IAPP®

AIGP

Artificial Intelligence Governance Professional

STUDY GUIDE

2025 EXAM OBJECTIVES

Includes interactive online learning environment and study tools:

2 custom practice exams

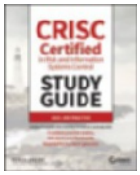
100 electronic flashcards

Searchable key term glossary

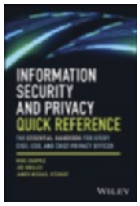
PETER H. GREGORY, CRISC, CISM, CISA, CISSP, CDPSE, CIPM, CCSK

SYBEX®
A Wiley Brand

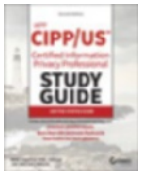
Other Information Security and Governance Study Guides and References from Sybex and Wiley



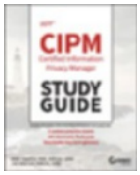
CRISC Certified in Risk and Information Systems Control Study Guide — ISBN 978-1-394-37366-6 (coming May 2026)



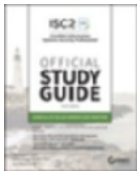
Information Security and Privacy Quick Reference: The Essential Handbook for Every CISO, CSO, and Chief Privacy Officer — ISBN 978-1-394-35331-6, May 2025



IAPP CIPP / US Certified Information Privacy Professional Study Guide, 2nd Edition — ISBN 978-1-394-28490-0, January 2025



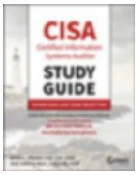
IAPP CIPM Certified Information Privacy Manager Study Guide — ISBN 978-1-394-15380-0, January 2023



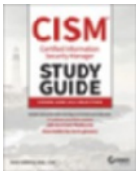
ISC2 CISSP Certified Information Systems Security Professional Official Study Guide, 10th Edition — ISBN 978-1-394-25469-9, June 2024



ISC2 CCSP Certified Cloud Security Professional Official Study Guide, 3rd Edition — ISBN 978-1-119-90937-8, October 2022



CISA Certified Information Systems Auditor Study Guide Covering 2024–2029 Exam Objectives — ISBN 978-1-394-28838-0, December 2024



CISM Certified Information Security Manager Study Guide — ISBN 978-1-119-80193-1, May 2022

IAPP® AIGP®

Artificial Intelligence Governance Professional Study Guide

2025 Exam Objectives



IAPP® AIGP®

Artificial Intelligence Governance Professional Study Guide

2025 Exam Objectives



Peter H. Gregory, CRISC, CISM, CISA,
CISSP, CDPSE, CIPM, CCSK

SYBEX®
A Wiley Brand

Copyright © 2026 by John Wiley & Sons, Inc. All rights reserved, including rights for text and data mining and training of artificial intelligence technologies or similar technologies.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at http://www.wiley.com/go/permission.

The manufacturer's authorized representative according to the EU General Product Safety Regulation is Wiley-VCH GmbH, Boschstr. 12, 69469 Weinheim, Germany, e-mail: Product_Safety@wiley.com.

Trademarks: Wiley and the Wiley logo, and the Sybex logo, are trademarks or registered trademarks of John Wiley & Sons, Inc. and/or its affiliates in the United States and other countries and may not be used without written permission. All other trademarks are the property of their respective owners. John Wiley & Sons, Inc. is not associated with any product or vendor mentioned in this book.

Limit of Liability/Disclaimer of Warranty: While the publisher and the authors have used their best efforts in preparing this work, including a review of the content of the work, neither the publisher nor the authors make any representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials or promotional statements for this work. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for your situation. You should consult with a specialist where appropriate. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002. For product technical support, you can find answers to frequently asked questions or reach us via live chat at https://sybexsupport.wiley.com.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our website at www.wiley.com.

Library of Congress Cataloging-in-Publication Data has been applied for:

Paperback ISBN: 9781394363940
ePDF ISBN: 9781394363964
ePub ISBN: 9781394363957
Obook ISBN: 9781394363971

Cover Design: Wiley
Cover Image: © Jeremy Woodhouse/Getty Images

To all business professionals who aspire to improve their organizations' efficiency, integrity, and safety through the use of AI.

Acknowledgments

Books like this involve the work of many people, and as an author, I truly appreciate the hard work and dedication that the team at Wiley demonstrates. I want to extend special thanks to my acquisitions editor, Jim Minatel, who worked diligently to make this project possible.

I also greatly appreciate the editing and production team for the book, including Patrick Walsh, the project manager who kept the train on the tracks; Kezia Endsley, the copy editor who made sure all of the language and formatting were just right; Navin Vijayakumar, the managing editor, who managed the various phases of editing, and Maduramuthu Krishnaraj, the Content Refinement Specialist, who guided us through layouts, formatting, and final cleanup to produce a great book. Finally, special thanks to Ejona Preçi, the technical editor, who provided insightful advice and gave excellent improvement suggestions and feedback throughout the book. I would also like to thank the many behind-the-scenes contributors, including the graphic, production, and technical teams, who helped bring the book and companion materials to a finished product.

My long-time literary agent, Carole Jelen of Waterside Productions, continues to provide me with fantastic opportunities, advice, and assistance throughout my writing career.

I have difficulty expressing my gratitude to my wife and love of my life, Rebekah, for tolerating my frequent absences (in the home office) while I developed this manuscript. This project could not have been completed without her loyal and unfailing support and encouragement.

About the Author

Peter H. Gregory (CISSP, CISM, CISA, CRISC, CIPM, CDPSE, CCSK, A/CCRF, A/CCRP, A/CRMP) is the author of more than 60 books on security, privacy, and technology, including *Solaris Security* (Prentice Hall, 2000), *The Art of Writing Technical Books* (Waterside, 2022), *CISA Certified Information Systems Auditor Study Guide* (Wiley, 2025), *Chromebook For Dummies* (Wiley, 2023), and *Elementary Information Security* (Jones & Bartlett Learning, 2024).

Peter is a recently retired career technologist and cybersecurity executive. He has held security leadership positions at GCI Communications, Optiv Security, and Concur Technologies since the early 2000s, after a celebrated career in software, systems, and security engineering at Bally Gaming, World Vision, and AT&T Wireless. Peter serves as an advisory board member for the University of Washington and Seattle University, with a focus on cybersecurity education programs. He is a 2008 graduate of the FBI Citizens' Academy.

Peter resides in Central Washington State and can be found at www.peterhgregory.com.

About the Technical Editor

Ejona Preçi (CISSP, CISM, CRISC, ITIL, ISO/IEC 42001 LI) is a renowned global CISO and a leading voice in cybersecurity and AI governance. She works at the forefront of advancing responsible technology, developing frameworks that guide the safe, ethical, and trustworthy use of AI systems. A multiple award winner, she has been honored as Cybersecurity Woman of the Year 2024 (Barrier Breaker), named among the Global 40 Under 40 in Cybersecurity, and recognized as one of the Top 20 Women in Cybersecurity 2025. As a thought leader, writer, and keynote speaker, she is dedicated to shaping global policy discourse and promoting trust, resilience, and accountability in the era of AI.

Contents at a Glance

Acknowledgments *viii*

About the Author *ix*

About the Technical Editor *x*

Introduction *xxxiii*

Assessment Test *xliv*

Answers to Assessment Test *lii*

Part I **Foundations of AI and Governance** **1**

Chapter 1 AI and AI Governance 3

Chapter 2 Organizational Readiness 29

Chapter 3 Updating Policies for AI 71

Part II **Legal and Regulatory Obligations** **99**

Chapter 4 Privacy and Data Protection Law 101

Chapter 5 Sectoral and Civil Laws 147

Chapter 6 The EU AI Act 193

Chapter 7 AI Standards and Frameworks 221

Part III **Governing the AI Lifecycle** **255**

Chapter 8 AI System Design 257

Chapter 9 Data Governance and Model Training 301

Chapter 10 Deployment and Monitoring 331

Chapter 11 AI Risk Management and Assurance 353

Chapter 12 Ongoing AI Operations 379

Appendix

Appendix Answers to the Review Questions 401

Index *411*

Contents

Acknowledgments *viii*

About the Author *ix*

About the Technical Editor *x*

Introduction *xxxiii*

Assessment Test *xliv*

Answers to Assessment Test *lii*

Part I Foundations of AI and Governance 1

Chapter 1 AI and AI Governance 3

The Types of AI 4
- Definition of AI 4
- AI Capabilities 5
- AI Functionalities 5
- AI Techniques 6

Risks and Harms Posed by AI 7
- Risks and Harms to Individuals 8
- Risks and Harms to Groups 9
- Risks and Harms to Organizations 9
- Risks and Harms to Society 11

Characteristics of AI Requiring Governance 12
- Complexity 13
- Opacity 13
- Autonomy 14
- Speed and Scale 14
- Potential for Harm and Misuse 15
- Data Dependency 15
- Privacy 16
- Intellectual Property 16
- Probabilistic vs. Deterministic Outputs 17

Principles of Responsible AI 17
- Ethics 17
  - Challenges with Ethics 18
  - Applications of Ethics 18
- Fairness 18
  - Challenges with Fairness 19
  - Applications of Fairness 19

Safety and Reliability 19
Challenges with Safety 19
Applications of Safety and Reliability 19
Privacy 20
Challenges with Privacy 20
Applications of Privacy 20
Security 20
Challenges with Security 21
Applications of Security 21
Transparency and Explainability 21
Challenges with Transparency and Explainability 21
Applications of Transparency and Explainability 21
Accountability 22
Challenges with Accountability 22
Applications of Accountability 22
Human-centricity 23
Challenges with Human-centricity 23
Applications of Human-centricity 23
Summary 23
Exam Essentials 25
Review Questions 26

Chapter 2 Organizational Readiness 29

Roles and Responsibilities for AI Governance Stakeholders 30
The Strategic Importance of Defined Governance Roles 31
Stakeholder Categories and Their Governance Roles 31
Executive Leadership 31
AI Governance Board or Council 32
Legal, Ethics, and Compliance Teams 32
Technical and Engineering Teams 33
Cybersecurity and Risk Management Teams 34
Product and Business Unit Leaders 34
End Users and Frontline Staff 35
External Parties 35
Assigning Responsibilities Across the AI Lifecycle 35
Implementation Tools and Techniques 36
Cross-functional Collaboration in the AI Governance Program 37
Why Cross-functional Collaboration Is Critical 37
Principles of Effective Cross-functional Governance 38
Building the Collaborative Structure 39
Working Groups and Advisory Boards 39
Establish Cross-functional Use Case Reviews 40
Collaboration Templates and Artifacts 40
Fostering a Collaborative Culture 41

External and Cross-organizational Collaboration 42
Sustaining Collaboration Over Time 42
Training and Awareness Program on AI Terminology, Strategy, and Governance 43
Why Training and Awareness Are Foundational 43
The Scope of the Training Program 44
Curriculum Structure 44
Example Structure and Content Areas 44
Example Module Titles 45
Delivering the Training 45
Blended Learning 46
Mandatory vs. Elective Content 47
Regional and Cultural Considerations 47
Maintaining Training Records 47
Building Awareness Beyond Formal Training 47
Tracking and Measuring Success 48
Common Pitfalls and How to Avoid Them 48
Sustaining the Program Over Time 49
Tailoring AI Governance to Organizational Context 49
Company Size: Scaling Governance to Organizational Footprint 50
Small Companies and Startups (1–200 Employees) 50
Mid-sized Organizations (200–5,000 Employees) 51
Large Enterprises (5,000+ Employees) 51
Organizational Maturity and AI Governance 52
Maturity Models 52
Low Maturity 53
Moderate Maturity 53
High Maturity 54
Industry Sector-specific Governance Imperatives 54
Regulated Industries 54
Business- and Consumer-facing Technology 55
Industrial and Manufacturing 55
Public Sector and Nonprofits 55
Products and Services: Governance Based on Use Case Impact 56
Internal Use AI 56
Decision Support AI 56
Automated Decision-making 56
Real-time/High-stakes AI 57
Risk Tolerance: Calibrating Governance Based on Appetite for Uncertainty 57
Low Risk Tolerance 57

Medium Risk Tolerance 57
High Risk Tolerance 58
Business Alignment 58
Innovation-driven Organizations 58
Reputation-focused Organizations 59
Cost-reduction or Efficiency-oriented Organizations 59
Developers, Deployers, and Users in AI Governance 59
Definitions and Role Boundaries 60
AI Developers 60
AI Deployers 60
AI Users 60
Responsibilities Across the AI Lifecycle 61
Opportunities and Leverage Points 62
Resource Needs and Expectations 63
AI Developer 63
AI Deployer 64
AI User 64
Governance Conflicts and Misalignments 64
Developers vs. Users 65
Developers vs. Deployers 65
Users vs. Deployers 65
Role-based Governance Controls 65
Controls for Developers 65
Controls for Deployers 65
Controls for Users 66
Role Coordination and Integration 66
Summary 66
Exam Essentials 68
Review Questions 69

Chapter 3 Updating Policies for AI 71

Oversight in the Age of Autonomous Decision-making 72
The Lifecycle Policy Model 73
Why Lifecycle Oversight Matters 73
Core Features of a Lifecycle-oriented Policy 73
Use Case Assessment: Aligning Purpose and Risk 74
Should AI Be Used at All? 74
Risk Rating Frameworks 75
Risk Management Tailored for AI 75
Ethics by Design: From Principle to Practice 76
Operationalizing AI Ethics 76
Building Ethics into Technical Design 76

Data Acquisition and Use 77
Data Lineage and Documentation 77
Consent, Privacy, and Synthetic Data 77
Model Development: Guardrails for Creation 77
Enforcing Reproducibility and Accountability 78
Safe and Interpretable Model Architectures 78
Training and Testing: Preparing for the Real World 78
Rigorous Validation Standards 79
Red Teaming and Adversarial Testing 79
Test Results Are a Vital Gate 80
Deployment and Monitoring: Guarding the Gate 80
Pre-deployment Controls 80
Post-deployment Monitoring 80
Documentation and Reporting: Building Institutional Memory 81
Living Documentation 81
Internal and External Reporting 81
Incident Management: Planning for Failure 81
Defining and Escalating Incidents 82
AI Playbooks Are Needed 82
Root Cause and Remediation 82
Evaluate and Update Existing Data Privacy and Security Policies for AI 83
Why AI Disrupts Traditional Data Governance 83
AI Is Data-hungry by Design 83
Inference and Reidentification Risks 84
Privacy Policy Gaps and AI-specific Threats 84
Common Policy Shortfalls 84
AI-specific Privacy Threats 84
Key Policy Areas to Evaluate and Update 85
Purpose Limitation in Dynamic Pipelines 85
Enhanced Anonymization and Pseudonymization Standards 85
Lifecycle-based Data Retention Controls 86
Data Subject Rights in AI Contexts 86
Right to Explanation and Access 86
Right to Be Forgotten (Data Deletion) 86
Security Policy Updates for AI Systems 87
Expanding the Threat Model 87
Securing the AI Supply Chain 87
Practical Steps for Policy Revision 87
Conduct a Gap Assessment 88
Build Cross-functional Review Teams 88

