



RESPONSIBLE

AI



BEST PRACTICES for Creating Trustworthy AI Systems

QINGHUA LU

LIMING ZHU

JON WHITTLE

XIWEI XU



FREE SAMPLE CHAPTER |



RESPONSIBLE AI

This page intentionally left blank

RESPONSIBLE AI

BEST PRACTICES FOR CREATING
TRUSTWORTHY AI SYSTEMS

Qinghua Lu, Liming Zhu,
Jon Whittle, and Xiwei Xu

◆◆ Addison-Wesley

Boston • Columbus • New York • San Francisco • Amsterdam • Cape Town • Dubai
London • Madrid • Milan • Munich • Paris • Montreal • Toronto • Delhi • Mexico City
São Paulo Sydney • Hong Kong • Seoul • Singapore • Taipei • Tokyo

Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book, and the publisher was aware of a trademark claim, the designations have been printed with initial capital letters or in all capitals.

The authors and publisher have taken care in the preparation of this book, but make no expressed or implied warranty of any kind and assume no responsibility for errors or omissions. No liability is assumed for incidental or consequential damages in connection with or arising out of the use of the information or programs contained herein.

For information about buying this title in bulk quantities, or for special sales opportunities (which may include electronic versions; custom cover designs; and content particular to your business, training goals, marketing focus, or branding interests), please contact our corporate sales department at corpsales@pearsoned.com or (800) 382-3419.

For government sales inquiries, please contact governmentsales@pearsoned.com.

For questions about sales outside the U.S., please contact intlcs@pearson.com.

Visit us on the Web: informit.com/community

Library of Congress Control Number: 2023915722

Copyright © 2024 Pearson Education, Inc.

Cover image: Shutterstock

All rights reserved. This publication is protected by copyright, and permission must be obtained from the publisher prior to any prohibited reproduction, storage in a retrieval system, or transmission in any form or by any means, electronic, mechanical, photocopying, recording, or likewise. For information regarding permissions, request forms, and the appropriate contacts within the Pearson Education Global Rights & Permissions Department, please visit www.pearson.com/permissions.

No patent liability is assumed with respect to the use of the information contained herein. Although every precaution has been taken in the preparation of this book, the publisher and author assume no responsibility for errors or omissions. Nor is any liability assumed for damages resulting from the use of the information contained herein.

ISBN-13: 978-0-13-807392-3

ISBN-10: 0-13-807392-9

\$PrintCode

Pearson's Commitment to Diversity, Equity, and Inclusion

Pearson is dedicated to creating bias-free content that reflects the diversity of all learners. We embrace the many dimensions of diversity, including but not limited to race, ethnicity, gender, socioeconomic status, ability, age, sexual orientation, and religious or political beliefs.

Education is a powerful force for equity and change in our world. It has the potential to deliver opportunities that improve lives and enable economic mobility. As we work with authors to create content for every product and service, we acknowledge our responsibility to demonstrate inclusivity and incorporate diverse scholarship so that everyone can achieve their potential through learning. As the world's leading learning company, we have a duty to help drive change and live up to our purpose to help more people create a better life for themselves and to create a better world.

Our ambition is to purposefully contribute to a world where

- Everyone has an equitable and lifelong opportunity to succeed through learning
- Our educational products and services are inclusive and represent the rich diversity of learners
- Our educational content accurately reflects the histories and experiences of the learners we serve
- Our educational content prompts deeper discussions with learners and motivates them to expand their own learning (and worldview)

While we work hard to present unbiased content, we want to hear from you about any concerns or needs with this Pearson product so that we can investigate and address them.

Please contact us with concerns about any potential bias at <https://www.pearson.com/report-bias.html>.

The authors have been influenced and inspired by many leading thinkers in the fields of AI, Responsible AI, and related areas. We would like to acknowledge many of these individuals here. This book has in part been shaped by conversations with the following people, writings of theirs we have pored over, or ideas of theirs we have listened to. Thanks to all of them for all the contributions they have made to making AI more responsible.

Colleagues we have worked with:

- John Grundy
- Rashina Hoda
- Waqar Hussain
- Aurelie Jacquet
- Arif Nurwidyantoro
- Gillian Oliver
- Harsha Perera
- Mojtaba Shahin
- Rifat Ara Shams
- Judy Slatyer
- Stela Solar
- Toby Walsh
- Chen Wang
- Zhenchang Xing
- Didar Zowghi

Leaders we have been inspired by:

- Yoshua Bengio
- Nick Bostrom
- Joy Buolamwini
- Kate Crawford
- Virginia Dignum
- Kay Firth-Butterfield
- John C. Havens
- Dan Hendrycks
- Ben Hutchinson
- Tim Hwang
- Fei-Fei Li
- Margaret Mitchell
- Andrew Ng
- Bashar Nuseibeh
- Meredith Ringel Morris
- Francesca Rossi
- Stuart Russell
- Ed Santow
- Iyad Rahwan
- Irene Solaiman
- Max Tegmark
- Cat Wallace
- Jennifer Wortman Vaughan

Contents

	Preface	xv
	About the Author	xix
Part I	Background and Introduction	1
	1 Introduction to Responsible AI	3
	What Is Responsible AI?	4
	What Is AI?	6
	Developing AI Responsibly: Who Is Responsible for Putting the “Responsible” into AI?	8
	About This Book	9
	How to Read This Book	11
	2 Operationalizing Responsible AI: A Thought Experiment—Robbie the Robot	13
	A Thought Experiment—Robbie the Robot	13
	Who Should Be Involved in Building Robbie?	14
	What Are the Responsible AI Principles for Robbie?	16
	Robbie and Governance Considerations	18
	Robbie and Process Considerations	19
	Robbie and Product Considerations	21
	Summary	22
Part II	Responsible AI Pattern Catalogue	23
	3 Overview of the Responsible AI Pattern Catalogue	25
	The Key Concepts	25
	The Multifaceted Meanings of <i>Responsible</i>	25
	Varied Understandings of <i>Operationalization</i>	26
	The Duality of Trust and Trustworthiness	28
	Why Is Responsible AI Different?	30
	A Pattern-Oriented Approach for Responsible AI	32

4	Multi-Level Governance Patterns for Responsible AI	39
	Industry-Level Governance Patterns	42
	G.1. RAI Law and Regulation	42
	G.2. RAI Maturity Model	44
	G.3. RAI Certification	46
	G.4. Regulatory Sandbox	48
	G.5. Building Code	50
	G.6. Independent Oversight	51
	G.7. Trust Mark	53
	G.8. RAI Standards	55
	Organization-Level Governance Patterns	56
	G.9. Leadership Commitment for RAI	57
	G.10. RAI Risk Committee	58
	G.11. Code of RAI	60
	G.12. RAI Risk Assessment	62
	G.13. RAI Training	64
	G.14. Role-Level Accountability Contract	66
	G.15. RAI Bill of Materials	68
	G.16. Standardized Reporting	70
	Team-Level Governance Patterns	72
	G.17. Customized Agile Process	72
	G.18. Tight Coupling of AI and Non-AI Development	74
	G.19. Diverse Team	75
	G.20. Stakeholder Engagement	77
	G.21. Continuous Documentation Using Templates	79
	G.22. Verifiable Claim for AI System Artifacts	80
	G.23. Failure Mode and Effects Analysis (FMEA)	82
	G.24. Fault Tree Analysis (FTA)	84
	Summary	85

5	Process Patterns for Trustworthy Development Processes	87
	Requirements	88
	P.1. AI Suitability Assessment	89
	P.2. Verifiable RAI Requirement	90
	P.3. Lifecycle-Driven Data Requirement	92
	P.4. RAI User Story	94
	Design	96
	P.5. Multi-Level Co-Architecting	96
	P.6. Envisioning Card	98
	P.7. RAI Design Modeling	99
	P.8. System-Level RAI Simulation	101
	P.9. XAI Interface	103
	Implementation	105
	P.10. RAI Governance of APIs	105
	P.11. RAI Governance via APIs	107
	P.12. RAI Construction with Reuse	108
	Testing	110
	P.13. RAI Acceptance Testing	110
	P.14. RAI Assessment for Test Cases	112
	Operations	114
	P.15. Continuous Deployment for RAI	114
	P.16. Extensible, Adaptive, and Dynamic RAI Risk Assessment	116
	P.17. Multi-Level Co-Versioning	118
	Summary	120
6	Product Patterns for Responsible-AI-by-Design	121
	Product Pattern Collection Overview	122
	Supply Chain Patterns	123
	D.1. RAI Bill of Materials Registry	123
	D.2. Verifiable RAI Credential	126

	D.3. Co-Versioning Registry	129
	D.4. Federated Learner	132
	System Patterns	134
	D.5. AI Mode Switcher	134
	D.6. Multi-Model Decision-Maker	137
	D.7. Homogeneous Redundancy	139
	Operation Infrastructure Patterns	141
	D.8. Continuous RAI Validator	141
	D.9. RAI Sandbox	144
	D.10. RAI Knowledge Base	146
	D.11. RAI Digital Twin	148
	D.12. Incentive Registry	151
	D.13. RAI Black Box	153
	D.14. Global-View Auditor	156
	Summary	158
7	Pattern-Oriented Reference Architecture for Responsible-AI-by-Design	159
	Architectural Principles for Designing AI Systems	160
	Pattern-Oriented Reference Architecture	161
	Supply Chain Layer	162
	System Layer	163
	Operation Infrastructure Layer	164
	Summary	165
8	Principle-Specific Techniques for Responsible AI	167
	Fairness	167
	T.1. Fairness Assessor	168
	T.2. Discrimination Mitigator	170
	Privacy	172
	T.3. Encrypted-Data-Based Trainer	173
	T.4. Secure Aggregator	174
	T.5. Random Noise Data Generator	176

Explainability	178
T.6. Local Explainer	178
T.7. Global Explainer	180
Summary	182
Part III Case Studies	183
9 Risk-Based AI Governance in Telstra	185
Policy and Awareness	186
Telstra’s Definition of AI	186
Awareness	187
Assessing Risk	188
Dimensions of Risk	188
Levels of Risk	189
Operation of the Risk Council	190
Learnings from Practice	192
Identifying and Registering Use Cases	192
Support from Technology Tools	193
Governance over the Whole Lifecycle	193
Scaling Up	194
Future Work	195
10 Reejig: The World’s First Independently Audited Ethical Talent AI	197
How Is AI Being Used in Talent?	198
Aggregating Siloed Data Across Multiple Sources	198
Providing Decision-Making Support at Scale	199
Reducing Unconscious Bias	199
What Does Bias in Talent AI Look Like?	200
Data Bias	200
Human Bias	200
Regulating Talent AI Is a Global Issue	201
US Legislation Being Introduced	201

European Legislation Being Introduced	202
Reejig’s Approach to Ethical Talent AI	202
Debiasing Strategies	202
How Ethical AI Evaluation Is Done: A Case Study in Reejig’s World-First Independently Audited Ethical Talent AI	204
Overview	204
The Independent Audit Approach	204
A Summary of the Results	205
Recognition and Impact	205
Project Overview	206
About the Reejig Algorithm	206
The Objectives of the Project	206
The Approach	206
The Ethical AI Framework Used for the Audit	207
Ethical Principles	207
Ethical Validation	208
Functional Validation	209
The Benefits of Ethical Talent AI	210
Building Stronger and More Diverse Teams by Removing Bias	210
Maintaining Privacy and Security	211
Demonstrating Leadership Against Competitors	211
Reejig’s Outlook on the Future of Ethical Talent AI	211
Reejig Has Led the Way in AI Ethics from Day One	211
A New Independent Audit of Reejig Is Already Underway	212
The Future of Workforce AI Will Unlock Zero Wasted Potential	212
11 Diversity and Inclusion in Artificial Intelligence	213
Importance of Diversity and Inclusion in AI	215
Definition of Diversity and Inclusion in Artificial Intelligence	216
Guidelines for Diversity and Inclusion in Artificial Intelligence	219
Humans	219
Data	222

Process	226
System	231
Governance	232
Conclusion	234
Human	235
Data	235
Process	235
System	235
Governance	236
Part IV Looking to the Future	237
12 The Future of Responsible AI	239
Regulation	241
Education	242
Standards	244
Tools	245
Public Awareness	246
Final Remarks	246
Part V Appendix	249
Index	271

This page intentionally left blank

Preface

Writing a book is not a minor undertaking. The authors of this book know this from experience: Collectively, we've labored through the trials of book-writing multiple times, and we've also had many failed attempts and false starts along the way. When an idea for a new book comes along, then, it is a brave soul that not only agrees to write it, but immerses him/herself into it, body and soul. Nevertheless, that's what happened with this book. The topic of Responsible AI is so important to society, so topical in the current zeitgeist, and so needed, that it would be folly not to take on the challenge.

