6	PREFERENCES	324
6.6.1	Revealed Preferences	325
6.6.2	Stated Preferences	327
6.6.3	Idealized Preferences	330
6.7	HAPPINESS	333
6.7.1	The General Approach to Happiness	334
6.7.2	Problems for Happiness-Focused Ethics	337
6.8	SOCIAL WELFARE FUNCTIONS	340
6.8.1	Measuring Social Welfare	342
6.9	MORAL UNCERTAINTY	350
6.9.1	Making Decisions Under Moral Uncertainty	350
6.9.2	Implementing a Moral Parliament in AI Systems	356
6.9.3	Advantages of a Moral Parliament	356
6.10	CONCLUSION	359
6.11	LITERATURE	360
6.11.1	Recommended Reading	360
CHAPTER	7 ■ Collective Action Problems	362
7.1	MOTIVATION	362
7.2	GAME THEORY	366
7.2.1	Overview	366
7.2.2	Game Theory Fundamentals	368
7.2.3	The Prisoner's Dilemma	369
7.2.4	The Iterated Prisoner's Dilemma	377
7.2.5	Collective Action Problems	392
7.2.6	Summary	399
7.3	COOPERATION	400
7.3.1	Summary	413

7.4	CONFLICT	414
7.4.1	Overview	414
7.4.2	Bargaining Theory	416
7.4.3	Commitment Problems	417
7.4.4	Information Problems	423
7.4.5	Factors Outside of Bargaining Theory	426
7.4.6	Summary	428
7.5	EVOLUTIONARY PRESSURES	429
7.5.1	Overview	429
7.5.2	Generalized Darwinism	429
7.5.3	Levels of Selection and Selfish Behavior	436
7.5.4	Summary	442
7.6	CONCLUSION	442
7.7	LITERATURE	444
7.7.1	Recommended Reading	444
CHAPTER	8 ■ Governance	446
8.1	INTRODUCTION	446
8.1.1	The Landscape	447
8.2	ECONOMIC GROWTH	449
8.3	DISTRIBUTION OF AI	454
8.3.1	Distribution of Access to AI	456
8.3.2	Distribution of Power Among AIs	460
8.4	CORPORATE GOVERNANCE	465
8.4.1	What Is Corporate Governance?	465
8.4.2	Legal Structure	465
8.4.3	Ownership Structure	466
8.4.4	Organizational Structure	467
8.4.5	Assurance	468
8.5	NATIONAL GOVERNANCE	469
8.5.1	Standards and Regulations	469
8.5.2	Liability for AI Harms	471
8.5.3	Targeted Taxation	472
8.5.4	Public Ownership over AI	473
8.5.5	Improving Resilience	473
8.5.6	Not Falling Behind	474
8.5.7	Information Security	475
8.6	INTERNATIONAL GOVERNANCE	477
8.6.1	Forms of International Governance	478
8.6.2	Four Questions for AI Regulation	481
8.6.3	What Can Be Included in International Agreements?	483

8.7	COMPUTE GOVERNANCE	486
8.7.1	Compute Is Indispensable for AI Development and Deployment	486
8.7.2	Compute Is Physical, Excludable, and Quantifiable	488
8.8	CONCLUSION	492
8.9	LITERATURE	494
8.9.1	Recommended Reading	494
<hr/> Acknowledgments		495
<hr/> References		497
<hr/> Index		529



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 (wwwaisafetybook.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