Policy Compliance Artifacts 88
Policies to Manage Third-party Risk 88
Why AI Multiplies Third-party Risk 89
Beyond the Traditional Vendor Model 89
Hidden Dependencies and Sub-tier Risks 89
Core Policy Objectives for Third-party AI Risk Management 90
Procurement Policy Controls for AI-enabled Solutions 90
AI Risk Flagging at Procurement Intake 90
AI-specific Due Diligence 91
Vendor Risk Tiers 91
Contracting and Legal Safeguards 91
AI-specific Contractual Clauses 92
Liability and Incident Handling 92
AI Supply Chain Governance 92
Mapping the AI Supply Chain 93
Open-source AI Risk Policies 93
Other Third-party AI Issues 93
Contracting with Human Annotators and AI Workers 93
Employee Use of Generative AI Tools 94
Monitoring and Reviewing Third-party AI Risk 94
Ongoing Vendor Oversight 94
Triggers for Reassessment 94
Summary 95
Exam Essentials 96
Review Questions 97

Part II Legal and Regulatory Obligations 99

Chapter 4 Privacy and Data Protection Law 101

Notice, Choice, Consent, and Purpose Limitation in AI 102
Notice 102
Unique Challenges in AI 103
Choice and Consent 104
Defining Consent and Its Variants 104
Obstacles to Meaningful Consent in AI 104
Purpose Limitation 105
Purpose Creep and Reuse 106
AI-specific Purpose Challenges 106
Data Minimization and Privacy by Design in AI 107
Data Minimization 107
Data Minimization Across the AI Lifecycle 107
Balancing Utility and Minimization 108

Privacy by Design (PbD) 108
Seven Foundational Principles of Privacy by Design 109
Implementing Privacy by Design in AI Development 109
Aligning Privacy by Design with Organizational Roles 110
Practical Tools and Frameworks 110
Practical Implications and Governance 111
Operationalizing Privacy Principles in AI 111
Governance Touchpoints Across the AI Lifecycle 111
Privacy Roles and Responsibilities 112
Documentation and Auditability 112
Scaling Governance 113
Data Controller Obligations in the AI Context 114
Privacy Impact Assessments and Risk Management 114
The Role of DPIAs in AI Development 114
DPIA Lifecycle in AI Projects 115
Common Pitfalls in AI DPIAs 116
Aligning with Emerging Frameworks 116
Using Third-party Processors in AI Projects 116
Controller vs. Processor in the AI Context 116
Contractual Requirements 117
Processor Drift and Compliance Gaps 118
Cross-border Data Transfers 118
Global Data Flows in AI Systems 118
AI-specific Data Transfer Challenges 119
Transfer Impact Assessments (TIAs) 119
Data Subject Rights and AI 120
AI-specific Implementation Challenges 120
Practical Solutions 121
Incident Management and Breach Notification 122
Reframing Breaches in AI Systems 122
GDPR Breach Notification Requirements 122
AI Incident Detection and Preparedness 123
Recordkeeping and Accountability 123
Controllers Must Demonstrate Compliance 124
What to Document in AI Projects 124
Tools and Frameworks for Governance 125
Understanding the Requirements That Apply to Sensitive or Special Categories of Data 125
Legal and Regulatory Foundations 126
GDPR Article 9 and the Prohibition Baseline 126
Key International Equivalents 126
AI's Unique Risk Profile for Sensitive Data 128

Biometrics and Other High-risk Categories of Sensitive Data 128
Biometrics 128
Healthcare and Genetic Data 129
Religious and Philosophical Beliefs 130
Political Opinions and Affiliations 131
Race, Ethnicity, and Union Membership 132
Legal Basis and Consent for Sensitive Data 132
Explicit Consent: A High Bar 132
When Consent Is Not Enough 133
Technical and Organizational Safeguards 134
Core Technical Measures 134
Organizational Safeguards 135
Designing for Data Minimization 135
Special Considerations in AI Applications 135
Bias and Fairness in Sensitive Attributes 135
Inference of Sensitive Data from Non-sensitive Inputs 136
Purpose Creep and Secondary Use 137
DPIAs and Risk Management 137
What to Include in a DPIA for Sensitive AI 138
The Role of Risk Registers and Review Boards 138
Data Subject Rights and Sensitive Data 139
Enhanced Rights Expectations 139
Sensitive Data Adds Risk to AI Rights Requests 139
Opt-out Rights and Limitations in AI Contexts 140
Rights in High-risk Automated Decision-making 141
Summary 141
Exam Essentials 142
Review Questions 144

Chapter 5 Sectoral and Civil Laws 147

Intellectual Property Laws and AI 148
Foundations of Intellectual Property and AI 148
Copyright Law: The Center of AI Training Controversies 149
Patent Law and AI-generated Inventions 149
Trade Secrets and AI Models 150
Database Rights and Regional Variability 151
Use of Protected Works in AI Training 151
The Training Data Dilemma 151
Licensing Models for Training Data 152
AI Outputs and the Question of Infringement 152
When Is an AI Output Infringing? 152
Output Attribution and Legal Accountability 153

Ownership and Protection of AI-generated Content 153
Copyrighting AI-generated Works 153
Joint Authorship and AI as a Tool 154
Risk Mitigation and Compliance Strategies 154
Establishing an IP-aware AI Governance Framework 154
Embedding Contractual Protections 155
Investing in Explainability and Transparency 155
Cross-border IP Conflicts and Jurisdictional Complexities 155
Varying Legal Standards Across Jurisdictions 156
Jurisdictional Enforcement and Forum Shopping 156
International Treaties on Intellectual Property 157
The Future of AI Regulations 157
Proposed U.S. Legislation 157
European Union Initiatives 158
Industry Standards and Self-regulation 158
IP Law as a Central Governance Challenge 158
Non-discrimination Laws and AI 159
Legal Foundations of Anti-discrimination in Algorithmic Systems 160
Protected Classes and Prohibited Conduct 160
Algorithmic Bias as a Legal Risk 161
Employment: AI in Hiring, Promotion, and Termination 161
Legal Requirements for Hiring Algorithms 161
EEOC Guidance on AI in Employment 161
Local Laws and Fair Hiring Ordinances 162
Credit and Lending: Navigating FCRA and ECOA in the Age of AI 162
Equal Credit Opportunity Act (ECOA) 162
Fair Credit Reporting Act (FCRA) 163
Bias in Credit Scoring Models 164
Housing: AI and the Fair Housing Act 164
Protected Characteristics Under FHA 164
Algorithmic Advertising and Targeting 164
Tenant Screening and Credit-based Bias 165
Insurance: Risk Assessment in a Regulated Landscape 165
Discrimination in Insurance Underwriting and Pricing 165
Proxy Discrimination and Feature Selection 166
State-level Developments in Insurance AI 166
Risk Mitigation and Compliance Strategies 167
Perform Bias Audits and Impact Assessments 167
Ensure Explainability and Transparency 167
Establish Human Oversight and Accountability 168

Include Fairness Clauses in Vendor Contracts 168
Fairness as a Legal Mandate, Not Just an Ethical Goal 168
How Consumer Protection Laws Apply to AI 169
Legal Foundations: Unfair and Deceptive Acts or Practices 169
The FTC Act and Section 5 Authority 170
State-level Consumer Protection Laws 170
Emerging Focus Areas for AI Enforcement 171
FTC Enforcement and Policy in the AI Era 171
FTC Guidance on AI Fairness and Transparency 171
Key FTC Enforcement Actions Involving AI 172
AI-powered Deception and Manipulation 173
Dark Patterns and Manipulative Interfaces 173
Chatbots, Impersonation, and Misleading AI Interactions 174
Fake Reviews, AI-generated Endorsements, and Influencer Fraud 174
Emerging Risks: Deepfakes, Misinformation, and Emotional Manipulation 174
AI Misinformation in Health, Finance, and Safety 175
Emotional Manipulation and Vulnerable Populations 175
Romantic Relationships with AI 175
Risk Mitigation and Consumer Protection Compliance Strategies 176
Design for Transparency and Disclosure 176
Substantiate AI Claims 176
Monitor for Algorithmic Drift and Harm 177
Implement AI Governance Policies 178
Vendor and Third-party Risk Controls 178
A Proactive Approach to AI and Consumer Protection 178
How Product Liability Laws Apply to AI 179
Product Liability and Its Application to AI 180
Theories of Product Liability 180
What Constitutes a "Defect" in AI-driven Products? 181
Expanding the Definition of "Product" to Include Software and AI 181
Design Defects in AI Systems 182
Risk-utility Balancing and the "Reasonable Alternative Design" 182
Manufacturing Defects in AI-integrated Products 182
AI as an Assembly of Interdependent Parts 182
Challenges in Detecting AI Manufacturing Defects 183
Failure to Warn: Disclosing AI Risks 183
Human-machine Interface and User Misunderstanding 183

Updating Warnings Post-deployment 184
Preemption and Regulatory Tension 184
Continuing Legal Uncertainty 184
Multi-party Liability and the AI Supply Chain 185
The European Union's Approach to AI and Product Liability 185
Proactive Risk Mitigation Strategies for AI Product Liability 185
Conduct AI-specific Safety Testing 185
Maintain Traceability Across the AI Lifecycle 186
Build Cross-functional Product Safety Governance 187
Integrate AI Update, Governance, and Field Monitoring 187
Modernizing the Liability Mindset 187
Summary 188
Exam Essentials 189
Review Questions 191

Chapter 6 The EU AI Act 193

Risk Classification 194
Prohibited AI Systems 194
Definition and Legal Basis 195
Examples of Prohibited AI 195
Implications for Developers 195
High-risk AI Systems 195
Definition and Criteria 196
Limited-risk AI Systems 196
Definition 197
Transparency Obligations 197
Minimal-risk AI Systems 197
Definition 197
Encouraged Practices 198
Classification Requirements 199
High-risk AI: Full Compliance Framework 199
Risk Management System 199
Data and Data Governance 200
Technical Documentation 200
Record keeping and Logging 201
Conformity Assessment 201
Human Oversight 201
Transparency and User Instructions 201
Quality Management System 202
Limited-risk AI: Targeted Transparency Obligations 202
Disclosure Requirements 202
Notification and User Autonomy 202

Documentation and Accountability 203
Minimal-risk AI: Voluntary Best Practices 203
Requirements for General-purpose AI 204
Defining General-purpose AI Under the EU AI Act 204
Legal Definition 204
General Obligations for GPAI Providers 205
Technical Documentation and Transparency 205
Model Evaluation and Testing 205
Data Governance 205
Additional Obligations for GPAI with Systemic Risk 206
What Qualifies as Systemic Risk? 206
Enhanced Requirements 206
Governance and Accountability 206
Obligations for Distributors and Deployers of GPAI 207
Encouraged Practices Across All GPAI Models 208
Enforcement and Penalties 208
Oversight and Supervisory Authorities 208
Inspection Powers and Remedies 209
Penalties for Non-compliance 209
Escalation and Recurrence 210
Cross-border Enforcement and Cooperation 210
Mutual Assistance Framework 210
Jurisdictional Clarity for GPAI Providers 211
Extraterritoriality 211
Organizational Context 211
AI Providers: Core Compliance Burdens 211
Deployers: Operational Accountability in Practice 212
Importers: Bridging the Legal Jurisdictions 213
Distributors: Passive but Not Exempt 213
Shared and Overlapping Responsibilities 214
Horizontal Duties 214
AI-as-a-Service and Role Ambiguity 214
EU AI Act Implementation 214
Summary 216
Exam Essentials 217
Review Questions 218

Chapter 7 AI Standards and Frameworks 221

OECD Principles 222
Key OECD Principles for Responsible Stewardship of Trustworthy AI 222
Implementation Guidance for Government Policymakers 224
Relevance to AI Governance Professionals 224

NIST AI Risk Management Framework (AI RMF) 224
Overview of the NIST AI RMF 225
The Govern Function: Building the Foundation for AI Risk Management 226
Category 1: Policies, Processes, and Procedures 226
Category 2: Roles and Responsibilities 227
Category 3: Risk Tolerance and Risk Appetite 227
Category 4: Organizational Culture and Internal Communication 228
The Map Function: Understanding the Context of an AI System 228
Category 1: Context and Intended Purpose 228
Category 2: System Design and Functionality 229
Category 3: Stakeholders and Affected Populations 229
Category 4: Risk Identification and Prioritization 229
The Measure Function: Assessing AI Risk and Trustworthiness 230
Category 1: Risk Metrics and Indicators 230
Category 2: Evaluation Methods 230
Category 3: System Performance Monitoring 231
Category 4: Risk Documentation and Reporting 231
The Manage Function: Taking Action to Address AI Risk 231
Category 1: Risk Treatment and Mitigation 232
Category 2: Incident Response and Recovery 232
Category 3: Risk Management Integration 233
Category 4: Risk Evolution and System Retirement 233
Putting It All Together: Integrating the AI RMF into Governance Practice 234
AI RMF vs. NIST Cybersecurity Framework (CSF) 234
NIST ARIA: Layered AI Model Evaluation 236
Purpose and Positioning of ARIA 236
The Three Levels of Testing 236
Model Testing 236
Red Teaming 236
Field Testing 237
Practical Application of ARIA 237
ARIA and Executive Branch Compliance 237
ARIA Beyond the Federal Government 237
ISO Standards 238
ISO/IEC 22989: Concepts and Terminology for AI 238
Why Terminology Matters in Governance 238
Key Definitions and Concepts in ISO 22989 239
AI Capabilities and System Behaviors 240

AI System Lifecycle Stages 240
ISO/IEC 42001: AI Management Systems 240
The Purpose and Scope of ISO/IEC 42001 241
Core Requirements of ISO/IEC 42001 241
Risk Management in ISO/IEC 42001 241
Trustworthiness Dimensions in 42001 242
Implementing ISO/IEC 42001 in Practice 242
Conducting a Readiness Assessment 242
Integrating with Existing Management Systems 243
Developing AI-specific Policies and Procedures 243
Training, Awareness, and Organizational Culture 244
Other Relevant ISO Standards for AI Governance 244
ISO/IEC 23894: AI Risk Management 244
ISO/IEC 38507: Governance of IT (the Governance Implications of AI) 244
ISO/IEC TR 24028: Trustworthiness in AI 245
ISO/IEC 42005: Artificial Intelligence (AI) (the AI System Impact Assessment) 245
ISO AI Standards in the Global Governance Landscape 245
Other AI-related Standards 246
The EU High-level Expert Group (HLEG) Ethics Guidelines 246
The G7 Hiroshima Process and Code of Conduct 247
The IEEE Ethically Aligned Design Framework 247
World Economic Forum (WEF) AI Governance Toolkits 248
OECD Framework for Classification of AI Systems 248
United Nations UNESCO Recommendation on the Ethics of AI 248
Singapore Model AI Governance Framework 249
ISO/IEC JTC 1/SC 42 Roadmap and Future Standards 249
Summary 250
Exam Essentials 251
Review Questions 253

Part III Governing the AI Lifecycle 255

Chapter 8 AI System Design 257

Documenting the Lifecycle 259
Documenting Requirements 260
Documenting the Designing and Building Process 261
Documentation in the Design Phase 261
Documentation in the Development Phase 262
Documenting the Training and Testing Process 263
Training Documentation 263
Testing and Validation Documentation 264