But this is not just any book on Responsible AI. There are quite a few books covering the topic already. Some of these are deeply technical books that look only at the technology aspects of AI systems, such as how to manipulate AI models and their data to try and ensure responsible AI. Other books are more philosophical in nature, citing examples where AI has already had an unfortunate impact on society, or exploring what terrible potentials of AI there are in the future.

This book, however, lies somewhere in the middle. It fills a gap between the highly technical advice and more philosophical thinking. It aims to provide concrete guidance to the AI practitioner, to the AI development teams, and to those who care about governing AI systems when they are developed, such as senior managers and boards. The emphasis is on *concrete* guidance. There are many AI ethics principles available nowadays, but there is still a lack of information on how to convert these principles into practice. This book, which can be thought of as a reference volume, provides a set of tried and tested patterns for doing just that. We gathered these patterns from an in-depth search of existing literature and practice: We didn't make them up, but bring together solutions that have been tried out in anger and we collect them in one place.

We hope that this book will serve its purpose, to inform and guide the reader toward responsible AI. Indeed, in a perfect world, our book—like all good reference books—will sit on the reader's shelf many years into the future, to be picked up when a reminder is needed of how best to handle a particular ethical issue in AI. There's no doubt that AI will evolve significantly and rapidly in the coming years. But the fundamentals of how to design, implement and use systems responsibly are somewhat more stable. And so, we also hope that these patterns, although undoubtedly they will be added to over the years, will stand the test of time.

Register your copy of *Responsible AI: Best Practices for Creating Trustworthy AI Systems* at informit.com for convenient access to downloads, updates, and corrections as they become available. To start the registration process, go to informit.com/register and log in or create an account. Enter the product ISBN 9780138073923 and click Submit. Once the process is complete, you will find any available bonus content under "Registered Products."

This page intentionally left blank

Acknowledgments

Writing a book is a long journey that requires the assistance and support of many. We would like to thank Stuart Powell, Shujia Zhang, Didar Zowghi, and Francesca da Rimini for their contributions to the case studies. We also appreciate Judy Slatyer, Aurelie Jacquet, Yue Liu, Boming Xia, and Pamela Finckenberg-Broman for their help with the pattern chapters.

Pearson did a professional and efficient job in the production process. This book has greatly benefited from their expertise.

We are grateful to the management of CSIRO's Data61. Without their generous support, this book would not have been written.

This page intentionally left blank

About the Authors

Dr. Qinghua Lu is a principal research scientist and leads the Responsible AI science team at CSIRO's Data61. She received her PhD from University of New South Wales in 2013. Her current research interests include responsible AI, software engineering for AI/GAI, and software architecture. She has published 150+ papers in premier international journals and conferences. Her recent paper titled "Towards a Roadmap on Software Engineering for Responsible AI" received the ACM Distinguished Paper Award. Dr. Lu is part of the OECD.AI's trustworthy AI metrics project team. She also serves a member of Australia's National AI Centre Responsible AI at Scale think tank. She is the winner of the 2023 APAC Women in AI Trailblazer Award.

Dr./Prof. Liming Zhu is a Research Director at CSIRO's Data61 and a conjoint full professor at the University of New South Wales (UNSW). He is the chairperson of Standards Australia's blockchain committee and contributes to the AI trustworthiness committee. He is a member of the OECD.AI expert group on AI Risks and Accountability, as well as a member of the Responsible AI at Scale think tank at Australia's National AI Centre. His research program innovates in the areas of AI/ML systems, responsible/ethical AI, software engineering, blockchain, regulation technology, quantum software, privacy, and cybersecurity. He has published more than 300 papers on software architecture, blockchain, governance and responsible AI. He delivered the keynote "Software Engineering as the Linchpin of Responsible AI" at the International Conference on Software Engineering (ICSE) 2023.

Prof. Jon Whittle is Director at CSIRO's Data61, Australia's national centre for R&D in data science and digital technologies. With around 850 staff and affiliates, Data61 is one of the largest collections of R&D expertise in Artificial Intelligence and Data Science in the world. Data61 partners with more than 200 industry and government organisations, more than 30 universities, and works across vertical sectors in manufacturing, health, agriculture, and the environment. Prior to joining Data61, Jon was Dean of the Faculty of Information Technology at Monash University.

Dr. Xiwei Xu is a principal research scientist and the group leader of the software systems research group at Data61, CSIRO. With a specialization in software architecture and system design, she is at the forefront of research in these fields. Xiwei is identified by the Bibliometric Assessment of Software Engineering Scholars and Institutions as a top scholar and ranked 4th in the world (2013–2020) as the most impactful SE researchers by JSS (*Journal of Systems and Software*), a well-recognized academic journal in software engineering research.

Credits

Cover: everything possible/Shutterstock

Figure 3-1, Figure 4-1, Figure 4-2, Figure 5-1, Figure 5-2, Figure 6-1–6-14, Figure 7-1, FIG11-1: CSIRO

Figure 9-1: Telstra

Figure 10-1: Reejig

PART I

BACKGROUND AND INTRODUCTION

What is responsible artificial intelligence (AI)? Why do we need responsible AI? How complex is the operationalization of responsible AI? How many different perspectives need to be taken into account? These are the questions we answer in Part I.

In Chapter 1, “Introduction to Responsible AI,” we introduce the history and motivation of responsible AI and give a definition from a systems perspective. We also discuss who should be responsible for responsible AI.

In Chapter 2, “Operationalizing Responsible AI: A Thought Experiment—Robbie the Robot,” we go through a thought experiment using Robbie the Robot to explain what organizations need to think about when it comes to responsible AI, including governance considerations, process considerations, and product considerations.

This page intentionally left blank

2

Operationalizing Responsible AI: A Thought Experiment—Robbie the Robot

Before we delve into the details of how to operationalize responsible AI principles, this chapter presents an example, designed to illustrate the complexity of responsible AI, and the broad range of stakeholders that need to be involved in the process. We hope you will find this example both fun and illustrative.

A Thought Experiment—Robbie the Robot

To illustrate just how complex the operationalization of responsible AI principles is—and how many different perspectives need to be taken into account—let’s walk through a thought experiment. For this experiment, we use Robbie the Robot, the nonspeaking robot introduced by Dr. Susan Calvin in Isaac Asimov’s classic book *I, Robot*.

Robbie is a children’s robot, designed to play with and take care of kids. Without the ability to speak, Robbie finds other ways to communicate. As Susan Calvin says in the prelude to the chapter on Robbie: “Robbie had no voice. He was a nonspeaking robot. Robbie was made to take care of children. He was a nanny...”¹

1. I. Asimov, *I, Robot* (Gnome Press, 1950), 1st Edition, page 11, 2 December 1950.

The first chapter in *I, Robot* then goes on to tell a story of a young girl, Gloria, and her friendship with Robbie. We first find them playing hide-and-seek in Gloria's garden. Gloria is incredibly fond of Robbie, remarking at one point in the chapter: "He was *not* only a machine. He was my *friend!*"² But Gloria's mother, Mrs. Weston, is suspicious of Robbie. Although Robbie has been with the family for two years—and there have been no issues—Mrs. Weston gradually starts to worry that Robbie might do something unexpected, and might even harm Gloria.

"I don't want a machine to take care of my daughter. Nobody knows what it's thinking." She tells her husband. And then: "I wasn't worried at first. But something might happen and that...that thing will go crazy and..."³

In the end, Mrs. Weston sends Robbie back to the manufacturer, US Robots. This action upsets Gloria, who really misses him. To try to show Gloria that Robbie is just "some pieces of metal with electricity," Mr. and Mrs. Weston take Gloria to the factory where Robbie was made and is now being used to manufacture other robots. Things don't go according to plan, however. When Gloria sees Robbie, she runs toward him, not noticing a huge tractor on the factory floor, which would have run her over were it not for Robbie, who, seeing Gloria in danger, rescues her. Mr. and Mrs. Weston are forced to take Robbie back to the house, and Gloria is reunited with her best friend.

Although the story of Robbie was originally published in 1940, and predicted a future where children would have robot nannies, we still don't. And to create one remains fiendishly difficult, both from a technical perspective (we still struggle to get robots to carry out seemingly simple tasks such as playing a game of hide-and-seek) and from a perspective of responsibility (how can Mrs. Weston be confident that Robbie won't go "crazy" and hurt her daughter?). The trope of kids befriending robots has since been explored extensively in popular entertainment, in movies such as *The Iron Giant*, *Big Hero 6*, and *Earth to Echo*. In many of these stories, the robot AI does indeed go "crazy" and bad things happen; the recent movie *M3GAN* is a good example in the horror genre.

Who Should Be Involved in Building Robbie?

In the remainder of this chapter, we use Robbie the Robot as an example to consider where responsible AI issues come up. Let's consider things from the perspective of US Robots, the company that created Robbie.

As discussed earlier in this chapter, a diverse set of stakeholders need to be involved in building, using, and managing an AI system such as Robbie the Robot. Each stakeholder has knowledge that will contribute to making sure that Robbie is designed responsibly. Table 2.1 lists some of the stakeholders that US Robots should include, as well as the key contributions each of these stakeholders can make when it comes to designing Robbie in a responsible way.

