# RESPONSIBLE AI

## BEST PRACTICES for Creating Trustworthy AI Systems

QINGHUA LU
LIMING ZHU
JON WHITTLE
XIWEI XU

# RESPONSIBLE AI

# Responsible AI: Best Practices for Creating Trustworthy AI Systems

## Table of Contents

Cover

Half Title

Title Page

Copyright Page

Contents

Preface

About the Author

Part I: Background and Introduction

- 1 Introduction to Responsible AI
  - What Is Responsible AI?
  - What Is AI?
  - Developing AI Responsibly: Who Is Responsible for Putting the Responsible into AI?
  - About This Book
  - How to Read This Book
- 2 Operationalizing Responsible AI: A Thought ExperimentRobbie the Robot
  - A Thought ExperimentRobbie the Robot
    - Who Should Be Involved in Building Robbie?
    - What Are the Responsible AI Principles for Robbie?
    - Robbie and Governance Considerations
    - Robbie and Process Considerations
    - Robbie and Product Considerations
  - Summary

# Table of Contents

Part II: Responsible AI Pattern Catalogue

3 Overview of the Responsible AI Pattern Catalogue

The Key Concepts

The Multifaceted Meanings of Responsible

Varied Understandings of Operationalization

The Duality of Trust and Trustworthiness

Why Is Responsible AI Different?

A Pattern-Oriented Approach for Responsible AI

4 Multi-Level Governance Patterns for Responsible AI

Industry-Level Governance Patterns

G.1. RAI Law and Regulation

G.2. RAI Maturity Model

G.3. RAI Certification

G.4. Regulatory Sandbox

G.5. Building Code

G.6. Independent Oversight

G.7. Trust Mark

G.8. RAI Standards

Organization-Level Governance Patterns

G.9. Leadership Commitment for RAI

G.10. RAI Risk Committee

G.11. Code of RAI

G.12. RAI Risk Assessment

G.13. RAI Training

G.14. Role-Level Accountability Contract

G.15. RAI Bill of Materials

G.16. Standardized Reporting

Team-Level Governance Patterns

G.17. Customized Agile Process

G.18. Tight Coupling of AI and Non-AI Development

G.19. Diverse Team

G.20. Stakeholder Engagement

# Table of Contents

G.21. Continuous Documentation Using Templates

G.22. Verifiable Claim for AI System Artifacts

G.23. Failure Mode and Effects Analysis (FMEA)

G.24. Fault Tree Analysis (FTA)

Summary

5 Process Patterns for Trustworthy Development Processes

Requirements

P.1. AI Suitability Assessment

P.2. Verifiable RAI Requirement

P.3. Lifecycle-Driven Data Requirement

P.4. RAI User Story

Design

P.5. Multi-Level Co-Architecting

P.6. Envisioning Card

P.7. RAI Design Modeling

P.8. System-Level RAI Simulation

P.9. XAI Interface

Implementation

P.10. RAI Governance of APIs

P.11. RAI Governance via APIs

P.12. RAI Construction with Reuse

Testing

P.13. RAI Acceptance Testing

P.14. RAI Assessment for Test Cases

Operations

P.15. Continuous Deployment for RAI

P.16. Extensible, Adaptive, and Dynamic RAI Risk Assessment

P.17. Multi-Level Co-Versioning

Summary

6 Product Patterns for Responsible-AI-by-Design

Product Pattern Collection Overview

Supply Chain Patterns

# Table of Contents

D.1. RAI Bill of Materials Registry

D.2. Verifiable RAI Credential

D.3. Co-Versioning Registry

D.4. Federated Learner

System Patterns

D.5. AI Mode Switcher

D.6. Multi-Model Decision-Maker

D.7. Homogeneous Redundancy

Operation Infrastructure Patterns

D.8. Continuous RAI Validator

D.9. RAI Sandbox

D.10. RAI Knowledge Base

D.11. RAI Digital Twin

D.12. Incentive Registry

D.13. RAI Black Box

D.14. Global-View Auditor

Summary

7 Pattern-Oriented Reference Architecture for Responsible-AI-by-Design

Architectural Principles for Designing AI Systems

Pattern-Oriented Reference Architecture

Supply Chain Layer

System Layer

Operation Infrastructure Layer

Summary

8 Principle-Specific Techniques for Responsible AI

Fairness

T.1. Fairness Assessor

T.2. Discrimination Mitigator

Privacy

T.3. Encrypted-Data-Based Trainer

T.4. Secure Aggregator

T.5. Random Noise Data Generator

# Table of Contents

Explainability

T.6. Local Explainer

T.7. Global Explainer

Summary

Part III: Case Studies

9 Risk-Based AI Governance in Telstra

Policy and Awareness

Telstras Definition of AI

Awareness

Assessing Risk

Dimensions of Risk

Levels of Risk

Operation of the Risk Council

Learnings from Practice

Identifying and Registering Use Cases

Support from Technology Tools

Governance over the Whole Lifecycle

Scaling Up

Future Work

10 Reejig: The Worlds First Independently Audited Ethical Talent AI

How Is AI Being Used in Talent?

Aggregating Siloed Data Across Multiple Sources

Providing Decision-Making Support at Scale

Reducing Unconscious Bias

What Does Bias in Talent AI Look Like?

Data Bias

Human Bias

Regulating Talent AI Is a Global Issue

US Legislation Being Introduced

European Legislation Being Introduced

Reejigs Approach to Ethical Talent AI

# Table of Contents

Debiasing Strategies

How Ethical AI Evaluation Is Done: A Case Study in Reejigs World-First Independently Audited Ethical Talent AI

Overview

The Independent Audit Approach

A Summary of the Results

Recognition and Impact

Project Overview

About the Reejig Algorithm

The Objectives of the Project

The Approach

The Ethical AI Framework Used for the Audit

Ethical Principles

Ethical Validation

Functional Validation

The Benefits of Ethical Talent AI

Building Stronger and More Diverse Teams by Removing Bias

Maintaining Privacy and Security

Demonstrating Leadership Against Competitors

Reejigs Outlook on the Future of Ethical Talent AI

Reejig Has Led the Way in AI Ethics from Day One

A New Independent Audit of Reejig Is Already Underway

The Future of Workforce AI Will Unlock Zero Wasted Potential

11 Diversity and Inclusion in Artificial Intelligence

Importance of Diversity and Inclusion in AI

Definition of Diversity and Inclusion in Artificial Intelligence

Guidelines for Diversity and Inclusion in Artificial Intelligence

Humans

Data

Process

System

Governance

# Table of Contents

Conclusion

Human

Data

Process

System

Governance

Part IV: Looking to the Future

12 The Future of Responsible AI

Regulation

Education

Standards

Tools

Public Awareness

Final Remarks

Part V: Appendix

Index

Pearson