Maintaining Traceability and Audit Readiness 264
Version Control and Change Management 264
Explainability and Documentation Artifacts 265
Integration with Broader Governance Frameworks 265
Business Context and Use Cases 265
Defining the Business Context 266
Understanding Organizational Objectives 266
Organizational Readiness and Governance 266
Stakeholder Identification and Engagement 267
Articulating the AI Use Case 267
Defining the Scope and Boundaries 267
Defining Success Metrics 268
Assessing Data Availability and Suitability 268
Inventorying Available Data 268
Legal and Ethical Constraints on Data Use 269
Data Readiness for Deployment 269
Ethical and Operational Considerations 269
Risk of Harm and Ethical Tradeoffs 269
Human Oversight and Decision Rights 270
Workforce Readiness and Culture 270
Alignment with Organizational Values and Strategy 270
Impact Assessments 271
Understanding AI Impact Assessments 271
Legal and Regulatory Drivers 271
When and How Often to Conduct Assessments 272
Preparing for an Impact Assessment 272
Define the Assessment Scope 272
Identify Stakeholders 273
Inventory Required Information 273
Conducting an Impact Assessment 273
Identify Potential Impacts 273
Evaluate Risk Likelihood and Severity 274
Analyze Discrimination and Fairness Risks 274
Assess Explainability and Transparency 274
Review Human Oversight Mechanisms 275
Document Mitigation Strategies 275
Reviewing and Operationalizing the Assessment 275
Internal Review and Signoff 276
Integration with Development and Deployment 276
External Transparency and Accountability 276
Continuous Reassessment 277
AI System Design Considerations 277
Defining the Purpose and Framing the Problem 277
Architecture and Model Selection 278

Designing for Modularity and Traceability 278
Choosing the Right Model Type 278
Human Oversight and Feedback Loops 279
Designing Human-in-the-loop Processes 279
Designing for Usability and Understanding 280
Feedback Integration 280
Metrics, Thresholds, and Controls 280
Metric Selection 280
Threshold Setting and Adjustment 281
Operational Controls 281
Identifying and Managing Design Risks 281
Framing AI Design Risk 282
Distinguishing Between Internal and External Risk Sources 282
Systematic Risk Identification Techniques 283
Risk Matrix 283
Risk Mitigation Hierarchy 283
Stakeholder Mapping and Risk Perception 284
Use Case Evaluation and Scenario Planning 284
Use Case Classification and Risk Profiling 284
Scenario Planning and Adverse Events 285
Benchmarking and Comparative Risk Analysis 285
Testing, Piloting, and Validation 285
Unit Testing 285
Pre-deployment Pilots 285
Shadow Mode Testing 286
Stress Testing and Red Teaming 286
Risk Governance Integration 286
Alignment with Enterprise Risk Frameworks 287
Documenting Residual Risk and Justifications 287
Preparing for Post-deployment Risk Monitoring 287
Proprietary Model Considerations 287
Ownership and Accountability 288
Legal Liability and Civil Exposure 288
Regulatory Classification as “Provider” 289
Data Rights, Quality, and Governance 289
Data Sourcing and Legal Entitlement 289
Data Quality and Labeling Integrity 290
Data Governance Integration 290
Model Explainability, Testing, and Monitoring 290
Explainability Obligations 291
Testing and Validation Requirements 291
Monitoring and Change Control 291

Intellectual Property and Competitive Risk 292
IP Ownership and Protectability 292
Reverse Engineering and Model Theft 292
Model Misuse and Access Control 293
Resourcing and Organizational Readiness 293
Technical Infrastructure 294
Talent and Skill Development 294
Governance and Ethical Review Capacity 294
Summary 295
Exam Essentials 296
Review Questions 298

Chapter 9 Data Governance and Model Training 301

Data Governance 302
Defining Data Governance for AI 303
Legal Authority and Lawful Use of Data 303
Confirming Lawful Rights to Use Data 303
Documentation Requirements 304
Secondary Use Risks 304
When Using Third-party AI Systems 304
Assessing Data Quality and Integrity 305
The Dimensions of Data Quality 305
Detecting and Preventing Data Drift 305
Ensuring Sufficient and Representative Data 305
Avoiding Underfitting and Overfitting 305
Representing Real-world Populations 306
Fit-for-purpose Evaluation 306
Data Stewardship and Roles 306
Security, Privacy, and Ethical Use 307
Privacy Risk Assessments 307
Security Controls 307
Ethical Considerations 307
Organizational Maturity and Implementation 308
Common Pitfalls and Failure Modes 308
Benefits of Governance in AI Development 309
Data Lineage and Provenance 309
Defining Data Lineage and Data Provenance 310
The Role of Lineage and Provenance in AI 310
Capturing Data Provenance 311
Tracking Data Lineage 312
Common Failure Modes 312
Building a Provenance and Lineage Framework 313
Use Cases and Industry Examples 314

Benefits of Strong Traceability 314
Traceability Challenges and Tradeoffs 314
AI System Testing 315
AI Testing Is a Discipline 315
Understanding AI Testing Objectives 315
Unit and Component Testing 316
Integration Testing 316
Validation Testing 317
Performance Testing 317
Security Testing 318
Bias and Fairness Testing 319
Interpretability and Explainability Testing 319
Human-in-the-loop and Scenario Testing 320
Documentation and Traceability 320
Organizational Practices for AI Testing 321
Challenges and Tradeoffs 321
Identifying and Managing Testing Risks 321
The Unique Nature of AI Testing 322
Types of Testing Risks in AI 322
Testing Risk Identification Techniques 324
Risk Mitigation Strategies 324
Managing Risk Throughout the Training Lifecycle 325
Testing Tools and Infrastructure 326
Common Testing Pitfalls 326
Case Studies and Examples 326
Hiring Algorithm Bias 326
Autonomous Vehicle Simulation 327
Financial Services Outlier 327
Summary 327
Exam Essentials 328
Review Questions 329

Chapter 10 Deployment and Monitoring 331

Pre-release Readiness 333
Assessing AI System Readiness for Production 334
Operational Readiness and Risk Evaluation 334
AI Model Deployment Considerations 335
AI Model Optimization for Production Fit 336
Human-in-the-loop and Release Protocols 337
Final Gate Reviews and Stakeholder Signoff 337
Deployment Considerations 338
Enforcing Deployment Policies and Procedures 338

Data Governance in the Deployment Phase 339
Risk Management During Deployment 340
Issue Management and Incident Response 341
User Training and Enablement 342
Cross-functional Governance Alignment 342
Continuous Monitoring 343
Foundations of AI System Monitoring 343
Technical Monitoring Frameworks 344
Detecting Drift and Performance Degradation 344
Fairness and Ethical Monitoring 345
Defining Maintenance Schedules 346
Handling Emerging Risks 346
Human Oversight in Monitoring 347
Monitoring Third Parties 347
Summary 348
Exam Essentials 349
Review Questions 350

Chapter 11 AI Risk Management and Assurance 353

AI Risk Management and Forecasting 355
Proactively Identifying and Categorizing AI Risks 356
Risk Management in AI Supply Chains 356
Forecasting Secondary and Downstream Harms 358
Risk Forecasting Techniques 358
Embedding Risk Management into Governance Structures 359
Documentation and Assurance 360
Regulatory Considerations 360
AI System Audits and Testing 361
Performance Audits and Model Validation 361
Red Teaming for Adversarial Robustness 362
Threat Modeling for AI-specific Attack Surfaces 362
Security Testing for Confidentiality and Integrity 363
Regulatory and Compliance Audits 364
Real-time Testing and Continuous Monitoring 365
Integrating Testing into the AI Lifecycle 365
AI Audit Roles and Responsibilities 366
Audit Documentation and Evidence Retention 366
Audit and Testing Tooling and Automation 367
Incident Management and Disclosures 367
Identifying and Managing AI Incidents 368
Categorizing and Escalating Incidents 369
Documenting Incidents, Issues, and Risks 369

Collaborating with Stakeholders to Understand Root Causes 370
Transparency and Public Disclosures 371
Organizational Preparedness and Readiness 372
Summary 373
Exam Essentials 374
Review Questions 376

Chapter 12 Ongoing AI Operations 379

Managing Business Records 380
Incident Documentation and Tracking 380
Issue Reporting and Lifecycle Management 381
Risk Documentation and Mitigation Tracking 382
Post-market Monitoring Plans 382
Governance and Accountability Structures 383
Retention and Disposal Policies 384
Tools and Technologies for Business Record Management 384
Legal and Regulatory Considerations 385
Integration with Broader Organizational Practices 386
External Communications 386
Establishing a Communication Governance Structure 386
Communicating AI System Capabilities and Limitations 387
Regulatory and Legal Communication Obligations 388
Managing Public and Media Relations 388
User-facing Communications and Transparency 389
Partner and Vendor Communications 390
Communication Planning Across the AI Lifecycle 390
Measuring Communication Effectiveness 391
AI System Retirement 391
Understanding the Need for AI System Retirement 392
Building an AI System Retirement Policy 392
Planning for Controlled Deactivation 393
Localization and Regional Retirement Strategies 394
Data and Model Artifacts: Archiving and Disposal 395
Communication and Transparency in AI Retirement 395
Lessons Learned and Organizational Feedback Loops 397
Summary 397
Exam Essentials 398
Review Questions 399

Appendix

Appendix Answers to the Review Questions 401

Index *411*

Introduction

Congratulations on choosing to become an AI Governance Professional (AIGP)! Whether you have worked for several years in AI or have just recently been introduced to AI governance, don't underestimate the hard work and dedication required to obtain and maintain AIGP certification. Although ambition and motivation are essential, the rewards of being AIGP certified can far exceed the effort.

You probably never imagined you would find yourself working in AI or seeking a professional AI certification. The massive increase in AI across every profession, combined with many organizations adding AI capabilities, has led to your involvement in this field. Or, you noticed that AI-related career options are increasing exponentially, and you have decided to get ahead of the curve. Or perhaps you have responsibilities in your organization for governance, and you need to get a handle on the adoption of AI to ensure it's done safely and responsibly. You aren't alone! Welcome to the journey and the fantastic opportunities that await you.

I have compiled this information to help you understand the commitment needed, prepare for the exam, and maintain your certification. Not only do I wish you to prepare for and pass the exam with flying colors, but I also provide you with the information and resources to maintain your certification and proudly represent yourself and the professional world of AI with your new credentials.

The International Association of Privacy Professionals (IAPP) is a recognized leader in the field of privacy. Founded in 2000, this nonprofit organization has over 50,000 members and has several certifications, including:

- Artificial Intelligence Governance Professional (AIGP)
- Certified Information Privacy Professional (CIPP)
- Certified Information Privacy Manager (CIPM)
- Certified Information Privacy Technologist (CIPT)
- Privacy Law Specialist (PLS)
- Fellow of Information Privacy (FIP)

Some of IAPP's certifications have been accredited under ISO/IEC 17024:2012, which means that IAPP's procedures for accreditation meet international requirements for quality, continuous improvement, and accountability. For the current accreditation status of AIGP, consult the IAPP certification handbook or website.

If you're new to IAPP, we recommend visiting the organization's website (www.iapp.org) and familiarizing yourself with the available guides and resources. In addition, if you're near one of the 150+ local IAPP chapters in dozens of countries worldwide, consider contacting a chapter for information on local and regional meetings, training days, conferences, or study sessions. You will meet other AI and privacy professionals who can give you additional insight into the AIGP certification and the privacy profession.

The AIGP certification primarily focuses on the governance of AI models and systems, with an emphasis on privacy. It certifies the individual's knowledge of establishing and

managing a governance function, AI models, and privacy. Better organizations that utilize governance structures will recognize the need to incorporate AI capabilities into their governance sphere, ensuring safe, ethical, and legal outcomes. An AIGP-certified individual is a great candidate for these positions.

If you're preparing to take the AIGP exam, you'll undoubtedly want to find as much information as possible about AI and governance. The more information you have, the better off you'll be when attempting the exam. This study guide was written with that in mind. The goal was to provide enough information to prepare you for the test, but not so much that you'll be overloaded with information outside the scope of the exam.

Bloom's Taxonomy is a model for categorizing educational goals, as shown in Figure 1. IAPP exam questions focus primarily on the Remember, Understand, Apply, and Analyze levels.

This book presents the material at an intermediate technical level. Experience with and knowledge of AI, governance, privacy, and lifecycle processes will help you fully understand the challenges you'll face as an AI governance professional.

I've included review questions at the end of each chapter to give you a taste of what taking the exam is like. I recommend you check out these questions first to gauge your level of expertise. You can then use the book mainly to fill in the gaps in your current knowledge. This study guide will help you expand your knowledge base before taking the exam.

If you can answer 80% or more of the review questions correctly for a given chapter, you can feel safe moving on to the next chapter. If you're unable to answer that many correctly, reread the chapter and try the questions again. Your score should improve.

FIGURE 1 Bloom's Taxonomy.

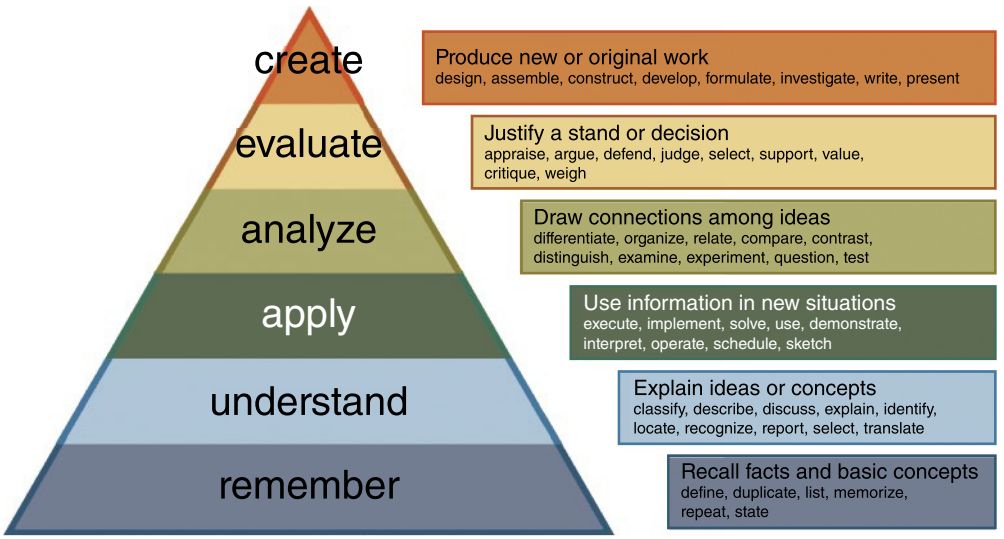


Image courtesy Tidema via Wikimedia Commons CC license



Don't just study the questions and answers! The questions on the actual exam may differ from those included in this book as practice questions. The exam is designed to test your knowledge of a concept or objective, so use this book to learn the objectives behind the questions.

Certification Requirements

To earn your AIGP certification, you are required to register for the AIGP exam, pay the exam fee, and agree to various IAPP terms and conditions, including the IAPP code of professional conduct and a confidentiality agreement, which appear in the IAPP certification handbook.

The exam costs $649 for IAPP members and $799 for non-members (subject to change; always confirm in the IAPP certification handbook). More details about the AIGP exam and how to take it can be found at `iapp.org/certify/aigp/`

IAPP does not require AIGP candidates to document and prove relevant work experience. The AIGP certification is strictly knowledge-based. That said, if you do not have any IT, cybersecurity, or privacy experience, you'll need to spend considerably more time studying for the exam, as you'll need to learn the fundamentals of governance and the fundamentals of AI models and systems.

The AIGP Exam

The AIGP exam is a vendor-neutral certification for business and IT professionals responsible for (or supporting) AI governance. While no prior experience is explicitly required for this certification, a background in IT, privacy, or cybersecurity governance will be particularly helpful, as the concepts of governance in those domains closely align with AI governance.

The exam covers four major domains:

1. The foundations of AI governance
2. How laws, standards, and frameworks apply to AI
3. How to govern AI development
4. How to govern AI deployment and use

These four areas include understanding the types of AI models, measuring and managing AI systems in production, privacy, and the systems development lifecycle. The exam places a strong emphasis on various AI model scenarios. You'll need to learn a lot of information, but you'll be well rewarded for possessing this credential. AI is taking the world by storm,

and better organizations will value the knowledge of leaders who understand how to build a governance function to ensure the best possible outcomes with new AI systems.