2. I. Asimov, *I, Robot* (Gnome Press, 1950), 1st Edition, page 16, 2 December 1950.

3. I. Asimov, *I, Robot* (Gnome Press, 1950), 1st Edition, page 15, 2 December 1950.

Table 2.1 Stakeholders of Robbie

Stakeholder Type	Role	Responsible AI Contribution	Type of Contribution
US Robots Company Board	To manage the reputation and market risks of developing Robbie	Ensures that a risk management framework is set up and monitored to assess, mitigate, and manage risks associated with Robbie's deployment in family settings	Governance
Government	To ensure the safety of the general public	Enacts laws that regulate how family robots are designed, manufactured, and used	Governance
Industry Bodies	To produce standards for robotics companies to follow	Creates standards for family robots that member organizations agree to follow	Governance
Parent Groups	To advise parents on the use of Robbie and to lobby government on appropriate legislation	Sets up information sharing for parents, e.g., workshops, web portals	Governance
VP Ethics	US Robots executive who ensures the company has a reputation for responsible AI	Defines and rolls out training and practices for responsible AI across the company; may include independent testing of Robbie features before release	Process
COO	US Robots executive responsible for effective and efficient processes within the company	Implements recommendations from the VP Ethics to ensure company practices include responsible AI considerations	Process
Product Manager	Team member who makes decisions as to which features go into Robbie (e.g., what its objectives are, what the constraints are)	Sets up and manages a process to get customer input on desired features and works with technical experts on feasibility	Process
Project Manager	Team member who manages the development of Robbie over time, ensuring delivery of features according to an agreed-upon schedule	Ensures that the project plan includes key check-in points to consider issues related to responsible AI	Process
Technical Manager	Team member who manages the technical teams developing Robbie to deliver agreed-upon features	Ensures best-practice responsible AI guidelines (coding practices, appropriate use of off-the-shelf components) are used	Product
Data Scientist	Team member who manages the data that Robbie is trained on to carry out tasks	Ensures, as far as possible, that training data is representative of the broader population and not biased to one segment of society	Product
AI Expert	Team member who develops AI models to process data	Monitors AI model performance for bias	Product
Software Engineer	Team member who manages the integration of Robbie into US Robots' larger software systems as required	Ensures that best-practice responsible AI software patterns are used	Process
General Public	Users of Robbie	Provides feedback to US Robots to ensure that any responsible AI issues are fixed	Governance
Suppliers	Other manufacturers or AI technology/solution providers	Ensures the supplied product components are without any responsible AI issues	Governance

As Table 2.1 shows, responsible AI is complex: many stakeholders need to be involved. The good news, however, is that this is no different to any complex systems engineering task. Building skyscrapers, flying airplanes, implementing large-scale government information systems—these are all examples of complex engineering projects that society operates routinely today. And, over time, society has agreed upon sets of rigorous processes and methods to ensure that such systems are safe, secure, and operate as expected. The only difference with responsible AI is that AI is a fast-moving technology, so we do not yet have a full set of rigorous practices. (This book, of course, partially fills that gap!)

What Are the Responsible AI Principles for Robbie?

The first step in ensuring that Robbie implements AI responsibly is for US Robots to agree to a high-level set of responsible AI principles. These could be Australia's AI Ethics Principles, as described in Chapter 1, or they could be something company- or context-specific. In his book, Asimov famously captured the operating principles of US Robots as the Three Laws of Robotics, codified in the *Handbook of Robotics, 2058 AD*:

1. A robot must not harm a human. And it must not allow a human to be harmed.
2. A robot must obey a human's order, unless that order conflicts with the First Law.
3. A robot must protect itself, unless this protection conflicts with the First or Second Laws.⁴

These Robot Laws were encoded in Robbie's positronic brain to ensure that they would be followed. For a modern engineering firm creating a robot like Robbie, these laws could well serve as high-level principles to follow. But to encode them in the design and operation of a robot, they need to be made more concrete (i.e., the laws must be operationalized).

To some extent, Asimov's laws can be related to modern AI ethics principles. Table 2.2, for example, maps them to Australia's AI Ethics Principles. Note that some of Asimov's laws map in a fairly straightforward manner. It becomes quickly clear, however, that Asimov's laws are actually quite narrow. Other than the safety of humans, they say nothing about what is considered societally appropriate behavior by Robbie. For example, one would expect Robbie, as a child's companion, to act and teach in a way that is considered proper. In modern-day AI systems, in contrast, there is a lot of concern about whether AI systems will exhibit behavior that is discriminatory, biased, unfair, or socially unacceptable. None of this concern is captured in Asimov's laws. Arguably, this kind of behavior could be included under the First Law, but this depends on the definition of *harm*, which in Asimov's book is largely focused on physical safety.

4. I. Asimov, *I, Robot* (Gnome Press, 1950), 1st Edition, page 9, 2 December 1950.

Table 2.2 Mapping Asimov's Laws to Australia's AI Ethics Principles

AI Ethics Principle	Description	Sample Problematic Behaviors in Robbie Context	Covered by Asimov's Laws?
Human, societal, and environmental well-being	AI systems should benefit individuals, society, and the environment.	Robbie causes problems with children, such as child safety or psychological dependency.	Partially—First Law covers physical safety but not broader well-being issues.
Human-centered values	AI systems should respect human rights, diversity, and the autonomy of individuals.	Robbie does not encourage children to assert their right of freedom of speech.	No—Laws say nothing about human rights.
Fairness	AI systems should be inclusive and accessible, and should not involve or result in unfair discrimination against individuals, communities, or groups.	Robbie treats children from different backgrounds differently.	No—Laws say nothing about diversity in end users.
Privacy protection and security	AI systems should respect and uphold privacy rights and data protection, and ensure the security of data.	Robbie collects data from the child and shares with the company.	No—Laws do not cover privacy.
Reliability and safety	AI systems should reliably operate in accordance with their intended purpose.	Robbie fails to rescue Gloria from the tractor due to a malfunction.	Partially—Second Law somewhat covers "intended purpose" but does not explicitly address malfunctions.
Transparency and explainability	There should be transparency and responsible disclosure so people can understand when they are being significantly impacted by AI, and can find out when an AI system is engaging with them.	Robbie fails to tell Mrs. Weston that he has been cheating Gloria at hide-and-seek.	Partially—Second Law guarantees that Robbie explains his actions but only if explicitly asked.
Contestability	When an AI system significantly impacts a person, community, group, or environment, there should be a timely process to allow people to challenge the use or outcomes of the AI system.	Mrs. Weston is unable to get Robbie to teach Gloria in a way that she wants.	Partially—Second Law guarantees a challenge of Robbie's outcomes but only if explicitly asked.
Accountability	People responsible for the different phases of the AI system lifecycle should be identifiable and accountable for the outcomes of the AI systems, and human oversight of AI systems should be enabled.	US Robots fails to put appropriate procedures in place to ensure Robbie follows the three laws.	No.

Robbie and Governance Considerations

Putting aside Asimov's Laws for a moment, as they are clearly incomplete for our purposes, let's move forward assuming the AI ethics principles in Table 2.2 are our driver.

Table 2.1 identifies six stakeholders relevant to Governance. Let's consider just one of these, the company board. Like any board, the main purpose of the board of US Robots is to set the strategic direction of the company and to ensure that the company is operating within all relevant laws, ethically, and in a way that safeguards the reputation and financial sustainability of the company.

Imagine, then, the position of the CEO of US Robots. She's just had a brilliant idea: to create a new robot, which will be called Robbie, that will act as a child's nanny. It could be a big money-spinner for the company and could really place US Robots on the map as a global leader in robotics technologies. The only remaining question is what will the board think? In many ways, the board's main job is to think about what can go wrong and make sure that the CEO has a plan to deal with any potential threats. In the case of Robbie, the board can imagine a *lot* that can go wrong. Robbie could accidentally hurt a child; he's a heavy piece of metal, after all, and could easily put one of his heavy metal feet in the wrong place. Or Robbie could inflict psychological damage on a child by inadvertently creating an emotional dependency. How will Robbie protect children from harm caused by others? Are Robbie's computer vision systems good enough to identify all harmful objects correctly, or will he miss one? Robbie can't speak, so there is less risk that he will fill the child's head with inappropriate thoughts, but there's still a risk of not being inclusive; he'll need to be programmed with all the different customs and traditions of children from different ethnic and religious backgrounds. And what if Robbie breaks the law? Will the company ultimately be responsible? What HR practices should the board ensure are in place to reprimand engineers who build the wrong mechanisms into Robbie?

It isn't the board's job to provide answers to all of these questions. That is the CEO's job. The board, however, needs to make sure that the questions are asked—and that someone has the answers.

Fortunately, the board is a sophisticated one. Board members gather all the relevant experts together and come up with a plan of action. The board directs the CEO to do the following:

- Develop a responsible AI risk assessment (see G.12. RAI Risk Assessment). One way to do this is to start with the AI ethics principles in Table 2.2 and then imagine all the things that can go wrong. Each of them represents a risk; the board agrees to a risk likelihood and impact severity in each case, and considers mitigation actions that can be put in place to reduce the overall risk rating.
- Introduce ethics training across Project Robbie (see G.13. RAI Training). The board is aware that their workforce is diverse. It includes graduates fresh out of college who are up to speed with the latest technological developments but, as primarily technical specialists, may not have any background or training in the social impacts of technology. The company also includes many staff who have worked for the company for years; they have a good sense of the company's core customer needs but may not be up to speed on the latest technological developments and, in particular, the ethical risks associated with them. So, the board decides that everyone working on Project Robbie should undergo mandatory ethics training.

- Set up an ethics committee as a subcommittee of the board (see G.10. RAI Risk Committee). The board realizes that it has too many things to worry about to leave ethics to the board itself. So it delegates responsibility to an ethics committee, whose job is to oversee the implementation of Robbie in a responsible way. But to make sure that the board has visibility and remains accountable, the ethics committee will be composed of a subset of board members and will be chaired by the most relevant board member. It is at this point that the board members realize they do not have enough ethics expertise on the board, so they go back and revise their board skills matrix to include ethics, and the chair goes out to recruit a new board member with the requisite experience who can chair the subcommittee. No work on Project Robbie will commence until this is done.

The CEO explains to the board that Project Robbie is complex, at a scale unlike anything the company has tackled before. “We can’t build Robbie by ourselves,” the CEO explains, and she goes on to explain that US Robots will need to procure components of Robbie from other providers. The board agrees, but to ensure that Robbie remains an exemplar of responsible AI, the board insists that all acquired components go through a rigorous responsible AI evaluation process before considering their use, including how they will interact with other components (see G.15. RAI Bill of Materials).

The board is happy with its decisions. It’s been a busy few weeks for board members, figuring out how all of this is going to work, but they are content with the outcome. They are happy to support this new idea from the CEO, and they agree that it could be a new future for the company. But they are also as confident as they can be that Robbie will be developed in a responsible way and that, in particular, there won’t be any adverse events that will come back to haunt the company.

The latest board meeting is about to finish. Everyone is happy. Until, almost as an afterthought, the CEO raises a question.

“Have we done enough?” she asks the chair.

“What do you mean? We’re implementing all these measures.”

“Yes,” continues the CEO. “But are they enough? Is there more we can do?”

“I can’t think of anything,” says another board member.

The board chair reflects for a moment and then, like the wise experienced executive that she is, she says: “I can’t think of anything either. But that doesn’t mean there isn’t anything. Maybe we are just not seeing it. Let’s do two things. First, we’ll get an independent review of our plan by experts in the field to make sure it holds water. Second, we’ll have a quarterly review at board meetings to make sure it’s working and there’s nothing we’re forgetting.”

The board meeting ends, and the exciting work on creating Robbie, the children’s nanny robot, begins.

Robbie and Process Considerations

Is the company’s work on responsible AI done? After all, the board and the CEO have put in place rigorous mechanisms to assess and track the risks associated with Robbie’s development. Things should be fine, right?