The AIGP exam consists solely of standard multiple-choice questions. Each question has four possible answer choices, and only one of those answers is correct. When taking the test, you'll likely find some questions where you think multiple answers might be correct. In those cases, remember that you're looking for the *best* possible answer to the question!

You'll have three hours to take the exam and be asked to answer 100 questions during that time. Your exam will be scored on a scale of 100–500, with a passing score of 300.

In-person Exams

IAPP partners with Pearson VUE testing centers, so your next step is to find a testing center near you. In the United States, you can do this based on your address or your postal code, while non-U.S. test takers may find it easier to enter their city and country. You can search for a test center near you at the Pearson VUE Exams website: `www.pearsonvue.com/us/en/iapp.html`

Once you know where you'd like to take the exam, set up a Pearson VUE testing account and schedule an exam on their site.

On the day of the test, bring a government-issued identification card or passport that contains your full name (exactly matching the name on your exam registration), your signature, and your photograph. Make sure to show up with plenty of time before the exam starts. Please note that you are not allowed to bring your notes, electronic devices (including earbuds, smartphones, and watches), or any other materials. Test centers I have visited have lockers for storing your personal effects, if you don't want to keep them in your vehicle.

Remote Exams

IAPP also offers online exam proctoring via OnVUE. Candidates using this approach will take the exam at home or in the office and be proctored remotely via webcam.

Due to the rapidly changing nature of the at-home testing experience, candidates wishing to pursue this option should check the OnVUE website for the latest details at: `https://home.pearsonvue.com/op/OnVUE-technical-requirements`

Exam candidates considering taking an at-home exam with a corporate-managed computer must carefully read these details, as many organizations place strict restrictions on permitted actions.

Develop a Study Plan

Whether this is your first professional certification, or if you have earned others, follow these steps to study and prepare for your AIGP exam.

- **Read the IAPP certification handbook:** This IAPP publication includes all the official steps required to earn IAPP certifications, including the AIGP.

- **Review or read this book in its entirety:** Depending on your background in governance or AI, read as much as you need to get an idea of the amount of study you'll need.
- **Self-assess:** Read the study questions in each chapter and take the practice exams online to confirm your knowledge level and identify areas of study.
- **Know your best learning methods:** Each person has a preferred learning style that works better for comprehension and retention and aligns with their overall lifestyle.
- **Consider my online video course:** Find my AIGP video training course at www.oreilly.com. This online course is a companion to this study guide.
- **Study iteratively:** Depending on your work experience in governance and AI, plan your study program to last at least two months, but no longer than six months. During this time, periodically take practice exams and note your strengths and weaknesses. Once you have identified your weak areas, focus on those areas weekly by rereading the related sections in this book, retaking practice exams, and noting your progress.
- **Avoid cramming:** You may find online or printed resources that involve last-minute cramming. Research suggests that exam cramming can lead to susceptibility to illness, sleep disturbances, overeating, and digestive issues. One thing is sure: many people find that good, steady study habits result in less stress and greater clarity and focus during exams.
- **Find or form a study group:** Many IAPP chapters and other organizations have formed specific study groups or offer less expensive exam review courses. Contact your local chapter to see whether these options are available to you. In addition, be sure to keep an eye on the IAPP website. Use your local network to find out whether there are other local study groups and helpful resources.
- **Schedule your exam:** After you decide whether to take your exam at a Pearson VUE test center or online at home or work, log on to iapp.org and purchase and schedule your exam.
- **Check the IAPP candidate guide:** The IAPP candidate guide is your source for the most up-to-date and official information about taking your exam, including test center rules, and what you must bring (including a government-issued photo ID that matches your exam registration) to the test center.
- **Logistics check:** A week or more before your exam, confirm the time and place of your exam, and confirm transportation, parking, and other details.
- **Sleep:** Be sure to get plenty of sleep in the days leading up to your exam. I recommend you relax the day before your exam to give your mind a rest from studying.

Day of the Exam

Follow these tips on the day of the exam.

- **Arrive early:** Plan to arrive early at the test center, just in case traffic or other unforeseen circumstances add to your travel time.

- **Observe test center rules:** Be sure you understand and comply with all test center rules. Be sure to ask questions of your exam proctor so that you can put your mind at ease during the exam.
- **Answer all exam questions:** Read each question carefully, but don't try to overanalyze. Select the best answer based on the AIGP job practice and the IAPP glossary. There may be more than one reasonable answer, but you are required to select the *best* answer. You should be able to flag any question you aren't sure about so you can return to it later. Make sure your pace is fast enough to comfortably get through the entire exam with time to review questions you flagged—but don't rush through it, or you'll be likely to misread some questions and answer them incorrectly.

If You Did Not Pass

Don't lose heart if you do not pass the exam. Instead, consider it a steppingstone to your eventual success! As soon as you leave the testing center, take some notes on what types of questions you feel you struggled with. Then, return to this book and my video study course and repeat those chapters and videos to reinforce your knowledge. Take at least a few weeks to study those topics and refresh your knowledge on other related subjects. Success is granted to those who persist and are determined.

After the AIGP Exam

Once you have taken the exam, you will be notified of your score immediately, so you'll know if you passed the test right away. You should keep track of your score report with your exam registration records and the email address you used to register for the exam.

Maintaining Your Certification

AI is a rapidly evolving field with new capabilities and models arising regularly. All AIGP holders are required to complete continuing professional education annually to maintain current knowledge and sharpen their skills. The guidelines around continuing professional education are somewhat complicated, but they boil down to two requirements:

- You must have a minimum of 20 hours of credit every year.
- You must submit evidence of your continuing education on the CPE portal of the IAPP website.

There are many acceptable ways to earn CPE credits, many of which do not require travel or attending a training seminar. The critical requirement is that you generally do not earn CPEs for work you perform as part of your regular job. CPEs are intended to cover professional development opportunities outside of your day-to-day work. You can earn CPEs in several ways, including:

- Attending conferences
- Attending training programs
- Attending professional association meetings and activities

- Taking self-study courses
- Attending vendor marketing presentations
- Teaching, lecturing, or presenting
- Publishing articles, monographs, or books
- Participating in the exam development process
- Volunteering with IAPP
- Mentoring

For more information on the activities that qualify for CPE credits, visit this site to download the latest IAPP CPE policy: `https://iapp.org/certify/cpe-policy`

Study Guide Elements

This study guide uses some common elements to help you prepare. These include the following:

Summaries The Summary section of each chapter provides a brief overview of the chapter, enabling you to easily understand its contents.

Exam Essentials The Exam Essentials focus on major exam topics and critical knowledge that you should take into the test. The Exam Essentials focus on the exam objectives provided by IAPP.

Chapter Review Questions A set of questions at the end of each chapter will help you assess your knowledge and whether you are ready to take the exam based on your understanding of that chapter's topics.

Additional Study Tools

This book includes additional study tools to help you prepare for the exam. They include the following.



Visit www.wiley.com/go/Sybextestprep to register and gain access to this interactive online learning environment and test bank, featuring study tools and resources.

Sybex Test Preparation Software

Sybex's test preparation software lets you prepare with electronic test versions of the review questions from each chapter, the practice exam, and the bonus exam included in this book.

You can build and take tests on specific domains, by chapter, or cover the entire set of AIGP exam objectives using randomized tests.

Electronic Flashcards

Our electronic flashcards are designed to help you prepare for the exam. Over 100 flashcards will ensure that you are familiar with critical terms and concepts.

Glossary of Terms

Sybex provides a glossary of terms in PDF format, allowing quick searches and easy reference to materials in this book.

Bonus Practice Exams

In addition to the practice questions for each chapter, this book includes two 100-question practice exams. We recommend using them both to test your preparedness for the certification exam.

AIGP Exam Objectives

IAPP publishes relative weightings for each exam objective. The following table lists the four AIGP domains and the extent they are represented on the exam.

Domain	% of Exam*
I. Understanding the foundations of AI governance	~23
II. Understanding how laws, standards, and frameworks apply to AI	~20
III. Understanding how to govern AI development	~29
IV. Understanding how to govern AI deployment and use	~28

*Figures are approximate, based on the AIGP Body of Knowledge and Exam Blueprint that provides ranges of the number of exam coverage in each domain and competency.

AIGP Certification Exam Objective Map

The AIGP body of knowledge is organized in a way that duplicates some topics across multiple domains; for example, Domain IV ("Understanding how to govern AI deployment and use") contains the performance indicator, "Identify laws that apply to the AI model." Because Domain II ("Understanding how laws, standards, and frameworks apply to AI") covers laws, we find it puzzling that AIGP put that performance indicator in Domain IV.

There are other such examples. Thus, this book's organization varies somewhat from the AIGP domains to a more logical sequence.

Table 1 illustrates how the domains, competencies, and performance indicators are addressed throughout this book.

TABLE 1 AIGP Domains, Competencies, and Performance Indicators Mapping to This Book

AIGP Domains, Competencies, and Performance Indicators	Chapter
Domain I: Understanding the foundations of AI governance	
I.A Understand what AI is and why it needs governance (all performance indicators)	1
I.B Establish and communicate organizational expectations for AI governance (all performance indicators)	2
I.C Establish policies and procedures to apply throughout the AI lifecycle (all performance indicators)	3
Domain II: Understanding how laws, standards, and frameworks apply to AI	
II.A Understand how existing data privacy laws apply to AI (all performance indicators)	4
II.B Understand how other types of existing laws apply to AI (all performance indicators)	5
II.C Understand the main elements of the EU AI Act (all performance indicators)	6
II.D Understand the main industry standards and tools that apply to AI (all performance indicators)	7
Domain III: Understanding how to govern AI development	
III.A Govern the designing and building of the AI model (all performance indicators)	8
III.B Govern the collection and use of data in training and testing the AI model (all performance indicators)	9
III.C Govern the release, monitoring, and maintenance of the AI model	10, 11
- Assess readiness and prepare for release into production	10

(*Continued*)

TABLE 1 (Continued)

AIGP Domains, Competencies, and Performance Indicators	Chapter
- Conduct continuous monitoring of the AI model and establish a regular schedule for maintenance, updates, and retraining	10
- Conduct periodic activities to assess the AI model's performance, reliability, and safety	11
- Manage and document incidents, issues, and risks	11
- Collaborate with cross-functional stakeholders to understand why incidents arise from AI models	11
- Make public disclosures with to meet transparency obligations	11
Domain IV: Understanding how to govern AI deployment and use	
IV.A Evaluate key factors and risks relevant to the decision to deploy the AI model	1, 4, 8, 10
- Understand the context of the AI use case	8
- Understand the differences in AI model types	1
- Understand the differences in AI deployment options	10
- Understand the requirements that apply to sensitive or special categories of data	4
IV.B Perform key activities to assess the AI model	5, 8, 11
- Perform or review an impact assessment on the selected AI model	8
- Identify laws that apply to the AI model	5
- Identify and evaluate key terms and risks in the vendor or open-source agreement	11
- Identify and understand issues that are unique to a company deploying its own proprietary AI model	8
IV.C Govern the deployment and use of the AI model	10, 11, 12
- Apply the policies, procedures, best practices, and ethical considerations to the deployment of an AI model	10

AIGP Domains, Competencies, and Performance Indicators	Chapter
- Conduct continuous monitoring of the AI model and establish a regular schedule for maintenance, updates, and retraining	10
- Conduct periodic activities to assess the AI model's performance, reliability, and safety	11
- Document incidents, issues, risks, and post-market monitoring plans	12
- Forecast and reduce risks of secondary or unintended uses and downstream harms	11
- Establish external communication plans	12
- Create and implement a policy and controls to deactivate or localize an AI model as necessary	12

Assessment Test

1. An AI-driven medical diagnostic system misclassifies a patient's condition due to biased training data. Which principle of responsible AI is most directly violated?
   - **A.** Transparency
   - **B.** Fairness
   - **C.** Security
   - **D.** Human-centricity

2. An autonomous delivery robot is programmed to make decisions without human intervention. What governance safeguard should be in place to manage this AI system?
   - **A.** Implementation of a human-in-the-loop mechanism
   - **B.** Strict limitation of opacity
   - **C.** Use of supervised learning only
   - **D.** Elimination of data dependency

3. A company uses generative AI to design marketing materials trained on copyrighted online content. What AI characteristic primarily drives the need for governance in this case?
   - **A.** Probabilistic outputs
   - **B.** Complexity
   - **C.** Speed and scale
   - **D.** Data dependency

4. An organization's AI governance board notices recurring gaps between technical and legal interpretations of "acceptable model risk." Which structural action would best strengthen cross-functional collaboration?
   - **A.** Assign each department its own separate governance policy.
   - **B.** Require periodic joint review meetings and shared documentation templates.
   - **C.** Eliminate legal participation in technical discussions to reduce friction.
   - **D.** Move all governance decisions to the IT department.

5. A mid-sized financial services company wants to expand its AI program. Which governance maturity action should it prioritize next?
   - **A.** Create a lightweight checklist with no documentation requirements.
   - **B.** Focus exclusively on model accuracy metrics.
   - **C.** Formalize lifecycle checkpoints such as fairness reviews and deployment approval gates.
   - **D.** Replace the governance council with ad hoc task forces.

6. In a global healthcare enterprise, the AI governance team includes executives, clinicians, and data scientists. Which factor most justifies maintaining this diverse membership?
   - **A.** Diversity increases model training speed.
   - **B.** Cross-functional expertise ensures balanced evaluation of technical, ethical, and clinical risks.
   - **C.** It satisfies minimum staffing requirements for compliance audits.
   - **D.** It reduces total program cost through redundancy elimination.

7. An organization integrates AI governance into its existing IT governance framework rather than managing them separately. What is the primary benefit of this combined approach?
   - **A.** It avoids duplication and embeds AI oversight into established processes.
   - **B.** It eliminates the need for privacy and security policies.
   - **C.** It prevents non-technical staff from influencing AI decisions.
   - **D.** It removes the requirement for lifecycle documentation.

8. During a vendor assessment, a company discovers that a third-party AI model uses publicly scraped data with uncertain consent. Which governance control would best address this risk?
   - **A.** Requiring the vendor to provide documentation of data sources and licensing
   - **B.** Allowing deployment since the data is publicly available
   - **C.** Ignoring data provenance if the model performs accurately
   - **D.** Adding a confidentiality clause unrelated to data sourcing

9. An AI chatbot trained on historical customer interactions begins generating responses that reveal fragments of personal data. Which policy area most urgently needs revision?
   - **A.** Access control and password management
   - **B.** Network performance optimization
   - **C.** Facilities security and visitor access procedures
   - **D.** Traditional data retention and anonymization standards

10. An AI system trained on social media text begins predicting users' religious beliefs, even though this attribute was never directly collected. What obligation does this create under privacy law?
    - **A.** None, since the system never collected explicit religion data.
    - **B.** The inference counts as processing of special category data and triggers GDPR Article 9 safeguards.
    - **C.** Only a transparency notice update is required.
    - **D.** It qualifies as anonymous data because it was inferred.

11. A multinational bank relies on a cloud-based AI model hosted in another region. What must the controller do to comply with GDPR transfer obligations?

- **A.** Encrypt model outputs but skip documentation.
- **B.** Obtain verbal consent from customers for each transfer.
- **C.** Conduct a Transfer Impact Assessment (TIA) and ensure valid safeguards such as SCCs or BCRs.
- **D.** Rely on processor confidentiality clauses alone.

12. A company developing an emotion-recognition tool processes voice recordings containing biometric and health cues. Which governance control is most essential before deployment?

- **A.** Marketing approval from the product team
- **B.** Execution of a Data Protection Impact Assessment (DPIA) addressing risks to data subjects
- **C.** Annual employee privacy training
- **D.** Generic system penetration testing

13. A generative-AI platform trains on millions of online articles under the assumption that scraping constitutes "fair use." Which factor would most weaken its fair-use defense?

- **A.** The model's outputs are non-commercial.
- **B.** The model's training is transformative rather than duplicative.
- **C.** The system reproduces recognizable portions of protected works.
- **D.** The data come from publicly accessible sources.

14. An AI-based credit-scoring tool uses a consumer's postal code as a model feature. Under anti-discrimination law, what is the primary concern?

- **A.** Postal code can act as a proxy for protected classes, producing disparate impact.
- **B.** The feature may reduce model accuracy.
- **C.** The feature violates open-banking data-sharing rules.
- **D.** The feature requires explicit customer consent under the FCRA.

15. A home-robot manufacturer releases a firmware update that introduces erratic movement, causing property damage. Which legal theory most likely applies?

- **A.** Manufacturing defect
- **B.** Design defect
- **C.** Failure to warn
- **D.** Breach of warranty

16. An AI system used to rank citizens based on their social behavior would violate which element of the EU AI Act?
    - **A.** Limited-risk AI disclosure obligations
    - **B.** High-risk AI documentation requirements
    - **C.** Prohibited AI classification
    - **D.** Systemic-risk designation for GPAI

17. Before a high-risk AI system can be marketed within the EU, what must its provider complete?
    - **A.** Certification from the European AI Office only
    - **B.** A conformity assessment and CE marking process
    - **C.** An EU-level procurement review
    - **D.** Submission of its algorithm source code to regulators

18. A company outside the EU develops a general-purpose AI model that is integrated into multiple EU applications. Which step ensures compliance with the EU AI Act's legal requirements?
    - **A.** Applying for an EU data adequacy determination
    - **B.** Requesting exemption from systemic-risk classification
    - **C.** Registering the model only in the developer's home country
    - **D.** Appointing an EU-based legal representative and maintaining conformity documentation

19. Which statement best describes the relationship between the NIST AI RMF and NIST ARIA?
    - **A.** ARIA replaces the AI RMF as a legally required framework for federal agencies.
    - **B.** The AI RMF provides strategic guidance, while ARIA offers tools to operationalize it.
    - **C.** The AI RMF applies only to cybersecurity, while ARIA applies to AI ethics.
    - **D.** ARIA focuses on model training, while the AI RMF addresses licensing.

20. According to ISO/IEC 42001, what is the purpose of integrating AI governance with existing management systems, such as ISO/IEC 27001 or ISO 9001?
    - **A.** To streamline certification processes and promote unified governance
    - **B.** To eliminate redundant controls by removing human oversight
    - **C.** To replace cybersecurity management with AI management
    - **D.** To prioritize algorithmic speed over quality management

21. How does ISO/IEC 22989 help organizations engaged in AI governance?

- **A.** It defines a legal enforcement framework for AI-related misconduct.
- **B.** It establishes certification rules for high-risk AI applications.
- **C.** It provides a standard lexicon to ensure consistency in terminology and communication.
- **D.** It mandates disclosure of proprietary algorithms to regulators.

22. During the design phase, an organization selects a highly complex deep learning model even though a simpler one would meet performance goals. Which governance issue does this decision most directly raise?

- **A.** Transparency and explainability risk
- **B.** Data sovereignty compliance
- **C.** Vendor lock-in
- **D.** Infrastructure scalability

23. A financial institution completes its AI impact assessment but fails to implement the mitigation steps it identified. What governance failure does this represent?

- **A.** Poor system documentation
- **B.** Lack of operationalization and review of assessment results
- **C.** Overclassification of model risk
- **D.** Duplication of regulatory filings

24. In designing a proprietary AI system, which practice most effectively reduces downstream legal and operational risk?

- **A.** Allowing data scientists to change model parameters without recordkeeping
- **B.** Using internal data without confirming its lawful basis for processing
- **C.** Outsourcing data collection to anonymous contractors
- **D.** Embedding documentation, testing, and explainability into the design process

25. A company trains an AI model using data purchased from a third-party broker. Later, it discovers that the broker lacked the legal authority to sell the data. Which governance control should have prevented this issue?

- **A.** Bias and fairness testing
- **B.** Verification of lawful data rights and provenance documentation
- **C.** Integration testing before deployment
- **D.** Regularization during model training

26. An AI system trained on clean, representative data begins producing biased results several months after deployment. What governance issue is the most likely cause?

- **A.** Lack of data drift monitoring and retraining protocols
- **B.** Overfitting during initial training
- **C.** Failure to anonymize personal data
- **D.** Use of synthetic data augmentation

27. During model validation, developers notice that results differ each time the model is retrained, even with identical data. Which testing or governance control addresses this risk?

- **A.** Ensuring model determinism through seed control and documented configurations
- **B.** Expanding the dataset with more examples
- **C.** Applying dropout regularization
- **D.** Replacing validation testing with performance testing

28. An AI customer support chatbot begins generating inaccurate responses after six months in production. Which governance control should have detected this issue earlier?

- **A.** Regular penetration testing
- **B.** Human-in-the-loop decision approval
- **C.** Scheduled retraining and performance monitoring
- **D.** Fine-tuning with anonymized data

29. Before deploying a high-risk AI model, a governance team confirms that documentation, risk mitigation plans, and monitoring plans are complete. What governance mechanism is being applied?

- **A.** Post-market surveillance
- **B.** Final gate review and stakeholder signoff
- **C.** Shadow modeling for drift detection
- **D.** Automated rollback testing

30. A healthcare provider deploys an AI diagnostic tool that runs on hospital devices without requiring cloud access. Which deployment model does this represent, and why is it appropriate?

- **A.** Cloud deployment—ensures centralized control and scalability
- **B.** Edge deployment—provides low latency and preserves patient privacy
- **C.** On-premises deployment—simplifies external vendor management
- **D.** Hybrid deployment—balances accuracy and interpretability

31. An organization integrates an open-source AI model licensed under AGPL into its commercial product. What is the most significant risk if the licensing terms are not appropriately reviewed?

A. The organization may be required to disclose proprietary source code.

B. The model could fail to meet expected performance benchmarks.

C. Users could access the model through unauthorized APIs.

D. The license automatically converts to the public domain after release.

32. During an AI system audit, a red team successfully manipulates a chatbot into revealing fragments of its training data. Which risk domain does this finding most directly relate to?

A. Operational risk

B. Ethical and social risk

C. Deployment risk

D. Data-related risk

33. A financial institution's fraud detection model begins incorrectly flagging legitimate transactions following a data feed update. What assurance practice would most likely have prevented this?

A. Periodic retraining using customer feedback

B. Continuous monitoring with drift detection alerts

C. Limiting model access to authorized employees only

D. Increasing the volume of test data

34. An AI-driven hiring platform is being retired due to new legal requirements under the EU AI Act. What is the most critical first step in the deactivation process?

A. Delete all model artifacts immediately to prevent misuse.

B. Replace the model with a temporary open-source version.

C. Notify affected stakeholders and plan for controlled deactivation.

D. Suspend all employee access without communication.

35. A global company operates an AI translation system that performs differently across languages due to cultural differences. To maintain compliance and effectiveness, which governance strategy is most appropriate?

A. Implement a single universal model trained on all languages equally.

B. Disable translations in low-accuracy regions.

C. Localize and retrain regional models to meet linguistic and cultural requirements.

D. Outsource compliance entirely to regional vendors.

36. An organization's AI monitoring system detects a recurring issue that doesn't meet the threshold for an incident. What is the most appropriate governance response?

A. Ignore the issue until an incident occurs.

B. Record it in the issue tracking system for trend analysis and escalation if repeated.

C. Immediately shut down the affected AI model.

D. Report it publicly as a compliance event.

Answers to Assessment Test

1. B. Fairness requires that AI systems produce equitable outcomes and avoid discrimination caused by biased data. When an AI diagnostic tool generates inconsistent or discriminatory results, it reflects bias present in the training dataset. Transparency helps explain the AI system's decision, but fairness is the principle that ensures equitable treatment for all individuals.

2. A. Autonomous systems introduce governance challenges because they act independently. A human-in-the-loop mechanism ensures oversight and allows intervention when autonomous behavior produces unsafe or unintended outcomes. This control supports accountability and reduces operational and ethical risks.

3. D. AI systems rely heavily on large volumes of training data, which may include copyrighted or proprietary material. This dependence introduces legal and ethical risks, making data dependency a key governance concern. Effective AI governance frameworks require processes to verify data provenance, manage intellectual property exposure, maintain data lineage, and ensure lawful use of all training inputs.

4. B. Cross-functional collaboration depends on shared processes and transparent communication. Regular joint review meetings supported by standard templates ensure alignment between technical precision and regulatory requirements. Separate or siloed policies create inconsistencies, and removing stakeholders can reduce oversight rather than improve it.

5. C. Organizations transitioning from informal to managed maturity should embed governance checkpoints throughout the AI lifecycle. Structured reviews, including bias testing and documentation gates, institutionalize accountability and readiness. Accuracy alone is insufficient, and ad hoc governance erodes traceability.

6. B. AI governance in regulated sectors requires multiple perspectives to detect complex risks. Including technical, operational, and ethical experts ensures decisions account for safety, fairness, and compliance obligations. Diversity of expertise is a risk-control mechanism.

7. A. Integrating AI governance into existing IT governance creates operational efficiency by leveraging existing review, approval, and documentation processes. These processes ensure AI-specific controls, such as fairness testing and model monitoring, fit within established workflows, reducing duplication and improving consistency and scalability across systems and processes.

8. A. Vendor transparency regarding training data sources and usage rights is an essential AI governance control. Policies and contracts should require vendors to disclose provenance and licensing to ensure compliance with data protection and intellectual property laws. The public availability of data does not exempt individuals from the obligation to verify lawful use.

9. D. Leakage of memorized personal data from AI outputs highlights weaknesses in traditional data retention and anonymization policies. AI governance must expand these controls to include differential privacy, data minimization, and lifecycle-based retention, recognizing that models can unintentionally preserve sensitive information long after data ingestion.

10. B. Under GDPR Article 9, processing special-category (sensitive) data, including religious or political beliefs, is prohibited unless a lawful exception applies. Inferred attributes are still personal data if they reveal protected categories. Controllers must identify and document a legal basis, conduct a DPIA, and implement technical and organizational safeguards to prevent discrimination or misuse.

11. C. Cross-border AI processing constitutes an international data transfer under GDPR. Controllers must assess the destination country's privacy laws, apply an approved safeguard (typically Standard Contractual Clauses—SCCs or Binding Corporate Rules—BCRs), and document their findings in a TIA. These safeguards ensure lawful transfer and accountability for data processed outside the European Economic Area (EEA).

12. B. Processing biometric or health data represents a high-risk activity under GDPR Article 35 and mandates a DPIA. The assessment documents processing purpose, data sensitivity, potential harms, and mitigation measures such as encryption, consent management, and bias testing. Without a DPIA, controllers cannot demonstrate compliance or justify the lawful basis for handling sensitive data.

13. C. Fair use depends partly on whether the new use is transformative and whether it affects the market for the original work. If an AI system reproduces identifiable or substantial parts of copyrighted material, the use is not transformative and undermines the fair-use claim—even if the data were publicly available.

14. A. Postal codes often correlate with protected traits such as race, income, and other characteristics. Even without intent to discriminate, their inclusion can result in disparate impacts prohibited by laws such as the Equal Credit Opportunity Act (ECOA) and the Fair Housing Act. Models must test and justify any variable that could serve as a proxy for protected traits.

15. A. When a product deviates from its intended design due to an error during production or update deployment, it constitutes a manufacturing defect. Even if the original design was sound, an improperly executed firmware change that causes harm triggers strict liability under product liability law. If the update was improperly executed, it is considered a manufacturing defect; if the update design itself was unsafe, it could be a design defect. For exam purposes, a manufacturing defect is the best fit.

16. C. Social scoring systems by governments are explicitly prohibited under the EU AI Act. These systems are considered incompatible with EU values and fundamental rights because they create discriminatory and manipulative effects that cannot be mitigated by controls or transparency.

17. B. High-risk AI systems must undergo conformity assessment (either with a qualified internal resource or via a third-party Notified Body) and achieve CE marking before entering the market. This process verifies compliance with safety, transparency, and oversight obligations. Regulators do not require the disclosure of proprietary source code, but they do require traceable documentation and auditability.

18. D. The EU AI Act has extraterritorial reach. Non-EU providers placing AI systems on the EU market must designate a legal representative within the EU to liaise with regulators, ensure conformity assessments, and maintain documentation. This requirement extends to general-purpose AI (GPAI) models integrated into downstream EU-facing applications.

19. B. The NIST AI Risk Management Framework (AI RMF) defines strategic goals and functions (Govern, Map, Measure, and Manage) for managing AI risk. The NIST Assurance of AI Systems (ARIA) program builds on it by providing concrete tools, metrics, and implementation pathways to translate those principles into operational practice, particularly for public-sector organizations. ARIA is a research initiative providing methods and metrics to operationalize the AI RMF; it does not replace or supersede the RMF.

20. A. The ISO/IEC 42001 standard mirrors the structure of other ISO standards, allowing organizations to integrate AI governance into established quality or information security management systems. This approach consolidates documentation, audits, and improvement processes, strengthening organizational coherence while reducing redundancy.

21. C. ISO/IEC 22989 standardizes definitions of key AI terms, such as transparency, autonomy, and explainability, thereby creating a shared vocabulary for developers, policymakers, and auditors. Adopting these standard terms reduces ambiguity in technical documentation and contracts, thereby promoting cross-sector and cross-border alignment in AI governance.

22. A. Overly complex models often reduce explainability and transparency. Governance frameworks emphasize model interpretability as a cornerstone of accountability, particularly in regulated sectors. When simpler, equally effective alternatives exist, complexity adds unnecessary governance burden and can obscure bias or error analysis. Additional complexity also increases the attack surface, potentially making it more difficult to protect such a system from malicious input and other forms of attack. Finally, additional complexity increases the governance burden by requiring additional effort to assess, monitor, and mitigate issues.

23. B. Impact assessments are only effective if their findings are reviewed, approved, and integrated into development and deployment processes. Failing to act on assessment results undermines the purpose of the AIA and reflects a breakdown in the feedback loop essential to continuous governance improvement. Note that a decision to accept any individual risk is permissible, provided it is made in accordance with defined governance roles and responsibilities.

24. D. Governance, testing, and explainability controls included in the design phase create defensibility and traceability throughout the lifecycle. Proper documentation and transparency enable organizations to demonstrate compliance and respond to audits or litigation.

25. B. Before any dataset is used for AI development, governance controls must verify lawful collection, usage rights, and provenance documentation. These controls include reviewing contracts, consent records, and regulatory compliance. Failure to confirm these rights exposes organizations to legal and ethical liabilities, even when the model performs correctly.

26. A. Data drift occurs when live input data diverges from the distribution of training data, degrading performance and fairness. Governance frameworks should require ongoing monitoring and retraining to ensure the continued validity of AI models. Without these checks, models can silently accumulate bias or inaccuracies over time.

27. A. Reproducibility is a cornerstone of responsible AI governance. Controlling random seeds, recording hyperparameters, and documenting configurations ensure that models produce consistent results across retraining. These efforts support auditability, transparency, and regulatory defensibility when AI decisions are challenged.