Of course, the company's work is far from done. In fact, it is just beginning. Governance considerations have been taken care of, but what about process issues? The CEO summons her VP Ethics and COO.

"I have some exciting news," starts the CEO. "The board has just approved that we can go ahead with Robbie!"

"That's fantastic," says the VP Ethics. "But now we have some *real* work to do."

The CEO, VP Ethics, and COO agree to put a working group together, containing key experts and stakeholders from across the company, to define a process approach to developing Robbie. It takes a few months, and some in the company are frustrated that development on Robbie can't start until the process considerations are resolved, but the CEO is firm: "We must get the processes right before starting."

The working group reports back to the CEO, who takes the recommendations to the board. Recommendations include

- **Verifiable Responsible AI Requirements** (see P.2. Verifiable RAI Requirement): The first issue the working group addresses is that the definition of AI Ethics is too vague to measure. The working group's recommendation is that the business analyst team develop a set of verifiable ethical AI requirements. For example, the group says, the ethical AI principle, transparency, could partially be satisfied by a requirement that Robbie includes a parent app where parents can review all Robbie's interactions with their children.
- **A Rigorous Data Lifecycle** (see P.3. Lifecycle-driven Data Requirement): The working group realizes that Robbie needs to respect different cultural traditions (as captured in one of the verifiable ethical AI requirements!). So the group defines a process for careful management of the data lifecycle—what data is collected, how it is managed, who has access, and so on—so that the data loaded into Robbie initially, as well as the way that Robbie collects additional data through sensing, is diverse and treats people from different cultural backgrounds equally.
- **Responsible Design** (see P.7. RAI Design Modeling): The working group also recommends that the responsible AI requirements are considered throughout the design process. They suggest a suite of processes for designing features that ensures the designers put responsible AI first, not let it be an afterthought.
- **Responsible AI Simulation** (see P.8. System-Level RAI Simulation): The working group strongly recommends that the company's simulation platforms, which it currently uses to simulate robotic interactions before deployment, are updated to build in ethical AI considerations. The working group is excited by the prospects here; they suggest using an AI simulator to run what-if scenarios and measure compliance to the verifiable ethical requirements over as many scenarios as possible. "We're using AI to test AI," they muse.
- **Software Engineering Process** (see P.10. RAI Governance of APIs, P.12. RAI Construction with Reuse, P.16. Extensible, Adaptive, and Dynamic RAI Risk Assessment): The working group takes a good look at the company's existing software engineering processes. Group members quickly realize that responsible AI is not built in. So the working group consults with relevant stakeholders and comes up with adaptations to existing engineering processes to make sure

that responsible AI is the primary consideration. Changes include the reuse of AI assets (to ensure that best-practice responsible AI is reused across the development), AI risk assessment at all levels of development (not just done once and forgotten), and a new process for testing Robbie's APIs to ensure there are no privacy leaks.

The board invites the working group to a special meeting of the board, where it runs a rigorous process to test the assumptions and recommendations of the working group. The careful probing of the board leads to some improvements, but, ultimately, the board members are happy. The board chair, however, wants visibility of the process implementation.

"Let's introduce regular review points," she says. "We'll do this quarterly so we can see how well the new process is working out, and if there need to be any changes."

Robbie and Product Considerations

At this point, many of the developers and AI experts within US Robots are getting very excited. They've been hearing about this new robot project for months. There are rumors, but there never seems to be any indication of a timeline for starting work on the project. Until, one day, the CEO sends an internal communication to the teams:

Dear Team,

I am very pleased to inform you that the board has now approved a start date for the development of our latest robot, Robbie. Robbie will be a children's companion robot. It will revolutionize the way that families interact with robots. This is an opportunity to change the world! But we must do this responsibly. And so, we have spent the last few months being rigorous about how we will ensure that Robbie does no harm.

We are now ready to embark on this adventure, and I look forward to working with you all on what will be a challenging but exciting initiative.

US Robots is abuzz with enthusiasm.

But the development teams know there is a lot of hard work ahead. They also know that the first, and most important, consideration is to make sure Robbie is developed ethically. The teams have been undergoing mandatory ethics training for many weeks now. There have been constant communications from the executive team about the importance of responsible AI—not just in the Robbie project, but in all projects. And line managers have asked all their staff to write clear objectives in their annual plans about how they will contribute to responsible AI.

The product manager and project manager for Robbie get together to agree on a way forward. They have been briefed on the new process, with responsible AI built in, that they will follow. But many system-level design decisions still need to be made. And the product and project managers are insistent that these also should put responsible AI first. They decide to do the following:

- Ensure responsible AI is built into Robbie's supply chain (see D.1. RAI Bill of Materials Registry). Robbie's development will be highly dependent on external providers, both of hardware and software components. A project as complex as Robbie can't be delivered by a single

company, even one as large as US Robots. “We need to make sure all external components are developed to the same high standards when it comes to responsible AI,” says the product manager, sensibly.

- Build in *kill switches* at multiple levels (see D.5. AI Mode Switcher). The project manager is concerned that, even if rigorous responsible AI practices are properly followed, situations outside the team’s control may still come up once Robbie is active. “We should build in *kill switches*, both local and remote ones, so that, if anything doesn’t look right, we can shut down different parts of the AI before things get out of hand.”
- Build redundancy into critical AI systems (see D.6. Multi-Model Decision-Maker). The product manager: “Any time that Robbie could potentially put a child in harm’s way—even if that potential is very remote—we should make sure multiple AI models are running in parallel. This will give us confidence that Robbie is only making critical decisions if all the models agree.” The project manager: “We could go further than that, and if the models disagree, activate a *kill switch*.”
- Quarantine new features (see D.9. RAI Sandbox). The product manager: “We’ll need to introduce new features once Robbie is active in society. There’s no way around this; at the very least, it will be needed to fix issues without recalling all versions of Robbie. The project manager agrees and replies, “We should quarantine new features when they are rolled out by isolating it from other critical AI components wherever possible—at least until it’s fully tested in the field.”

Summary

As you can see, when it comes to responsible AI, there is a lot to think about. Responsible AI isn’t the job of a single group of people. Rather, it needs to be embedded at all levels across a company. Neither is responsible AI something you do once and then forget. It is a constant challenge to review and re-review the approach. And, of course, there is a tension between the need to be responsible—and therefore, cautious—and the need to get features out the door and into a product. All of these considerations need to be taken seriously.

The example in this chapter is obviously an idealized scenario. There is no mention of the downsides of introducing governance, process, and product measures to ensure responsible AI. In practice, these measures cost money, and these costs may need to be balanced with the need to get a product out to market—although this, in itself, is an important decision to discuss in the context of responsible AI. One might argue that for-profit companies only care about profit, so many of these measures won’t be implemented. However, public and government opinion about responsible AI is clearly changing. It is becoming a competitive advantage to be responsible. And we are likely to see companies measured for it in the same way that they are measured—either formally through Environmental, Social, Governance (ESG) metrics or informally through reputation—for impacts on society.

Good luck, Robbie! We hope that US Robots has done a good job in building your AI responsibly.

Index

A

- AAAI, Code of Professional Ethics and Conduct, 62
- Accenture, risk committees, 60
- acceptance testing, 110–112
- accountability
 - ethical AI, 208
 - role-level accountability contracts, 66–68
- Active Directory Verifiable Credentials, 129
- ADIA Trust Marks, 54
- Administrative Provisions on Algorithmic Recommendations for Internet Information Services, 71
- Adobe, risk committees, 60
- aggregating
 - secure aggregators, 174–176
 - siloes data across multiple sources, 198–199
- agile processes, customized, 72–73
- AI (Artificial Intelligence)
 - defined, 6–9, 186–187, 198
 - diversity/inclusion, 216, 234
 - data, 217–218, 222–226, 235
 - Equity Fluent Leadership Playbook, 234
 - governance, 218–219, 232–234, 236
 - guidelines, 219
 - humans, 217, 219–222, 235
 - processes, 218, 226–232, 235
 - systems, 218
 - impact of, 198
 - improving customer service (example), 191
 - patterns of AI development
 - governance patterns, 33, 34–35
 - overview, 9–10, 31–32, 33–34
 - process patterns, 33, 36
 - product patterns, 33, 37
 - regulation, 241–242
 - as a spectrum, 7–8
 - speed of development, 240–241
 - system design, architectural principles, 160–161
 - talent/workforce management, 198
 - Telestra
 - definition of AI, 186–187
 - significant AI-informed decisions, 187
- AI Act, 7, 44, 49, 71, 202, 241–242
- AI Bill of Rights, 202
- AI ethics boards
 - Axon, 58
 - IBM, 58
- AI Ethics Principles, 5.0085, 16–17

- AI for Good, 9, 26
- AI Global, Responsible AI Community Portal, The, 148
- AI Regulatory Sandbox, 49
- AI Risk Management Framework, 64, 117
- AI Sandbox, 146
- AI service factsheets, 80
- AI Suitability Toolkit for Nonprofits, 90
- AI/Non-AI development, coupling, 74–75
- AirSim, digital twins, 151
- algorithms
 - Algorithmic Accountability Act of 2022, 44
 - Algorithmic Impact Assessment tool, 64, 117
 - Reejig ethical AI, 206
- Amazon
 - AWS fraud detection, 139
 - metadata tracking, 131
 - model training, 119
 - Rekognition, 108
 - SageMaker
 - continuous validators, 143
 - RAI continuous deployments, 116
 - SageMaker Pipelines, 75
- API governance
 - of API, 105–107
 - via API, 107–108
 - Robbie the Robot thought experiment, 20–21
 - security, 107
- APM, stakeholder engagement, 78
- Apptio Targetprocess, customized agile processes, 73
- architectures
 - architectural principles of system design, 160–161
 - co-architecting AI/non-AI components, 160
 - complexity, 160–161
 - multi-level co-architecting, 96–98
 - pattern-oriented reference architectures, 161
 - reference architectures, 159
 - operation infrastructure layer, 164–165
 - supply chain layer, 162–163
 - system layer, 163–164
 - reusability, 160–161
- Asimov, Isaac
 - I, Robot*, 13–14
 - Three Laws of Robotics, 16–17
- assessing
 - fairness, 168–170
 - RAI assessments for test cases, 112–114
 - risk, 62–64, 116–118
 - Robbie the Robot thought experiment, 18
 - Telestra, 188
- Association for Computing Machinery, computer ethics, 3
- Atola Technology, customized agile processes, 73
- audits
 - global-view auditors, 156–157, 164
 - FG-AI4H audit platform, 157
 - NVIDIA, 157
 - Seclea, 157
 - independent audits
 - Reejig ethical AI, 211–212
 - Reejig ethical AI case study, 204–205
- Australia
 - ADIA Trust Marks, 54
 - AI Ethics Principles, 5–6, 16–17
 - CDR, 242
 - CSIRO
 - Data61’s multidisciplinary and diverse team, diversity/inclusion, 213–214
 - future trends of AI development, 9
 - Responsible AI Question Bank, 39, 64
 - Enhanced Regulatory Sandbox, 50
 - National AI Centre, 6–7, 214, 245
 - National Data Commissioner, 68