28. C. Continuous monitoring and scheduled retraining are core governance practices to maintain AI system performance and accuracy over time. As data distributions and user interactions evolve, models degrade without routine evaluation. Governance frameworks should require regular retraining intervals and performance thresholds to detect drift or loss of accuracy before it harms customers.

29. B. Final gate reviews are structured checkpoints to ensure all regulatory, technical, and ethical requirements are met before release. This governance step provides traceability, formal approval, and accountability from all stakeholders—including legal, engineering, business, and risk—before production deployment.

30. B. Edge deployment processes data locally, minimizing latency and protecting sensitive information by avoiding data transmission to cloud servers. This approach aligns with healthcare privacy and safety obligations, where both speed and confidentiality are critical governance priorities. On-premises would have been the next-best strategy and may be used when edge devices constrain or prevent edge deployment. On-premises deployment also enhances privacy, but edge deployment is preferred for latency-sensitive healthcare use.

31. A. The GNU Affero General Public License (AGPL) requires that derivative works and network-accessible implementations make their source code available to users. Failing to comply exposes the organization to intellectual property violations, threatening reputational risk. A legal review of open-source license terms (and all other relevant legal agreements) is a critical element of AI vendor and component risk management.

32. D. Exposing fragments of training data represents an AI-specific data-related risk, involving data leakage and privacy breaches. Red team testing helps uncover these vulnerabilities before adversaries can exploit them. Strong data governance and differential privacy mechanisms are key controls to prevent the exposure of training data.

33. B. Continuous monitoring with statistical drift detection can identify shifts in data distribution or model outputs that could degrade accuracy. Detecting drift early allows organizations to retrain or recalibrate models before significant harm occurs. Continuous monitoring is a key principle of AI assurance and lifecycle governance.

34. C. Responsible AI retirement begins with planning and stakeholder communication. Controlled, methodical deactivation ensures that users, partners, and regulators are informed, fallback systems are in place, and audit trails are preserved. Immediate deletion or suspension without planning risks operational disruption and non-compliance.

35. C. Localization helps AI systems align with regional laws, data residency rules, linguistic nuances, and cultural expectations. Retraining AI models on local data improves fairness, accuracy, and contextual sensitivity. A single, universal AI model may fail to accurately reflect regional linguistic nuances or local regulatory environments.

36. B. Higher governance maturity requires capturing lower-severity issues to identify emerging risk patterns. Issue tracking systems enable categorization, assignment, and escalation when trends signal potential degradation or harm. Proactive documentation supports continuous improvement and early risk detection.

PART I

Foundations of AI and Governance





Chapter

1

AI and AI Governance

THIS CHAPTER COVERS AIGP DOMAIN I, "UNDERSTANDING THE FOUNDATIONS OF AI GOVERNANCE," SUBDOMAIN I.A, "UNDERSTAND WHAT AI IS AND WHY IT NEEDS GOVERNANCE." THE FOLLOWING TOPICS ARE COVERED:

- ✔ Definitions and types of AI systems
- ✔ The types of risks and harms that AI systems may cause
- ✔ Why AI systems are different from other systems, and why governance is needed
- ✔ The principles of responsible AI

THE AI GOVERNANCE PROFESSIONAL (AIGP) PERFORMANCE INDICATORS IN THIS SUBDOMAIN INCLUDE:

- ✔ Know the generally accepted definitions and types of AI.
- ✔ Understand the differences in AI model types.
- ✔ Identify the types of risks and harms posed by AI to individuals, groups, organizations and society (e.g., misalignment with objectives, ethics and bias risk, and complexity and scalability).
- ✔ Identify the unique characteristics of AI that require a comprehensive approach to governance (for example, complexity, opacity, autonomy, speed and scale, potential for harm or misuse, data dependency, and probabilistic versus deterministic outputs).
- ✔ Identify and apply the common principles of responsible AI (for example, fairness, safety and reliability, privacy and security, transparency and explainability, accountability and human-centricity).



In all its forms, AI brings unprecedented capabilities to organizations in every business sector to improve efficiency and quality. But, like other emerging technologies, organizations are apt to rush into significant, impactful projects to bring about sweeping business transformation, even before they understand its capabilities and risks. This chapter includes a brief survey of the types of AI systems, with discussions on capabilities, risks, and reasons why organizations should develop a governance structure to ensure their AI systems deliver the results they expect.

The Types of AI

Artificial Intelligence (AI) is an interdisciplinary field that combines computer science, statistics, mathematics, and engineering to create systems capable of performing tasks that typically require human intelligence. The primary goal of AI is to design machines that can think, learn, and problem-solve autonomously.

This section includes definitions and examples of various types of AI.

Definition of AI

Artificial Intelligence (AI) is the simulation of human intelligence processes by machines, generally computer systems. These processes include learning, reasoning, problem-solving, perception, and language understanding.

Some AI systems are designed to perform specific tasks, such as image recognition and generation, speech processing, natural language understanding, decision-making, and more. For instance, tools like HeyGen create realistic videos of a specific subject from a photograph and a script.

Other AI systems are designed to perform a broad range of tasks, including large language models (LLMs) like Claude from Anthropic and Microsoft Copilot. For example, virtual assistants like Siri and Alexa use AI to interpret spoken language, understand requests, and provide relevant responses.

AI Capabilities

AI is often categorized into levels of capability, based on the extent of the system's ability to replicate human cognitive abilities. These are:

- *Narrow AI* (or *Weak AI*) refers to a class of AI systems designed and trained to perform a specific or narrow set of tasks. These systems do not possess general intelligence or consciousness and focus only on one function type. One example is IBM Deep Blue, an AI designed to play chess at an expert level. Another example is the *large language model (LLM)* AI (also known as *generative AI*), as implemented in ChatGPT, Claude, Grok, and others.
- An *AI agent* is an autonomous system that perceives its environment through incoming sensor information and then takes action to achieve specific goals.
- *General AI* (or Artificial General Intelligence—AGI) is a form of AI that can understand, learn, and apply intelligence across various tasks, much like a human being can. AGI is capable of solving nearly any problem it encounters, just as humans can. For example, an AI system capable of managing all or part of an organization, including performing tasks like strategic decision-making and customer service, and even creative tasks like writing or creating product or process designs.
- *Artificial Superintelligence (ASI)* (or *Superintelligent AI*) is an AI that surpasses human intelligence in all aspects—creativity, problem-solving, emotional understanding, and decision-making. While this level of AI does not yet exist, it is often discussed in AI research and ethical discussions, due to its potential risks and benefits.
- *God-like AI* is an AI at a level beyond Artificial Superintelligence (ASI), where AI capabilities exceed human cognition so profoundly that they become incomprehensible to humans across all domains.



As of this writing, no AGI or Superintelligent AI system has been disclosed to the general public.

AI Functionalities

AI systems are also categorized in terms of their functionality, based on the types of tasks or processes they can perform. These include:

- *Reactive machines* are AI systems designed to respond to specific stimuli or inputs. They do not retain memories of past interactions, nor do they learn from them. Reactive machines operate based on pre-programmed rules and cannot improve on their own. An example is IBM's Deep Blue chess player, which is trained to play chess at an expert level but does not remember past games.

- *Limited memory AI* systems with limited memory retain information for short periods and use that information to improve their performance. They learn from past experiences or historical data to enhance their decision-making capability. Still, because their memory is limited, they typically discard information when it is no longer relevant. For example, autonomous vehicles use limited memory AI to process traffic data, road conditions, and previous driving patterns. They use this information for navigation and decisions about speed and direction, but they soon discard information due to limited resources.
- *Reasoning AI systems* possess advanced abilities to analyze and correlate patterns and make informed predictions. Most common LLMs today fall in this category. They are even capable of passing standardized tests (for example, bar testing) with performance exceeding that of the average human.
- *Theory of mind AI* systems are a theoretical, advanced form of AI. This AI could understand and simulate human emotions, beliefs, intentions, and thoughts. In the future, an AI system could interact with humans in a manner similar to human relationships, including understanding and responding to emotions. A fictional example is the HAL-9000 computer in the movie *2001: A Space Odyssey*.
- *Self-aware AI* is the highest level of AI that possesses consciousness and self-awareness. Self-aware AI would be able to recognize itself in its environment and understand its own existence. Like superintelligent AI, self-aware AI is currently a theoretical concept. An example of self-aware AI would be a humanoid robot capable of introspection, understanding its own limitations, and even questioning its purpose.
- *Robotics* is a type of AI system capable of performing kinetic tasks, such as carrying and moving objects, and more advanced tasks, such as meal cooking.

AI Techniques

Another way to classify AI systems is by their techniques or methods to perform tasks. These methods help us understand how AI systems work. These include:

- *Machine learning (ML)* is a technique that enables systems to learn from data, improve their performance, and make decisions without explicit programming. In ML, *algorithms* are trained on large datasets and "learn" to make predictions or decisions based on patterns identified in the data. Spam filters, for example, use ML to analyze past emails and identify patterns (such as specific words or sender behaviors) to help classify new emails as spam.
- *Deep learning* is a subset of machine learning that uses artificial neural networks consisting of multiple layers. A deep learning system is designed to process large volumes of data, often unstructured, and is particularly effective at tasks such as speech and image recognition. A good example is Google Translate, which can translate spoken or written text into different languages.

- *Natural language processing (NLP)* systems can understand, interpret, and generate human language. NLP combines linguistics and machine learning to process text and speech, making interactions between humans and machines more intuitive. An example is ChatGPT, a chatbot that uses NLP to process and respond to user inputs in human language.
- *Supervised learning* is a type of AI system training that uses labeled data. For example, an ML-powered spam filter is trained with examples of spam messages and legitimate messages. In another example, a robot is trained with a dataset of images to later identify various types of objects.
- *Unsupervised learning* is a type of AI system training that uses unlabeled data. The AI system is expected to detect patterns on its own. For example, given a customer relationship database, an AI system is directed to classify it as it sees fit.
- *Semi-supervised learning* is a type of AI system training that uses a small amount of labeled data and a large amount of unlabeled data. The AI system first learns from the labeled data and is then trained on the unlabeled data to predict labels. Over time, performance will improve. For example, an organization is training an AI-powered image recognition system but only has a small collection of images to train. The AI system is trained on the labeled data and continues learning with the unlabeled data.
- *Reinforcement learning* is a type of machine learning where an AI agent learns through interaction and receives feedback in the form of rewards or penalties. The objective is for the AI system to maximize its rewards through trial and error. An example is a robot that learns to navigate a maze.
- *Expert systems* are AI systems that simulate the decision-making ability of a human expert in a particular field, using rules and knowledge bases to make inferences and solve complex problems. For example, a medical diagnostic system can analyze patient symptoms and medical history to suggest a diagnosis or treatment.
- *Zero-shot learning (ZSL)* is a machine learning use case in which an AI model is trained to recognize and categorize objects or concepts without having seen any examples of those categories or concepts beforehand.

Understanding these and other definitions is critical for navigating and governing AI. Because some AI governance professionals do not have deep technical knowledge, they must understand concepts like these to help them achieve better credibility with their technical peers, while striving toward better business outcomes.

Risks and Harms Posed by AI

AI has brought transformative changes across many industries by enhancing efficiency, creating new opportunities, and solving complex problems. However, as AI becomes more integrated into business and society, it introduces risks and potential harms. Understanding these risks is essential for professionals managing and governing AI systems.

Let's explore the various types of risks and harms that AI can pose to individuals, groups, organizations, and society. Each risk is explained with examples to provide clarity and context.

Risks and Harms to Individuals

The potential risks and harms to individuals take on two forms: when individuals interact directly with AI and when organizations use AI-driven processing and decision-making with information about individuals.

- **Privacy violations:** One of the most significant risks AI poses to individuals is the potential for privacy violations. AI systems, particularly those that rely on large datasets and *Retrieval Augmented Generation (RAG)* systems, can expose sensitive personal information, resulting in privacy breaches. An AI that has been trained on or has access to personal or *personally identifiable information (PII)* without anonymization can potentially leak sensitive data when prompted with related queries. For example, if an AI has been exposed to usernames, passwords, or email addresses, it might be able to reproduce this information in its responses. Similarly, it could attempt to predict a user's login details when asked. Another example can be facial recognition systems used by retail stores to track customer movements. These systems could collect biometric data without consent, leading to concerns about surveillance and personal privacy.
- **Bias and discrimination:** AI systems can perpetuate or even amplify existing societal biases, which can harm individuals, particularly those from marginalized groups. This occurs when AI models are improperly designed or trained on biased data. Recent examples include AI-powered decision-making systems that help screen and qualify employment candidates, loan applicants, and financial aid enrollees. Organizations are often unaware of the bias in their systems until victims speak out.
- **Cybersecurity risks:** Many IT organizations do a substandard job of protecting their systems and the sensitive data they store and process. This seems more common when organizations introduce new technologies like AI—they get excited about bringing in the new toy without adequate thought on protecting it. Individuals become victims of identity theft, financial loss, and the exposure of their private information. As organizations race to build and implement AI systems, fundamental security practices are often being overlooked. We are seeing organizations fail big; from training and fine-tuning AI on non-anonymized data to deploying it in systems that lack proper security controls and violate principles like least privilege and segregation of duties. Ultimately, it's the individual who pays the price, as their privacy and data security are often neglected.
- **Loss of visibility, autonomy, and control:** AI-powered decision-making and processing systems often take humans out of the loop, reducing a person's autonomy and ability to influence outcomes. Healthcare and financial services are two industries where people

feel like they are being treated like a number instead of a person. For instance, AI-driven diagnostic tools might override a healthcare provider's decision or misinterpret a patient's condition, causing the individual to lose control over their treatment plan and potentially leading to adverse health outcomes. This situation can occur further up the food chain when a health insurance company must approve a procedure or treatment before it can commence.

I'm just scratching the surface here. As AI proliferates into more organizations and performs more functions, I think it is likely that we will see more risks like these.

Risks and Harms to Groups

Beyond the impact on individuals, AI can also exact significant consequences for groups, such as communities, families, and social networks. AI poses risks to groups that tend to be more systemic, with far-reaching social implications.

- **Amplification of social inequality:** AI systems that operate at scale can exacerbate existing inequalities. Decision-making AI systems that derive profile data from social media can exhibit biases that affect specific groups. For example, an AI-powered loan qualification system might offer higher interest rates or deny loans to individuals in particular communities or groups.
- **Erosion of social cohesion:** AI-powered platforms may inadvertently create echo chambers with little diversity of ideas, favoring sensational content that increases polarization and reduces tolerance of viewpoints. For example, some social media platforms may prioritize emotionally charged content, creating more division and eroding societal trust.
- **Cybersecurity risks:** AI adoption is more widespread in critical infrastructure (transportation, utilities, healthcare) and business. As stated earlier, organizations are not great at securing new technologies like AI until security problems occur. For example, a large bank might implement an AI-driven chatbot that can be tricked into revealing information about other customers. Cybersecurity-related risks are discussed throughout this book.

As before, these are notable examples but not the only ones.

Risks and Harms to Organizations

Organizations are adopting AI systems to improve efficiencies, quality, and decision-making and to reduce certain operational costs. AI poses unique risks to the organizations using them, including:

- **Operational risks:** Organizations relying on AI for critical processes, including inventory management, fraud detection, and customer service, are vulnerable to disruptions if the

AI system fails or behaves unpredictably. The more integral the AI system, the greater the potential impact of failure. For example, an AI-driven retail store inventory system malfunction could result in too much or too little stock of some products, affecting customer satisfaction, revenue, and brand damage.

- **Legal and regulatory risks:** Laws and regulations concerning privacy and AI are emerging and changing. Organizations not paying attention to these developments are at risk of violating relevant regulations, some of which could lead to public sanctions and reputation damage. For example, an organization lacking effective data privacy governance can run afoul of privacy laws if its AI-enhanced systems process PII unlawfully. On the other hand, regulatory bodies worldwide are also struggling to keep up with the rapid development and deployment of AI systems.
- **Financial risks:** Organizations employing AI in resource management or decision-making systems may experience poor outcomes. Further, adopting new technologies such as AI can exceed cost estimates, affecting financial results. For instance, an organization using AI-driven securities trading could experience significant losses if the AI system misjudges market conditions and events.
- **Cybersecurity risks:** Organizations are increasingly adopting AI-powered cybersecurity tools to improve their defenses and reduce the likelihood or impact of cyberattacks. Like other types of AI-powered systems, it's critical that the organization understand how the AI-powered system is supposed to work, and do so consistently. For example, an organization purchased an AI-powered network intrusion prevention system, blocking network traffic from a critical business partner, resulting in an hours-long outage and negative publicity. Further, organizations must protect their AI systems from poisoning that could enable them to produce erroneous output.
- **Supply chain risks:** Organizations using vendor-supplied AI systems are at risk of attacks on those vendors, where the attackers insert malicious code into the software that is distributed to its customers. The 2020 SolarWinds attack is a textbook example of this risk.
- **Misalignment risks:** Due to limited AI literacy and a lack of understanding of how these models operate, behavior that appears benign in testing environments can lead to unintended consequences when deployed. For example, an AI tasked with maximizing productivity, measured by the number of items produced per worker, might conclude that reducing the number of workers is the most effective way to achieve that goal.
- **Reputation risks:** If realized, all the risks cited in this section can lead to public disclosure, resulting in public backlash and brand damage. For example, an AI-driven targeted ad system could select "tone-deaf" ads, leading to public outcry and reputation damage.

The risks and harms to organizations in this section can be considered from the organization's perspective, which includes the same risks to individuals and groups described in the earlier sections. For example, an organization using an AI-powered employment application screening tool can cause harm to the individuals unfairly treated, as well as to the organization itself, if the harm is publicized.

Risks and Harms to Society

The impact of AI on society is the most profound, as the risks and harms brought by AI can reshape social, economic, and political structures. These effects can transcend the scope of individual organizations and groups, affecting entire nations, regions, and global systems.

- **Job displacement:** The use of automation in organizations can lead to the displacement of workers in industries, including manufacturing, transportation, information technology, and customer service. Indeed, as I am writing this chapter, I see news articles portending decreased demand for software engineers, transportation, customer service representatives, and other jobs. Further, some organizations are no longer hiring entry-level personnel in some industries, leading to new college grads struggling to find employment. For example, the adoption of autonomous vehicles reduces the demand for drivers.
- **Economic inequality:** Biases in training data could result in unequal treatment of persons, resulting in or exacerbating economic inequality. For example, employment and loan applicants could be treated unequally.
- **Erosion of human autonomy and agency:** Society could become overly dependent on automated, AI-powered decision-making, further eroding human autonomy, when decisions about individuals, groups, and society could be made by algorithms rather than people. For instance, governments could increasingly rely on AI algorithms to make public policy decisions, rather than engaging with the electorate in traditional ways.
- **Ethical and moral challenges:** As AI systems become more powerful and are relied upon to make decisions, ethical dilemmas will result, such as when systems are tasked with making life-or-death decisions. These challenges become particularly pronounced when AI decisions conflict with societal values or norms. For example, an autonomous taxi with passengers could be faced with an unavoidable accident in which its passengers or nearby pedestrians may be injured or killed. How should the autonomous vehicle respond?
- **Cybersecurity and privacy:** The potential scope of cybersecurity and privacy issues is bounded only by the scope of the systems themselves. In the event of an AI-related malfunction, large-scale public utilities, control systems, and surveillance systems could inflict considerable harm, including bodily injury and death. For instance, a large electric power grid using AI decision-making to manage electrical loads could misinterpret incoming signals, resulting in widespread power failures.
- **Manipulation of public opinion:** Today's social media platforms use AI-driven algorithms to control information feeds to increase engagement and ad revenue. The potential societal and political power may be too compelling to resist. For example, contrary to the law, a national government could manipulate a popular social media platform to favor certain election candidates.

This section reviewed the potential impact of AI at the individual, group, organization, and societal levels. The risks discussed here are rarely isolated to just one level but will spill over and impact many more areas. For example, a large employer utilizing an AI-driven employment screening function to identify suitable candidates for interviews may have improperly trained its selection algorithm, resulting in harm to specific applicants, people groups, the organization itself, and potentially society (at least in the region(s) where the organization operates).

Understand me: I am not against AI. On the contrary, I am an optimist and believe AI presents incredible opportunities for individuals, groups, organizations, and society. But anywhere you find great opportunities, you also find significant risks. I believe that this cosmic balance is unavoidable.

For readers who are (or will be) responsible for AI governance, I hope this section has provided practical, high-level information to help you better understand the range and types of risks associated with AI. Later chapters in this book explore these and other risks in greater detail.

We've Seen This Mania Before

It is said that there is nothing new under the sun. When it comes to AI, with all the hype, mania, and promises to solve every problem, this is not the first time that a new technology has made otherwise rational people a little crazy. Stepping back in time, we had the advent of the personal computer in the 1980s, the Internet in the 1990s, the cloud in the 2000s, and now AI. Each case involved organizations rushing to implement these technologies, with little regard for rational thought and decision-making. Organizations can get drunk on the prospect of first-mover advantage. The dot-com crash of the early 2000s is evidence that organizations rushed to "go online" without first creating sound business plans and defining specific problems to be solved. With AI, it's the same but also different—it is embedded in everything today, and bolting on AI without carefully understanding the risks is fraught with more danger than in previous iterations of new, shiny technology.

Characteristics of AI Requiring Governance

AI systems have evolved rapidly, becoming integrated into every level of humanity, from everyday individual lives to organizations and national governments. As AI continues to transform governments and industries, its complexity and the potential consequences of its deployment necessitate the creation of governance structures to manage them and ensure favorable outcomes. Unlike traditional information systems, AI introduces new challenges

that require tailored governance strategies to ensure ethical use, minimize risks, and optimize benefits. With great power comes great responsibility. *Governance* provides the guardrails.

This section explores the unique characteristics of AI that necessitate a comprehensive approach to governance. These characteristics include:

- Complexity
- Opacity
- Autonomy
- Speed and scale
- Potential for harm and misuse
- Data dependency
- Privacy
- Intellectual property
- Probabilistic vs. deterministic outputs

Each of these is discussed in more detail in this section.

Complexity

AI systems are often incredibly complex regarding their underlying models and the complex environments in which they operate. This complexity requires governance frameworks that can address their technical and operational intricacies. Complexity arises from advanced techniques such as neural networks and deep learning, where the interrelationships between inputs and outputs are not easily understood.

For example, the healthcare industry uses AI-powered diagnostic tools with deep learning algorithms trained on vast amounts of medical data. These systems can identify patterns that humans might miss, and their complexity makes it difficult for healthcare providers to understand exactly *why* the system made a particular diagnosis. This raises concerns about explainability, interpretability, accountability, and transparency.

A governance structure must ensure that AI systems are explainable, that management is accountable for all outcomes, and that systems are auditable. This includes developing transparency standards and documentation practices, ensuring that AI systems can be monitored effectively to assess and mitigate risks.

Opacity

Some AI systems are, by their very nature, "black boxes," particularly those using machine learning and deep learning models. Essentially, it can be challenging to understand how inputs are transformed into outputs, making the decision-making process of the AI system opaque and mysterious. This opacity presents significant governance challenges, especially in high-stakes fields like finance, healthcare, and criminal justice, where decisions based on

AI predictions can have serious consequences. The opposite of opacity is *explainability*, in which humans can understand and explain how an AI system made a particular decision or created a specific output. Explainability is similar to *interpretability*, which is an AI system's ability to understand the relationships between patterns in data.

For example, a bank's AI-powered loan approval system might reject an application without clearly explaining the factors influencing its decision. If the rejection is based on biased data, the opacity of the AI system makes it difficult for an applicant to contest the decision or understand why it was made. Likewise, the bank may have difficulty explaining precisely why the loan was denied, potentially putting the bank in regulatory jeopardy.

A governance structure must ensure that certain AI systems incorporate explainability and *transparency* in their design and training. This includes requiring that AI systems provide clear, understandable reasons for decisions, especially in areas like finance and healthcare, where transparency can protect individuals' rights, prevent discrimination, and protect organizations against legal action.

Autonomy

Some AI systems are *autonomous*, meaning they make decisions or take actions without human initiation or intervention, increasing the complexity of managing AI systems, as it can be challenging to predict how an *AI agent* will behave in every situation. Autonomous systems may be able to evolve and adapt over time, based on new data, adding challenges to governance.

For example, an autonomous vehicle may make a split-second decision to avoid a collision by swerving, potentially putting pedestrians at risk. The vehicle's AI-driven decision-making process is autonomous and cannot be easily overridden or influenced by human operators (if present) during critical moments.

A governance structure should identify appropriate limits on the actions of AI systems, ensure safety measures are in place, and regularly monitor performance. The governance function must enact clear policies to define what AI systems can and cannot do and address and establish *accountability* for decisions made by autonomous systems.

Speed and Scale

AI systems process data and make decisions much faster than humans, enabling them to operate at scale across a wide range of applications. The speed at which AI systems operate is both a strength and a risk. AI's ability to process vast amounts of data at high speeds makes it possible to solve problems impossible for humans to address manually. However, the speed and scale of AI can also lead to unintended consequences (particularly electricity and computing costs) if systems are not properly governed.

For example, a social media platform's AI algorithm automatically recommends content to millions of users, optimizing for engagement and, ultimately, advertising revenue. However, this can also spread harmful misinformation before it is detected and addressed.

Governance structure must be prepared for the speed and scale and the resource impacts of AI decision-making, necessitating real-time monitoring, response protocols, and appropriate controls to prevent and mitigate large-scale risks such as misinformation, financial fraud, and security breaches.

The Global Impact of AI

According to *Scientific American,* data centers (including those running AI systems) accounted for about 1.5% of global electricity usage, and this is projected to double by 2030. At that time, AI will be consuming as much electricity as the entire country of Japan. In the United States, AI-driven data centers are expected to account for fully half of the increase in electricity usage by 2030.

Potential for Harm and Misuse

As discussed earlier in this chapter, AI systems have the potential to be used for harmful purposes and cause unintended negative consequences. Whether through malicious actors or unforeseen behavior, AI's potential for harm requires strong governance structures to detect and prevent abuse and manage risk. AI systems can be tested in the design phase to detect the potential for harm.

For example, AI-driven surveillance systems could be used by authoritarian governments to monitor citizens' activities, suppress dissent, and infringe upon privacy rights (this is not theoretical but is already the case in some countries). Similarly, AI-powered weapons in military applications could malfunction or be misused, leading to ethical and safety concerns.

Governance structures must establish safeguards to detect and prevent harmful uses of AI, including mechanisms for *ethical oversight*, *accountability*, and *security controls*. In some jurisdictions, such as the EU, regulatory frameworks define the permissible uses of AI and ensure that AI is not deployed in ways that violate human rights or present risks to public safety.

Data Dependency

Most types of AI systems are heavily dependent on data, sometimes in vast amounts. The quality, volume, and diversity of AI training data directly affect the performance and reliability of these systems. Poor or biased data can lead to inaccurate or unfair outcomes.

For example, AI systems used in criminal justice, such as sentencing tools, may be trained on biased historical data, leading to discriminatory outcomes. In another example, if historical arrest data disproportionately represents certain demographic groups, the AI system may unfairly flag individuals from those groups as having a higher risk of reoffending.

Governance structures must address matters of *data quality* and *security*. Organizations should establish standards for collecting, storing, and processing data to ensure it is accurate, complete, and protected.

Privacy

Many AI systems depend heavily on data about employees, customers, and other constituents. Often, these datasets contain one or more elements of PII.

For example, organizations collect PII from customers and buy more from data brokers. They may train their AI systems with this data for various purposes, including sales forecasts, targeted marketing, product and service preferences, and demographic analysis. Some organizations may violate their own *privacy policy, notice of privacy practices (NOPP)*, and/or applicable privacy laws if they don't first examine them to determine what's permitted.

We still lack a reliable method for untraining an AI once it has ingested data. This introduces a serious challenge: after the training phase is complete, complying with data protection regulations becomes almost impossible. It underscores a critical point: Data must be anonymized and safeguarded before it is ever introduced into an AI model.

Governance structures must address matters of bias, fairness, privacy, legality, and protection of PII. It is essential to establish data collection, storage, and processing standards to avoid bias, ensure compliance with privacy regulations (for example, GDPR), include notices of privacy practices (NOPPs), and protect sensitive information.

Intellectual Property

Some public AI systems, including LLMs like ChatGPT, Copilot, Grok, and Claude, have been known to train their AI models with vast amounts of information, including copyrighted content such as newspaper articles, books, images, and art. As of mid-2025, there are no established legal precedents in the United States to determine whether training AI systems with copyrighted content is ethical or legal.

For example, Getty Images filed a lawsuit against Stability AI in 2023, alleging that Stability AI used millions of Getty Images photographs, along with captions and metadata, to train its AI image-generation tools without authorization.

Existing copyright frameworks also fail to protect artist's personal styles, techniques, and methods. By design, today's AI learns the patterns in the data it is trained on, which means if an AI is trained on a specific artist's work, it may adopt their style and not necessarily replicate their exact work. Thus, copyright law leaves artists woefully unprotected.

Governance structures must employ measures to ensure that all data used for AI systems is vetted for copyright and other intellectual property issues. Similarly, AI governance also needs to ensure that an organization's workforce is acutely aware of the business rules regarding the use of public AI models and whether any organization data may be uploaded to them.

Probabilistic vs. Deterministic Outputs

Some AI systems, especially neural networks and those based on machine learning, often produce *probabilistic* outputs, meaning their predictions or decisions are based on probabilities rather than certainties. In contrast, traditional *deterministic* systems' outputs are entirely predictable given the same inputs. AI's probabilistic nature introduces an element of uncertainty, which can be challenging when outcomes have significant consequences and must be explained.

For example, based on probabilities, a healthcare AI diagnostic tool predicts a patient's likelihood of developing any particular condition. The prediction is not definitive, and relying on it could lead to unnecessary treatments or missed diagnoses if the probabilities are misinterpreted or the AI system's accuracy is not high enough.

Governance structures must address the *uncertainty* inherent in probabilistic AI systems. Management may need to define acceptable thresholds for AI decision-making, provide strict guidelines for interpreting probabilities, and ensure transparency in decision-making. Like *risk appetite* and *risk tolerance*, organizations should also determine the level of certainty required in high-risk contexts, as defined by their AI classification model.

Principles of Responsible AI

As AI becomes more integrated into everyday life, society, and critical business sectors such as energy, healthcare, finance, and criminal justice, the principles of *responsible AI* have emerged as core principles for ethical and effective uses of AI. Responsible AI ensures that AI systems are designed, developed, and deployed to respect human rights, promote fairness, and minimize risks while benefiting individuals, organizations, and society.

The remainder of this section identifies and explains the key principles of responsible AI, including ethics, fairness, safety and reliability, privacy, security, transparency and explainability, accountability, and human-centricity. These principles are the core elements that should responsibly govern AI's acquisition, development, and use. Each principle is illustrated with examples to demonstrate its practical application.