- NSW AI Assurance Framework, 64, 117–118
 - Reejig, 197
 - role-level accountability contracts, 68
 - Telestra development, 185
 - UTS, 66, 197
 - Autopilot
 - homogeneous redundancy, 141
 - mode switchers, 136
 - awareness
 - public awareness
 - future of responsible AI, 246
 - Instagram, 246
 - Volkswagen emissions testing, 246
 - Telestra, 186, 187
 - Awesome AI Guidelines, knowledge bases, 148
 - AWS fraud detection, 139
 - Axon, AI ethics board, 58
 - Azure
 - Active Directory Verifiable Credentials, 129
 - Azure Monitor, continuous validators, 143
 - Azure Pipelines, 75
 - DevOps, customized agile processes, 73
 - Machine Learning, 116
- B**
-
- Baidu autonomous mini-buses
 - homogeneous redundancy, 141
 - mode switchers, 137
 - BARD conversational AI interface, 239–240
 - Bell Laboratories, FTA, 85
 - bias
 - assessing fairness, 167–168
 - debiasing strategies, 210
 - Reejig
 - debiasing strategies, 202–203
 - gender debiasing, 202–203
 - talent/workforce management
 - data bias, 200
 - human bias, 200–201
 - unconscious bias, 199–200
 - Bill of Rights, AI, 202
 - Bills of Material Registry, Robbie the Robot
 - thought experiment, 21–22
 - black boxes, 153–155, 164, 180
 - DHT, 155
 - RoBoTIPS, 155
 - blockchains
 - Securekey, 129
 - variable claims, 82
 - Blueprint for an AI Bill of Rights, 44
 - Blueprint for Equity and Inclusion in Artificial Intelligence, A*, 214
 - BMW, code of RAI, 62
 - Boeing, FTA, 85
 - BOM (Bill of Materials), 68–70, 161, 163
 - RAI BOM registry, 123–126
 - SBOM, 126
 - Bosch, code of RAI, 62
 - Boston Consulting Group, maturity models, 46
 - building code, governance, 50–51
- C**
-
- Caine, Mark, 205
 - California, CPRA, 242
 - Canada, Algorithmic Impact Assessment tool, 64, 117
 - case studies
 - Reejig, 183–184
 - algorithms, 206
 - approaches, 206
 - ethical principles, 207–208
 - independent audits, 204–205
 - objectives, 206
 - overview, 204
 - Telestra, 183–184
 - awareness, 186, 187
 - customer service improvements (example), 191

- definition of AI, 186–187
 - development of, 185–186
 - dimensions of risk, 188–189
 - future of, 195
 - identifying/registering use cases, 192–193
 - learnings from practice, 192
 - levels of risk, 189–190
 - lifecycle governance, 193–194
 - policies, 186
 - risk assessments, 188
 - risk council operations, 190–191
 - risk exposure matrix, 189–190
 - scalability, 194–195
 - significant AI-informed decisions, 187
 - tool support, 193
 - CDR (Consumer Data Rights), 242
 - centralized learners, 162–163
 - CEO, governance of responsible AI, 18–19
 - certification, 46–48
 - CertifyAI, certification, 48
 - chapters overview (reader's guide), 10–11
 - ChatGPT
 - continuous validators, 143–144
 - conversational AI interface, 239–240
 - checklists, operationalization, 28
 - China
 - Administrative Provisions on Algorithmic Recommendations for Internet Information Services, 71
 - Cyberspace Administration of China, 71
 - Internet Information Service Algorithmic Management Regulations, 44
 - co-architecting
 - AI/non-AI components, 160
 - multi-level co-architecting, 96–98
 - code building, governance, 50–51
 - Code of Ethics, 3
 - Code of Professional Ethics and Conduct, 62
 - code of RAI, 60–62
 - Codenotary, SBOM, 126
 - Colossus, 3
 - complexity, AI system design, 160–161
 - computer ethics, 3
 - concrete practices, operationalization, 28
 - contents overview (reader's guide), 10–11
 - continuous documentation with templates, 79–80
 - continuous RAI deployments, 114–116
 - continuous validators, 141–144, 164–165
 - conversational AI interfaces
 - BARD, 239–240
 - ChatGPT, 239–240
 - GPT-4, 240–241
 - Tay, 239–240
 - coupling AI/Non-AI development, 74–75
 - co-versioning, 162–163
 - multi-level co-versioning, 118–120
 - registries, 129–132
 - CPRA (California's Privacy Rights Act), 242
 - critical systems redundancy, Robbie the Robot thought experiment, 22
 - CSIRO (Commonwealth Scientific and Industrial Research Organization)
 - Data61's multidisciplinary and diverse team, diversity/inclusion, 213–214
 - future trends of AI development, 9
 - Responsible AI Question Bank, 39, 64
 - customer service, improving with AI, 191
 - customized agile processes, 72–73
 - Cyberspace Administration of China, 71
 - CycloneDX
 - bills of materials, 70
 - RAI BOM registry, 126
- D**
- da Rimini, Francesca. *See* diversity/inclusion
 - DALL-E text-to-image generator, 239
 - DARPA, XAI, 178
 - Dartmouth Workshop, diversity/inclusion, 214
 - data

- bias, talent/workforce management, 200
- diversity/inclusion
 - AI definition of, 217–218
 - guidelines, 222–226, 235
 - governance, 94
 - lifecycles, Robbie the Robot thought experiment, 20
 - lineage, Pachyderm, 131
- DataBricks, MLflow Model Registry, 119, 131
- debiasing strategies, 202–203, 210
- decisions
 - decision-making support, talent/workforce management, 199
 - significant AI-informed decisions, 187
- Deloitte, stakeholder engagement, 78
- Dependency-Track
 - bills of materials, 70
 - RAI BOM registry, 126
- deployments, continuous RAI, 114–116
- descriptive responsibilities, 25–26
- design
 - modeling
 - Robbie the Robot thought experiment, 20
 - trustworthy development processes, 99–101
 - system-level simulation, 101–103
 - trustworthy development processes, 96
 - design modeling, 99–101
 - envisioning cards, 98–99
 - XAI interface, 103–105
- development of AI, speed of development, 240–241
- development processes, 87–88
 - design, 96
 - design modeling, 99–101
 - envisioning cards, 98–99
 - system-level simulation, 101–103
 - XAI interface, 103–105
 - implementation, 105
 - RAI construction with reuse, 108–110
 - RAI governance of API, 105–107
 - RAI governance via API, 107–108
 - operations, 114
 - multi-level co-versioning, 118–120
 - RAI continuous deployments, 114–116
 - risk assessments, 116–118
 - requirements
 - lifecycle-driven data requirements, 92–94
 - suitability assessments, 89–90
 - user stories, 94–95
 - verifiable RAI requirements, 90–92
 - testing, 110–112
- DHT (Distributed Hash Tables), 155
- DI-AI guidelines. *See* diversity/inclusion
- Differential Privacy Library
 - Google, 177
 - IBM, 177
- digital humanism. *See* responsible AI
- digital twins, 148–151, 164–165
 - AirSim, 151
 - NVIDIA DRIVE Sim, 151
 - rfProis, 151
- Digital.ai, role-level accountability
 - contracts, 68
- discrimination mitigators, 170–172
- diverse teams, 75–77
- diversity/inclusion, 234
 - 2022 Diversity Annual Report, 77
 - AI definition of, 216
 - data, 217–218
 - governance, 218–219
 - humans, 217
 - processes, 218
 - systems, 218
 - Blueprint for Equity and Inclusion in Artificial Intelligence, A*, 214
 - CSIRO Data61’s multidisciplinary and diverse team, 213–214
 - Dartmouth Workshop, 214
 - Equity Fluent Leadership Playbook, 234

- ethical AI, 203, 210
 - guidelines, 219
 - data, 222–226, 235
 - governance, 232–234, 236
 - humans, 219–222, 235
 - processes, 226–232, 235
 - importance of, 215–216
 - Meta, 77
 - Microsoft, 77
 - National Artificial Intelligence Centre’s Think Tank on Diversity and Inclusion in AI, Australia, 214
 - Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*, 214
 - DNN (Deep Neural Networks), 180
 - documentation, continuous documentation with templates, 79–80
 - DPTM (Data Protection Trustmarks), 54
 - DRFBM (Design Review Based on Failure Modes), 83
 - DVC (Data Version Control), 119–120, 131
- E**
- ECCOLA, user stories, 95
 - education, future of responsible AI, 242–244
 - encryption
 - data-based trainers, 173–174
 - homomorphic encryption, 173, 174
 - Enhanced Regulatory Sandbox, 50
 - envisioning cards, 98–99
 - Equity Fluent Leadership Playbook, 236
 - Ethereum Verifiable Claims, blockchain variable claims, 82
 - ethical AI, 197
 - accountability, 208
 - benefits of, 210–211
 - debiasing strategies, 210
 - diversity/inclusion, 210
 - ethical validation, 208–209
 - fairness, 207
 - functional validation, 209–210
 - future of, 212
 - leadership, 211
 - principles of, 207–208
 - privacy, 208, 211
 - Reejig, 202
 - aggregating siloed data across multiple sources, 199
 - algorithms, 206
 - approaches, 206
 - Caine, Mark, 205
 - case studies, 204–210
 - debiasing strategies, 202–203
 - diversity/inclusion, 203
 - ethical principles, 207–208
 - ethical validation, 208–209
 - functional validation, 209–210
 - gender debiasing, 202–203
 - IEEE Symposium on Computational Intelligence, 205
 - independent audits, 204–205, 211–212
 - objectives, 206
 - World Economic Forum, 205
 - Zero Wasted Potential, 212
 - regulation, 201
 - AI Act, 202
 - AI Bill of Rights, 202
 - EU legislation, 202
 - NYC Local Law 144, 201–202
 - U.S. legislation, 201–202
 - security, 208, 211
 - transparency, 207
 - University of Helsinki, 66
 - Ethical AI: from Principles to Practice, 66
 - EthicalML-XAI, 182
 - ethics
 - AI ethics boards
 - Axon, 58
 - IBM, 58
 - Australian AI Ethics Principles, 5–6, 16–17
 - BMW code of RAI, 62
 - Bosch code of RAI, 62

- Code of Ethics, 3
 - Code of Professional Ethics and Conduct, 62
 - computer ethics, 3
 - Ethics of AI: Safeguarding Humanity, 66
 - Wiener, Norbert, 3
 - EU (European Union)
 - AI Act, 7, 44, 49, 71, 202, 241–242
 - AI Regulatory Sandbox, 49
 - ethical AI legislation, 202
 - GDPR, 71, 241, 242
 - EULA, Viz.ai, 68
 - explainability, 178–182
 - DARPA, 178
 - EthicalML-XAI, 182
 - global explainers, 180–182
 - Google Vertex Explainable AI, 180
 - IBM AI Explainability 360, 180
 - local explainers, 178–180
 - Microsoft InterpretML, 180, 182
 - RISE, 178
 - tf-explain, 182
 - Explainable AI (XAI), 103–105, 178
- F**
-
- Facebook Fairness Flow, 170
 - failure modes, FMEA, 82–84
 - FairLearn, 172
 - fairness, 167–168
 - assessing, 168–170
 - bias, 167–168
 - discrimination mitigators, 170–172
 - ethical AI, 207
 - Facebook Fairness Flow, 170
 - Google Fairness Gym, 170
 - Google Fairness Indicators, 170
 - IBM AI Fairness 360, 172
 - LiFT, 172
 - Microsoft FairLearn, 172
 - fallback loops, 164
 - Fastcase AI Sandbox, 146
 - FATE, federated learners, 134
 - features (Robbie the Robot thought experiment), quarantining, 22
 - federated learners, 132–134, 162–163, 176
 - FG-AI4H audit platform, 157
 - FinTech Regulatory Sandbox, 50
 - FloBC, incentive registries, 153
 - Flowers, Tommy, 3
 - Flywheel, federated learners, 134
 - FMEA (Failure Mode and Effects Analysis), 82–84
 - Ford Motor Company, FMEA, 83
 - FTA (Fault Tree Analysis), 84–85
 - functional validation, ethical AI, 209–210
 - future of ethical AI, 212
 - future of responsible AI
 - education, 242–244
 - public awareness, 246
 - regulation, 241–242
 - skills, 242–244
 - speed of development, 239–241
 - stakeholders, 242–243
 - standards, 244–245
 - STEM education, 242–243
 - tools, 245
 - upskilling, 242–244
- G**
-
- GDPR (General Data Protection Regulation), 71, 241, 242
 - NAI suite, 107
 - RAI governance of API, 106–107
 - gender debiasing, Reejig ethical AI, 202–203
 - general (strong) AI, 6–7
 - global explainers, 180–182
 - global-view auditors, 156–157, 164
 - FG-AI4H audit platform, 157
 - NVIDIA, 157
 - Seclea, 157
 - Google
 - 2022 Diversity Annual Report, 77

- BARD conversational AI interface, 239–240
- continuous documentation with templates, 80
- dataset requirements, 93
- Differential Privacy Library, 177
- Fairness Gym, 170
- Fairness Indicators, 170
- Fully Homomorphic Encryption, 174
- PAIR Guidebook, 90
- SecAgg, 176
- Vertex Explainable AI, 180
- Vision AI, 108
- governance
 - AI Act, 44
 - Algorithmic Accountability Act of 2022, 44
 - API, 20–21
 - bills of materials, 68–70
 - Blueprint for an AI Bill of Rights, 44
 - building code, 50–51
 - certification, 46–48
 - code of RAI, 60–62
 - continuous documentation with templates, 79–80
 - coupling AI/Non-AI development, 74–75
 - customized agile processes, 72–73
 - data, 94
 - diverse teams, 75–77
 - diversity/inclusion
 - AI definition of, 218–219
 - guidelines, 232–234, 236
 - FMEA, 82–84
 - FTA, 84–85
 - independent oversight, 51–53
 - industry-level governance patterns
 - certification, 46–48
 - independent oversight, 51–53
 - laws/regulations, 42–44
 - maturity models, 44–46
 - regulatory sandboxes, 48–50
 - standards, 55–56
 - trust marks, 53–54
 - Internet Information Service Algorithmic Management Regulations, 44
 - laws/regulations, 42–43
 - leadership commitment, 58–59
 - maturity models, 44–46
 - Microsoft Azure DevOps, 73
 - multi-level governance patterns, 39–42
 - organization-level governance patterns
 - bills of materials, 68–70
 - code of RAI, 60–62
 - leadership commitment, 56–58
 - RAI training, 64–66
 - risk assessments, 62–64
 - risk committees, 58–60
 - role-level accountability contracts, 66–68
 - standardized reporting, 70–71
 - patterns of AI development, 33, 34–35
 - RAI governance
 - of API, 105–107
 - via API, 43
 - RAI training, 64–66
 - regulatory sandboxes, 43, 48–50
 - risk assessments, 62–64
 - risk committees, 60
 - Robbie the Robot thought experiment, 18–19, 20–21
 - role-level accountability contracts, 66–68
 - stakeholder engagement, 77–78
 - stakeholders, 39–42
 - standardized reporting, 70–71
 - standards, 55–56
 - team-level governance patterns, 72
 - continuous documentation with templates, 79–80
 - coupling AI/Non-AI development, 74–75
 - customized agile processes, 72–73
 - diverse teams, 75–77
 - FMEA, 82–84
 - FTA, 84–85

stakeholders, 77–78
 verifiable claims for AI system artifacts, 80–82
 Telestra, 183–184
 awareness, 186, 187
 customer service improvements (example), 191
 definition of AI, 186–187
 development of, 185–186
 dimensions of risk, 188–189
 future of, 195
 identifying/registering use cases, 192–193
 learnings from practice, 192
 levels of risk, 189–190
 lifecycle governance, 193–194
 policies, 186
 risk assessments, 188
 risk council operations, 190–191
 risk exposure matrix, 189–190
 scalability, 194–195
 significant AI-informed decisions, 187
 tool support, 193
 trust marks, 53–54
 verifiable claims for AI system artifacts, 80–82
 GPT-3 language model, 108
 GPT-4 model
 conversational AI interface, 240–241
 incentive registries, 153
 graphs, knowledge, 165
 Guide for Artificial Intelligence Ethical Requirements Elicitation, The, 95
 guidelines, diversity/inclusion, 219
 data, 222–226, 235
 governance, 232–234, 236
 humans, 219–222, 235
 processes, 226–231, 235
 systems, 231–232, 235
 guiding principles of responsible AI, Robbie the Robot thought experiment, 16–17

H

Harvard Institute for Quantitative Social Science, SmartNoise, 177
 homogeneous redundancy, 139–141, 163–164
 homomorphic encryption, 173, 174
 humans
 bias, talent/workforce management, 200–201
 diversity/inclusion
 AI definition of, 217
 guidelines, 219–222, 235

I

I, Robot, 13–14
 IBM
 AI ethics board, 58
 AI Explainability 360, 180
 AI Fairness 360, 172
 AI service factsheets, 80
 Differential Privacy Library, 177
 Federated Learning, 176
 HE layers, 174
 Watson Natural Language Understanding, 139
 ICO (Information Commissioner's Office), Regulatory Sandbox, 50
 identifying use cases, Telestra, 192–193
 IEEE
 2830–2021 standard, 56
 3652.1–2020 standard, 56
 Artificial Intelligence Standards Committee, 56
 Building Code for Medical Device Software Security, 51
 Building Code for Power System Software Security, 51
 Building Code for the Internet of Things, 51
 Cybersecurity Initiative, 51

- P7000 standard, 56, 244–245
 - Symposium on Computational Intelligence, Reejig ethical AI, 205
- IJCAI, AI for Good, 9
- image generation tools
 - DALL-E, 239
 - Midjourney, 239
 - Stable Diffusion, 239
- IMDA, DPTM, 54
- impact of AI, 198
- implementation, trustworthy development
 - processes, 105
 - RAI construction with reuse, 108–110
 - RAI governance
 - of API, 105–107
 - via API, 107–108
- improving customer service (example), 191
- incentive registries, 151–153
 - FloBC, 153
 - GPT-4 model, 153
 - Open Science Rewards and Incentives Registry, The, 153
 - RBRM, 153
 - RLHF, 153
- inclusion/diversity, 234
 - 2022 Diversity Annual Report, 77
 - AI definition of, 216
 - data, 217–218
 - governance, 218–219
 - humans, 217
 - processes, 218
 - systems, 218
 - Blueprint for Equity and Inclusion in Artificial Intelligence, A*, 214
 - CSIRO Data61’s multidisciplinary and diverse team, 213–214
 - Dartmouth Workshop, 214
 - Equity Fluent Leadership Playbook, 234
 - ethical AI, 203, 210
 - guidelines, 219
 - data, 222–226, 235
 - governance, 232–234, 236
 - humans, 219–222, 235
 - processes, 226–232, 235
 - importance of, 215–216
 - Meta, 77
 - Microsoft, 77
 - National Artificial Intelligence Centre’s Think Tank on Diversity and Inclusion in AI, Australia, 214
 - Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*, 214
- independent audits, Reejig ethical AI, 204–205, 211–212
- independent oversight, 51–53
- industry-level governance patterns
 - certification, 46–48
 - independent oversight, 51–53
 - laws/regulations, 42–44
 - maturity models, 44–46
 - regulatory sandboxes, 48–50
 - standards, 55–56
 - trust marks, 53–54
- industry-level stakeholders, 40–41
- Instagram, public awareness, 246
- Internet Information Service Algorithmic Management Regulations, 44
- InterpretML, 180, 182
- IoT, IEEE Building Code for the Internet of Things, 51
- ISO/IEC 23894 on Artificial Intelligence and Risk Management, 64
- ISO/IEC 23894 standard, 117, 245
- ISO/IEC 42001 Information Technology - Artificial Intelligence - Management System, 56
- ISO/IEC JTC 1/SC 42 committee, 64, 117
- ISO/IEC JTC 1/SC42 standard, 56, 244–245

K

- Keras, pytorch2keras model migration tool, 110
- kill switches, Robbie the Robot thought experiment, 21–22

knowledge bases, 146–148
 Awesome AI Guidelines, 148
 Responsible AI Community Portal,
 The, 148
 Responsible AI Knowledge-base, 148
 knowledge graphs, 165

L

laws/regulations, 42–43
 leadership
 commitment, governance, 56–58
 Equity Fluent Leadership Playbook, 234
 ethical AI, 211
 learnings from practice, Telestra, 192
 legislation
 ethical AI, 201
 AI Act, 202
 AI Bill of Rights, 202
 EU legislation, 202
 NYC Local Law 144, 201–202
 U.S. legislation, 201–202
 future of responsible AI, 241–242
 levels of operationalization, 27–28
 Libal, NAI suite, 107
 lifecycle-driven data requirements, 92–94
 Robbie the Robot thought
 experiment, 20
 lifecycle governance, Telestra,
 193–194
 LiFT (LinkedIn Fairness Toolkit), 172
 LLM (Large Language Models)
 BARD, 239–240
 ChatGPT, 239–240
 GPT-4, 240–241
 Musk, Elon, 240–241
 Russell, Stuart, 240–241
 Tay, 239–240
 local explainers, 178–180
 loops, fallback loops, 164
 Luxton, XAI interface, 105

M

Machine Learning, Azure, 116
 Malta AI-ITA, 129
 Maner, Walter, 3
 Material Registry, Bills of, 21–22
 maturity models, 44–46
 medical devices, IEEE Building Code for
 Medical Device Software Security, 51
 Meta, diversity/inclusion, 77
 metadata tracking, Amazon, 131
 Microsoft
 AirSim, 151
 Azure Active Directory Verifiable
 Credentials, 129
 Azure DevOps, customized agile
 processes, 73
 Azure Machine Learning, 116
 Azure Pipelines, 75
 continuous documentation with
 templates, 80
 diversity/inclusion, 77
 FairLearn, 172
 InterpretML, 180, 182
 SEAL, 174
 SmartNoise, 177
 Tay conversational AI interface, 239–240
 Team Data Science Process, 75
 Midjourney text-to-image generator, 239
 MIL-STD-1629A, US Armed Forces Military
 Procedure 83
 minimum complexity, AI system design,
 160–161
*Minimum Elements for a Software Bill of
 Materials, The*, 69
 MIT, Ethics of AI: Safeguarding Humanity, 66
 MLflow Model Registry, 119, 131
 MLOps, 132
 mode switchers, 21–22, 134–137, 164
 model training, 119
 model-based design
 centralized learners, 162–163

federated learners, 162–163
 Robbie the Robot thought experiment, 20
 monitoring, 164
 Muccini, multi-level co-architecting, 98
 multi-level co-architecting, 96–98
 multi-level co-versioning, 118–120
 multi-level governance patterns, 39–42
 multi-model decision-makers, 137–139
 Musk, Elon 240–241

N

NAI suite, 107
 NAIAC (National Artificial Intelligence
 Advisory Committee), 53
 narrow (weak) AI, 6–7
 National AI Centre, 6–7, 245
 National Artificial Intelligence Centre's Think
 Tank on Diversity and Inclusion in AI,
 Australia, 214
 National Data Commissioner, Australia, 68
 National Institute of Standards and
 Technology, *Towards a Standard for
 Identifying and Managing Bias in Artificial
 Intelligence*, 214
 new features (Robbie the Robot thought
 experiment), quarantining, 22
 New Zealand, Privacy Trust Marks, 54
 NIST, AI Risk Management Framework, 64,
 117, 245
 nonsymbolic AI, 7–8
 normative responsibilities, 25, 26
 Norwegian Data Protection Agency,
 sandboxes, 146
 NSW AI Assurance Framework, 64, 117–118
 NTIA, *The Minimum Elements for a Software Bill
 of Materials*, 68–70
 NVIDIA DRIVE Sim
 digital twins, 151
 system-level simulation, 103
 NVIDIA, global-view auditors, 157
 NYC Local Law 144, 201–202

O

Open Science Rewards and Incentives
 Registry, The, 153
 OpenAI
 ChatGPT
 continuous validators, 143–144
 conversational AI interface, 239–240
 DALL-E text-to-image generator, 239
 GPT-3 language model, 108
 GPT-4 model
 conversational AI interface, 240–241
 incentive registries, 153
 risk assessments, 118
 OpenBOM, RAI BOM registry, 126
 OpenMined, PyGrid, 176
 operation infrastructure layer, reference
 architectures, 164–165
 operation infrastructure patterns, 141
 black boxes, 153–155
 continuous validators, 141–144
 digital twins, 148–151
 global-view auditors, 156–157
 incentive registries, 151–153
 knowledge bases, 146–148
 sandboxes, 144–146
 operationalization, 26–27, 28
 checklists, 28
 concrete practices, 28
 levels of, 27–28
 templates, 28
 operations, 114
 multi-level co-versioning, 118–120
 RAI continuous deployments, 114–116
 risk assessments, 116–118
 organization-level governance patterns
 bills of materials, 68–70
 code of RAI, 60–62
 leadership commitment, 56–58
 RAI training, 64–66
 risk assessments, 62–64
 risk committees, 58–60

- role-level accountability contracts, 66–68
 - standardized reporting, 70–71
- organization-level stakeholders, 41
- OWASP, verifiable claims for AI system artifacts, 82

P

- Pachyderm data lineage, 131
- PAIR Guidebook, suitability assessments, 90
- pattern-oriented reference architectures, 161
- patterns of AI development, 31–32
 - governance patterns, 33, 34–35
 - overview, 9–10, 33–34
 - process patterns, 33, 36
 - product patterns, 33, 37
- PDPA (Personal Data Protection Act), 54
- PDPC (Personal Data Protection Commission), 54
- pipelines
 - Amazon SageMaker Pipelines, 75
 - Azure Pipelines, 75
- policies, Telestra, 186
- posessive responsibilities, 25–26
- Powell, Stuart. *See* Telestra
- power systems, IEEE Building Code for Power System Software Security, 51
- practice (Telestra), learning from, 192
- practices, responsible AI, 31–32
- principles of responsible AI, 30, 167
 - explainability, 178–182
 - DARPA, 178
 - EthicalML-XAI, 182
 - global explainers, 180–182
 - Google Vertex Explainable AI, 180
 - IBM AI Explainability 360, 180
 - local explainers, 178–180
 - Microsoft InterpretML, 180, 182
 - RISE, 178
 - tf-explain, 182
 - fairness, 167–168
 - assessing, 168–170
 - bias, 167–168
 - discrimination mitigators, 170–172
 - Facebook Fairness Flow, 170
 - Google Fairness Gym, 170
 - Google Fairness Indicators, 170
 - IBM AI Fairness 360, 172
 - LiFT, 172
 - Microsoft FairLearn, 172
- privacy, 172–173
 - encrypted data-based trainers, 173–174
 - Google Differential Privacy Library, 177
 - Google SecAgg, 176
 - IBM Differential Privacy Library, 177
 - IBM Federated Learning, 176
 - IBM HE layers, 174
 - Microsoft SEAL, 174
 - OpenMined PyGrid, 176
 - random noise data generators, 176–177
 - secure aggregators, 174–176
 - SmartNoise, 177
- Robbie the Robot thought experiment, 16–17
- privacy, 172–173
 - CPRA, 242
 - encrypted data-based trainers, 173–174
 - ethical AI, 208, 211
 - Google Differential Privacy Library, 177
 - Google SecAgg, 176
 - IBM Differential Privacy Library, 177
 - IBM Federated Learning, 176
 - IBM HE layers, 174
 - Microsoft SEAL, 174
 - OpenMined PyGrid, 176
 - random noise data generators, 176–177
 - secure aggregators, 174–176
 - SmartNoise, 177
- Privacy Trust Marks, 54
- processes
 - diversity/inclusion
 - AI definition of, 218
 - guidelines, 226–232, 235

patterns of AI development, 33, 36
 Robbie the Robot thought experiment, 19–21
 trustworthy development processes, 87–88
 design, 96–105
 implementation, 105–110
 operations, 114–120
 requirements, 88–95
 testing, 110–114
 product considerations
 patterns of AI development, 33, 37
 Robbie the Robot thought experiment, 21–22
 product patterns, 121
 operation infrastructure patterns, 141, 151–153
 black boxes, 153–155
 continuous validators, 141–144
 digital twins, 148–151
 global-view auditors, 156–157
 knowledge bases, 146–148
 sandboxes, 144–146
 overview, 122–123
 supply chain patterns, 123
 co-versioning registries, 129–132
 federated learners, 132–134
 RAI BOM registry, 123–126
 verifiable RAI credentials, 126–129
 system patterns, 134
 homogeneous redundancy, 139–141
 mode switchers, 134–137
 multi-model decision-makers, 137–139
 tradeoffs, 123
 project management, APM and stakeholder engagement, 78
 public awareness
 future of responsible AI, 246
 Instagram, 246
 Volkswagen emissions testing, 246
 PyGrid, 176

Pynguin tool, RAI assessments for test cases, 114
 PyTorch, pytorch2keras model migration tool, 110

Q

Qualdo, continuous validators, 143
 quarantining new features, Robbie the Robot thought experiment, 22

R

RAI (Responsible AI)
 assessments for test cases, 112–114
 construction with reuse, 108–110
 continuous deployments, 114–116
 defined, 4–6
 descriptive responsibilities, 25–26
 future of
 education, 242–244
 public awareness, 246
 regulation, 241–242
 skills, 242–244
 speed of development, 239–241
 stakeholders, 242–243
 standards, 244–245
 STEM education, 242–243
 tools, 245
 upskilling, 242–244
 governance
 of API, 105–107
 via API, 43, 107–108
 normative responsibilities, 25, 26
 operationalization
 checklists, 28
 concrete practices, 28
 levels of, 28
 templates, 28
 patterns of AI development
 governance patterns, 33, 34–35
 overview, 9–10, 31–32, 33–34

- process patterns, 33, 36
 - product patterns, 33, 37
 - possessive responsibilities, 25–26
 - practices, 31–32
 - principles of, 30, 167
 - explainability, 178–182
 - fairness, 167–172
 - privacy, 172–177
 - requirements of, 30–31
 - responsibility for responsible AI, 8–9
 - Robbie the Robot thought experiment,
 - guiding principles of responsible AI, 16–17
 - training, 64–66
 - trust, 28–29
 - trustworthiness, 28–29
 - random noise data generators, 176–177
 - RBRM (Rule-Based Reward Models), 153
 - RCAID, Telestra
 - customer service improvements (example), 191
 - dimensions of risk, 188–189
 - future of, 195
 - identifying/registering use cases, 192–193
 - learnings from practice, 192
 - lifecycle governance, 193–194
 - risk assessments, 188
 - scalability, 194–195
 - tool support, 193
 - reader's guide (contents overview), 10–11
 - redundancy
 - critical systems, Robbie the Robot thought experiment, 22
 - homogeneous redundancy, 139–141, 163–164
 - Reejig ethical AI, 183–184, 197, 202
 - aggregating siloed data across multiple sources, 199
 - Caine, Mark, 205
 - case studies
 - algorithms, 206
 - approaches, 206
 - ethical principles, 207–208
 - independent audits, 204–205
 - objectives, 206
 - overview, 204
 - debiasing strategies, 202–203
 - diversity/inclusion, 203
 - gender debiasing, 202–203
 - IEEE Symposium on Computational Intelligence, 205
 - independent audits, case studies, 211–212
 - validation
 - ethical validation, 208–209
 - functional validation, 209–210
 - World Economic Forum, 205
 - Zero Wasted Potential, 212
- reference architectures, 159
 - operation infrastructure layer, 164–165
 - supply chain layer, 162–163
 - system layer, 163–164
 - registering use cases, Telestra, 192–193
 - regulations, 42–43
 - ethical AI, 201
 - AI Act, 202
 - AI Bill of Rights, 202
 - EU legislation, 202
 - NYC Local Law 144, 201–202
 - U.S. legislation, 201–202
 - future of responsible AI, 241–242
 - regulatory sandboxes, 43, 48–50
 - Rekognition, 108
 - reports, standardized, 70–71
 - requirements of
 - RAI, 20, 30–31
 - trustworthy development processes
 - lifecycle-driven data requirements, 92–94
 - suitability assessments, 89–90
 - user stories, 94–95
 - verifiable RAI requirements, 90–92
 - verifiable RAI, Robbie the Robot thought experiment, 20

Responsible AI Community Portal, The, 148
 Responsible AI Institute, certification, 48
 Responsible AI Knowledge-base, 148
 Responsible AI Question Bank, 39, 64
 reusability, AI system design, 160–161
 reuse, RAI construction, 108–110
 rfPro, system-level simulation, 103
 rfProis, digital twins, 151
 RISE (Randomized Input Sampling for Explanation), 178
 risk

- assessments, 62–64, 116–118
 - Robbie the Robot thought experiment, 18
 - Telestra, 188
- committees, 58–60
 - RCAID, Telestra risk assessments, 188
 - Robbie the Robot thought experiment, 18–19
- councils, Telestra operations, 190–191
- dimensions of, 188–189
- exposure matrix, Telestra, 189–190
- levels of, 189–190

 RLHF (Reinforcement Learning with Human Feedback), 153
 Robbie the Robot thought experiment

- API governance, 20–21
- Australian AI Ethics Principles, 16–17
- Bills of Material Registry, 21–22
- CEOs, 18–19
- critical systems redundancy, 22
- data lifecycles, 20
- governance, 18–19, 20–21
- guiding principles of responsible AI, 16–17
- I, Robot*, 13–14, 16–17
- kill switches, 21–22
- Mode Switchers, 21–22
- processes, 19–21
- product considerations, 21–22
- quarantining new features, 22
- risk assessments, 18

- risk committees, 18–19
- Robbie the Robot thought experiment, 20
- sandboxes, 22
- software engineering process, 20–21
- stakeholders, 14–16
- system-level simulation, 20
- training, 18
- verifiable RAI requirements, 20

 RoBoTIPS, black boxes, 155
 role-level accountability contracts, 66–68
 Russell, Stuart, 240–241

S

SageMaker

- continuous validators, 143
- RAI continuous deployments, 116

 SageMaker Pipelines, 75
 SAIAaaS, 110
 Salesforce, maturity models, 46
 sandboxes, 144–146, 164–165

- AI Sandbox, 146
- Fastcase AI Sandbox, 146
- Norwegian Data Protection Agency, 146
- regulatory sandboxes, 43, 48–50
- Robbie the Robot thought experiment, 22

 SBOM (Software Bill of Materials), 126
 scalability, Telestra, 194–195
 Schneider Electric, leadership commitment, 58
 Scikit-learn, 139
 SEAL, 174
 SecAgg, 176
 Seclea, global-view auditors, 157
 secure aggregators, 174–176
 Securekey, 129
 security

- API, 107
- ethical AI, 208, 211

 Shneiderman's classification, 39
 significant AI-informed decisions, 187
 siloed data, aggregating across multiple sources, 198–199

- simulation, Robbie the Robot thought experiment, 20
 - Singapore
 - DPTM, 54
 - FinTech Regulatory Sandbox, 50
 - PDPC, 54
 - skills, future of responsible AI, 242–244
 - SmartNoise, 177
 - software engineering process, Robbie the Robot thought experiment, 20–21
 - Sony Group AI Ethics Guidelines, 60
 - SPDX (Software Package Data Exchange)
 - bills of materials, 70
 - RAI BOM registry, 126
 - spectrum, AI as a, 7–8
 - speed of AI development, 240–241
 - Stable Diffusion text-to-image generator, 239
 - stakeholders
 - engagement, 77–78
 - future of responsible AI, 242–243
 - governance, 39–42
 - industry-level stakeholders, 40–41
 - organization-level stakeholders, 41
 - Robbie the Robot thought experiment, 14–16
 - team-level stakeholders, 42
 - standardized reporting, 70–71
 - standards
 - future of responsible AI, 244–245
 - industry-level governance patterns, 55–56
 - STEM education, future of responsible AI, 242–243
 - strong (general) AI, 6–7
 - structure of this book (reader's guide), 10–11
 - suitability assessments, 89–90
 - supply chain layer, reference architectures, 162–163
 - supply chain patterns, 123
 - co-versioning registries, 129–132
 - federated learners, 132–134
 - RAI BOM registry, 123–126
 - verifiable RAI credentials, 126–129
 - switching AI modes, 134–137, 164
 - symbolic AI, 7–8
 - system artifacts, verifiable claims for, 80–82
 - system design
 - architectural principles, 160–161
 - co-architecting AI/non-AI components, 160
 - complexity, 160–161
 - reusability, 160–161
 - system layer, reference architectures, 163–164
 - system patterns, 134
 - homogeneous redundancy, 139–141
 - mode switchers, 134–137
 - multi-model decision-makers, 137–139
 - system-level simulation, 101–103
 - systems, diversity/inclusion, 218
-
- ## T

- talent/workforce management
 - aggregating siloed data across multiple sources, 198–199
 - bias
 - data bias, 200
 - human bias, 200–201
 - unconscious bias, 199–200
 - decision-making support, 199
 - ethical AI
 - AI Act, 202
 - AI Bill of Rights, 202
 - benefits of, 210–211
 - debiasing strategies, 210
 - diversity/inclusion, 210
 - EU legislation, 202
 - future of, 212
 - leadership, 211
 - NYC Local Law 144, 201–202
 - privacy, 211
 - regulation, 201
 - security, 211
 - U.S. legislation, 201–202

- Reejig ethical AI, 202
 - algorithms, 206
 - approaches, 206
 - Caine, Mark, 205
 - case studies, 204–210
 - debiasing strategies, 202–203
 - diversity/inclusion, 203
 - ethical principles, 207–208
 - ethical validation, 208–209
 - functional validation, 209–210
 - gender debiasing, 202–203
 - IEEE Symposium on Computational Intelligence, 205
 - independent audits, 204–205, 211–212
 - objectives, 206
 - World Economic Forum, 205
 - Zero Wasted Potential, 212
 - role of AI in, 198
- Tay conversational AI interface, 239–240
- Team Data Science Process, 75
- team-level governance patterns, 72
 - continuous documentation with templates, 79–80
 - coupling AI/Non-AI development, 74–75
 - customized agile processes, 72–73
 - diverse teams, 75–77
 - FMEA, 82–84
 - FTA, 84–85
 - stakeholders, 77–78
 - verifiable claims for AI system artifacts, 80–82
- team-level stakeholders, 42
- Telestra, 183–184
 - awareness, 186, 187
 - customer service improvements (example), 191
 - definition of AI, 186–187
 - development of, 185–186
 - future of, 195
 - identifying/registering use cases, 192–193
 - learnings from practice, 192
 - lifecycle governance, 193–194
 - policies, 186
 - risk
 - assessments, 188
 - dimensions of, 188–189
 - levels of, 189–190
 - risk exposure matrix, 189–190
 - risk council operations, 190–191
 - scalability, 194–195
 - significant AI-informed decisions, 187
 - tool support, 193
- templates
 - continuous documentation with templates, 79–80
 - operationalization, 28
- terms of service of data.ai, 68
- Tesla Autopilot
 - homogeneous redundancy, 141
 - mode switchers, 136
- testing
 - acceptance testing, 110–112
 - RAI assessments for test cases, 112–114
 - trustworthy development processes, 110–112
- text-to-image generation tools
 - DALL-E, 239
 - Midjourney, 239
 - Stable Diffusion, 239
- tf-explain, 182
- TFF (TensorFlow Federated), federated learners, 134
- thought experiment, Robbie the Robot
 - API governance, 20–21
 - Australian AI Ethics Principles, 16–17
 - Bills of Material Registry, 21–22
 - CEOs, 18–19
 - critical systems redundancy, 22
 - data lifecycles, 20
 - governance, 18–19, 20–21
 - guiding principles of responsible AI, 16–17
 - I, Robot*, 13–14, 16–17

- kill switches, 21–22
 - Mode Switchers, 21–22
 - processes, 19–21
 - product considerations, 21–22
 - quarantining new features, 22
 - risk assessments, 18
 - risk committees, 18–19
 - sandboxes, 22
 - software engineering process, 20–21
 - stakeholders, 14–16
 - system-level simulation, 20
 - training, 18
 - verifiable RAI requirements, 20
 - Three Laws of Robotics, 16–17
 - tools
 - future of responsible AI, 245
 - Telestra support, 193
 - Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*, 214
 - Toyota, DRFBM, 83
 - tracking metadata, Amazon, 131
 - training
 - RAI training, 64–66
 - Robbie the Robot thought experiment, 18
 - transparency, ethical AI, 207
 - trust in responsible AI, 28–29
 - trust marks, 53–54
 - trustworthiness in responsible AI, 28–29
 - trustworthy development processes, 87–88
 - design, 96
 - design modeling, 99–101
 - envisioning cards, 98–99
 - system-level simulation, 101–103
 - XAI interface, 103–105
 - implementation, 105
 - RAI construction with reuse, 108–110
 - RAI governance of API, 105–107
 - RAI governance via API, 107–108
 - operations, 114
 - multi-level co-versioning, 118–120
 - RAI continuous deployments, 114–116
 - risk assessments, 116–118
 - requirements
 - lifecycle-driven data requirements, 92–94
 - suitability assessments, 89–90
 - user stories, 94–95
 - verifiable RAI requirements, 90–92
 - testing, 110–112
- ## U
- UC Berkeley, Equity Fluent Leadership Playbook, 236
 - UK, ICO Regulatory Sandbox, 50
 - unconscious bias, talent/workforce management, 199
 - University of Helsinki, ethical AI, 66
 - upskilling, future of responsible AI, 242–244
 - US Armed Forces Military Procedures, MIL-STD-1629A, 83
 - US Army Materiel Command's Engineering Design Handbook on Design for Reliability, FTA, 85
 - US Department of Commerce, NAIAC, 53
 - U.S. legislation
 - Algorithmic Accountability Act of 2022, 44
 - Blueprint for an AI Bill of Rights, 44
 - ethical AI, 201–202
 - US National Transportation Safety Board, 53
 - US Robots. *See* Robbie the Robot thought experiment
 - US White House Office of Science and Technology Policy, Blueprint for an AI Bill of Rights, 44
 - use cases, identifying/registering, 192–193
 - user stories, 94–95
 - UTS (University of Technology Sydney)
 - Ethical AI: from Principles to Practice, 66
 - Reejig case study, 197
 - algorithms, 206
 - approaches, 206

Caine, Mark, 205

- ethical principles, 207–208
- ethical validation, 208–209
- functional validation, 209–210
- IEEE Symposium on Computational Intelligence, 205
- independent audits, 204–205, 211–212
- objectives, 206
- overview, 204
- World Economic Forum, 205

V

validation

- continuous validators, 141–144, 164–165
- ethical AI
 - ethical validation, 208–209
 - functional validation, 209–210

values-driven AI. *See* RAI

verifiable claims for AI system artifacts, 80–82

Verifiable Credentials, 82

verifiable RAI

- credentials, 126–129
- requirements
 - Robbie the Robot thought experiment, 20
 - trustworthy development processes, 90–92
- versioning
 - co-versioning, 162–163
 - DVC, 119–120, 131
 - multi-level co-versioning, 118–120

Verta MLOps, 132

Vertex Explainable AI, 180

Vision AI, 108

Viz.ai, EULA, 68

Volkswagen, emissions testing and public awareness, 246

VSD Lab, envisioning cards, 99

W

W3C, Verifiable Credentials, 82

Watson Natural Language Understanding, 139

Waymo

- homogeneous redundancy, 141
- mode switchers, 136

weak (narrow) AI, 6–7

Wiener, Norbert, 3

Wizard of Oz study, XAI interface, 105

workforce/talent management

- aggregating siloed data across multiple sources, 198–199
- bias
 - data bias, 200
 - human bias, 200–201
 - unconscious bias, 199–200
- decision-making support, 199
- ethical AI
 - AI Act, 202
 - AI Bill of Rights, 202
 - benefits of, 210–211
 - debiasing strategies, 210
 - diversity/inclusion, 210
 - EU legislation, 202
 - future of, 212
 - leadership, 211
 - NYC Local Law 144, 201–202
 - privacy, 211
 - regulation, 201
 - security, 211
 - U.S. legislation, 201–202

Reejig ethical AI, 202

- algorithms, 206
- approaches, 206
- Caine, Mark, 205
- case studies, 204–210
- debiasing strategies, 202–203
- diversity/inclusion, 203
- ethical principles, 207–208
- ethical validation, 208–209

functional validation, 209–210
gender debiasing, 202–203
IEEE Symposium on Computational
Intelligence, 205
independent audits, 204–205,
211–212
objectives, 206
World Economic Forum, 205
Zero Wasted Potential, 212
role of AI in, 198
World Economic Forum

*Blueprint for Equity and Inclusion in
Artificial Intelligence, A*, 214
Reejig ethical AI, 205

X - Y - Z

XAI (Explainable AI), 103–105, 178
Xie, acceptance testing, 112

Zero Wasted Potential, Reejig ethical AI, 212
Zowghi, Didar. *See* diversity/inclusion