Ethics

Ethics should be considered the primary foundation of responsible AI. Ethical considerations must guide the development, deployment, and use of AI systems to align with core human values. AI systems must be designed to operate within ethical boundaries that reflect societal norms and sound moral principles.

Challenges with Ethics

AI is disrupting ethical and social norms. AI algorithms can make decisions that conflict with moral values, such as prioritizing profit over human well-being or disregarding human rights.

Like statistics, ethics can be twisted and turned to suit any outcome. Organizations must ensure that their ethical standards rest on solid foundations and are not driven by non-human-centric agendas.

For example, military organizations employ AI-powered drones that make decisions on the battlefield, with ethical implications that raise serious concerns about accountability and human dignity.

Applications of Ethics

Ethical AI requires sound, comprehensive guidelines to ensure that AI systems promote human well-being and not harm or exploit the vulnerable. AI systems should align with sound ethical principles, including human rights, dignity, and social good. Organizations should commission ethical audits, performed by objective, external parties without a stake in their outcomes. Impact assessments and stakeholder engagement are essential to identify and address ethical risks throughout the AI development lifecycle.

Questions for Further Consideration

Can an AI system that is developed without ethical considerations be considered responsible? Why or why not?

Who, inside or outside the organization, should establish ethical norms? How can they be determined to be sound and defendable?

How do ethical norms vary among societies? Can an AI system with a worldwide audience satisfy ethical norms in all locations?

Fairness

Humans seem to have a built-in instinct for fairness, as demonstrated by toddlers who are quick to point out differences in how they are treated with the exclamation, "That's not fair!" Fairness is a core component of our social fabric and is essential to responsible AI.

Fairness is a core principle of responsible AI that compels us to design and use AI systems that do not discriminate against individuals or groups based on characteristics such as race, gender, religious and political affiliation, ethnicity, socioeconomic status, and disability. When achieved, AI systems minimize or eliminate *bias* and promote equitable outcomes for all.

Challenges with Fairness

AI systems can unintentionally perpetuate or even amplify existing biases, particularly if their training data reflects historical inequities. Intentional or not, biases can lead to discriminatory outcomes, particularly when AI is used in decision-making processes, including hiring, lending, or criminal justice.

For example, a large retail organization employed an AI-powered system for screening new employment applications. After a short time, management discovered that the screening system rejected female applicants far more than male applicants. An investigation into the matter found that the training data used in the system consisted of the resumes and profiles of existing employees, who were predominantly and disproportionately male.

Not all societies and cultures have the same norms regarding fairness. Multinational organizations may need to understand potential differences and design their AI processes and systems accordingly.

Applications of Fairness

To ensure that AI systems exhibit fairness, organizations must ensure that the data used to train AI systems is diverse and representative, regularly audit AI systems for bias, and apply techniques such as *bias mitigation* to prevent further discrimination. AI systems should ensure *equitable access* and opportunities for all persons, regardless of their demographic characteristics.

Safety and Reliability

Safety and *reliability* are critical principles that ensure AI systems are safe for individuals, groups, organizations, and society. AI systems should function as intended, even when facing unexpected inputs in unpredictable environments. AI systems must also be robust to handle situations associated with errors or malfunctions.

Challenges with Safety

AI systems, especially autonomous systems with lives at stake, can pose significant risks if they malfunction or make bad decisions. The stakes are high in safety-critical areas like healthcare, emergency response, and autonomous transportation.

In autonomous vehicles, AI makes real-time navigation and driving decisions. A malfunction or failure in the AI system, such as the inability to identify a pedestrian in poor lighting conditions, could lead to accidents, injuries, and even fatalities.

Applications of Safety and Reliability

AI systems related to life safety should be developed with strict and comprehensive requirements. Those AI systems should be subject to rigorous testing and validation under various conditions. Further, such AI systems should be designed with *fail-safes,*

interruptability, and *fault tolerance*. Organizations should employ *continuous monitoring* of AI systems in life safety contexts to identify and address emerging risks.

Privacy

Many AI systems process large amounts of sensitive personal data and PII. To establish and maintain trust, organizations must take extra precautions so that all uses of AI comply with applicable privacy regulations. To ensure the *privacy* of PII, organizations must carefully track all uses of PII and provide means for responding to inquiries, corrections, and removal requests as required.

Challenges with Privacy

Even before AI, organizations were challenged to constrain themselves from using employee and customer data for uses beyond what was stipulated in privacy policies and applicable regulations. Some organizations behave as though there are no rules regarding AI systems and the data used to train them. Without strong data governance and AI governance functions, misbehavior may continue unabated until these practices can be reined in.

For example, an organization's mobile health app that uses AI to track fitness goals may inadvertently expose users' sensitive health data if its design is not analyzed for compliance with policies and regulations. Such violations could lead to public sanctions and brand damage.

Applications of Privacy

AI systems that use PII must be designed to comply with all applicable privacy laws. Privacy techniques, including *data minimization* (that is, collecting only the data necessary for the task at hand) and *deidentification* (various techniques to scrub datasets), should be employed. Security measures, discussed next, should be a part of an AI system's design.

Security

Many AI systems process large amounts of sensitive data, including personal information, intellectual property, and copyrighted content. Protecting this data from unauthorized access and compromise is a cornerstone of trust and is essential for applicable privacy laws (which require both the proper handling and use of personal data and its protection from theft).

Protecting personal information, intellectual property, and other content is foundational through long-established cybersecurity practices. This book summarizes these techniques, which are explored in great detail in other books published by the author.

For example, a retail organization is pushing to implement AI to improve the efficiency of its customer service function. Like many of its peers, the organization already struggles to understand where all its critical data lies and how it is used. This problem is only made worse when some of this data is also used to train its AI systems, which adds to the proliferation of sensitive data it struggles to protect.

Challenges with Security

Many organizations struggle to protect sensitive information and critical systems, as cybersecurity is often prioritized lower than profitability, time-to-market, and other business survival imperatives. Organizations lacking effective IT and cybersecurity governance will likewise struggle with AI if these foundational capabilities are substandard. Yet, organizations face tremendous pressure to innovate and drive efficiency using AI, as their peers are likely already doing. This is exacerbated by the fact that modern AI is subverting traditional security and trust boundaries.

Applications of Security

Organizations should first inventory their information systems and sensitive datasets, then perform risk assessments to identify risks of potential compromise and loss. Then, with management support, additional safeguards can be applied or improved to bring cybersecurity to a level of due care. This process is a standard, although brief, blueprint for modern cybersecurity.

Transparency and Explainability

The term *transparency* broadly refers to the ability of an organization to understand how its AI systems function. More specifically, *interpretability* is an organization's ability to understand how it reaches its conclusions. *Explainability* is the characteristic of an AI system that allows an organization to provide understandable justifications for the actions the system makes. These critical principles help build trust and ensure accountability, particularly for AI systems making impactful decisions.

Challenges with Transparency and Explainability

Many AI models, including neural networks and those employing deep learning, operate as "black boxes," making it difficult for an organization to understand its outputs and decisions. This opacity leads to trust issues and challenges to AI systems' decisions.

For example, a community bank has recently implemented an AI tool to examine loan applications. When the AI system rejects a loan application, the organization may not provide a concise reason for the rejection. The applicant is dissatisfied with the lack of a credible explanation and may complain to regulators.

Applications of Transparency and Explainability

To avoid situations in which an organization cannot adequately explain the reasoning behind an AI system's decisions and outputs, organizations must prioritize the development of *explainable AI (XAI)*, where AI models provide understandable reasons for their outputs. While this may appear to hamper an organization wishing to provide the most powerful AI capabilities, using XAI in decision-making contexts can help organizations avoid situations that can lead to financial, social, operational, and reputational damage.

Accountability

Accountability is that organizational trait that ensures that individuals and organizations are held responsible for the outcomes of their AI systems. The roles and responsibilities for AI system oversight and recourse are essential when problems arise.

Challenges with Accountability

AI systems often operate autonomously and make decisions independently, but this does not absolve anyone of responsibility when something goes wrong.

For example, a healthcare organization's AI-enhanced diagnostic system makes an incorrect diagnosis, leading to patient harm. The healthcare organization tried to blame the diagnostic system and its manufacturer, but the outcome in arbitration results in accountability resting solely with the healthcare provider, because it selected and configured the device, and it failed to conduct additional follow-up to confirm the diagnosis.

Applications of Accountability

Organizational *policy*, culture, and AI governance frameworks must establish and enforce clear accountability, including assigning responsibility for every aspect of AI system design, deployment, and monitoring. In some cases, organizations should establish mechanisms for *recourse* (including dispute resolution) when harm is alleged. Organizations also need to be transparent about the workings of their AI systems and provide individuals with avenues to challenge AI system decisions.

It's easy to pin the blame on AI developers and hold them accountable for their AI systems' actions. However, if developers are solely accountable for the costs of security, their products may become less attractive, which leads organizations to opt for cheaper, less secure alternatives. That's why accountability should ultimately lie with the deployers. When organizations are held responsible for the behavior of the AI systems they deploy and use, security quickly becomes a much higher priority.

The Culture of Accountability

Organizations are all over the place on the topic of accountability. In other words, are staff members, managers, leaders, and executives accountable for their actions and decisions? Are they required to face consequences when things don't go as intended? Organizations that struggle are more likely to blame others (vendors, service providers, and AI systems) when problems occur. If your AI system makes a decision that gets your organization in hot water, don't blame the AI system; after all, you selected it and are accountable for all outcomes.

Human-centricity

The principle of *human-centricity* states that AI systems should be designed with human well-being and dignity in mind. AI should augment human capabilities, improve people's lives, and preserve their dignity, rather than replace or diminish human roles. AI systems should be designed and implemented in alignment with human values and priorities.

Challenges with Human-centricity

As organizations strive to improve efficiency in all forms, they may cut too close to the bone and prioritize financial efficiency over human values. Organizations employing AI systems focusing on profit margins might forsake the well-being of their workers or customers.

For example, a factory has implemented AI to improve production efficiency, which results in worker displacement. The organization may improve its financial position but at the cost of losing its community-centric reputation.

Applications of Human-centricity

Organizations must emphasize human-centered design principles, ensuring that AI systems are developed to serve human interests. This may require *impact assessments* to determine how AI systems enhance, rather than displace, human capabilities. Processes in which AI systems make decisions should include a *human in the loop*, particularly in financial services, healthcare, and justice.

Summary

Artificial intelligence simulates human intelligence with capabilities like learning, reasoning, problem-solving, perception, and language understanding.

AI systems are classified in various ways, including their scope (narrow/weak AI, general AI, artificial general intelligence—AGI, and artificial superintelligence—ASI). AGI and ASI models have not yet been developed and announced to the general public.

AI functions include reactive machines, limited memory AI, theory of mind AI, self-aware AI, and robotics. Each type of function is suited to a range of tasks.

Some of the techniques related to AI systems include machine learning, deep learning, natural language processing, and expert systems. AI system training techniques include supervised, unsupervised, semi-supervised, and reinforcement learning.

AI systems pose certain risks to individuals, groups, organizations, and societies without proper safeguards. The risks to individuals include privacy violations, bias and discrimination, cyberattacks, and loss of autonomy and control. The risks to groups include the amplification of social inequality, erosion of social cohesion, and cyberattacks. Risks to organizations include operational disruption, legal and regulatory issues, financial risk, including cost overruns and legal costs, cyberattacks, and reputation damage.

The risks to society include job displacement, economic inequality, erosion of human agency, cyberattacks, and manipulation of public opinion.

Organizations often embrace and experiment with emerging technologies to gain first-mover advantage. Historic examples include the emergence of the personal computer (PC), the Internet, and cloud computing. There were winners and losers in every generation, partly because of good planning and partially due to good luck.

For most organizations, AI systems are the most complex and impactful systems they'll ever use. This complexity and power alone should convince any organization to put governance structures in place to ensure intended outcomes, as the consequences of mishaps can be substantial.

The characteristics of AI systems, particularly complexity and opacity, challenge organizations that employ AI systems in decision-making.

AI systems can deliver speed and efficiency but may warrant a human in the loop in high-risk, consequential roles, such as medical diagnosis, employment applicant analysis, and loan application analysis.

Autonomous AI systems, such as autonomous vehicles, need safeguards to minimize the risk of impactful, unavoidable circumstances.

Organizations whose AI systems process PII must be sure to comply with all applicable privacy laws, as well as organizations' own notices of privacy practices. Further, organizations must protect the data used by AI systems to resist cyberattacks.

Responsible AI is a cornerstone principle of AI. Great power calls for great responsibility, but AI systems also bring the risk of highly consequential errors and the lure of misuse. This and several characteristics of AI call for establishing effective governance structures to ensure only intended outcomes.

Ethics is a core tenet of responsible AI, but organizations must approach this with caution, as ethics can be subjective and twisted at will.

Fairness is a key part of responsible AI, as it's vital that AI systems in decision-making roles always produce outcomes that are seen as fair to all concerned.

The safety and reliability of AI systems is yet another facet of responsible AI, particularly in contexts like autonomous vehicles and robotic surgery.

Privacy and security are a part of responsible AI, where organizations are duty-bound to comply with privacy regulations, principles, and policies, and to protect all forms of data in an AI system from cyberattacks and other forms of abuse.

Responsible AI also includes transparency and explainability when AI systems make decisions that impact others, such as employment and loan applications. Organizations must understand (and explain to affected parties) how an AI system produces its outputs. Some types of AI systems are more explainable than others; organizations must choose wisely.

Responsible AI includes accountability—organizations cannot blame AI systems or their makers when things don't go as planned. Organizations need to have a culture of accountability, so that decision-makers take their roles seriously, since they will be bound by their consequences.

Finally, responsible AI includes human-centricity. AI can vastly improve the efficiency of its business processes, but when it comes to interaction with employees and customers,

it's unwise to make them "feel like a number." Further, AI systems in decision-making roles should involve a human in the loop to ensure fairness to affected parties.

Exam Essentials

Be able to explain the various types of AI systems. AI systems are classified and categorized in various ways, based on their capabilities, functions, and how they are trained and used. The types of AI systems include weak/narrow AI, general AI, and artificial superintelligence (not all exist today). AI functions include reactive machines, limited-memory AI, theory of mind AI, and robotics.

Know the characteristics of AI that necessitate governance. While AI systems can improve organizational efficiencies and quality, they are so potentially powerful that the consequences of unintended outcomes can destroy an organization. These characteristics include complexity, opacity, autonomy, speed and scale, and potential misuse and harm.

Know the principles of responsible AI. While AI systems can improve organizational efficiencies and quality, they are so potentially powerful that the consequences of unintended outcomes can harm citizens, groups, organizations, and society as a whole. Thus, AI systems must be responsibly developed. The principles of responsible AI include ethics, fairness, privacy, security, transparency, explainability, and accountability.

Review Questions

The following questions are designed to test your understanding of this chapter's material. For more information on how to obtain additional questions, see this book's Introduction. You can find the answers in Appendix A.

1. The Netflix Recommendation Engine and Google Search Autocomplete are examples of:
   - **A.** AI agent
   - **B.** General AI
   - **C.** Narrow AI
   - **D.** Training techniques

2. An organization has developed an AI system that receives and analyzes input from sensors and takes action as needed. This system is an example of:
   - **A.** AI agent
   - **B.** Robot
   - **C.** General AI
   - **D.** Operational technology

3. An EU-based retail organization retains data about individual transactions and maintains databases with detailed customer information. The organization wants to build an AI system to analyze its customers and transactions, to help better understand sales history, and to develop better sales and marketing strategies. What interaction, if any, is required with customers before launching this project?
   - **A.** Customers can opt out of this activity.
   - **B.** Customers must be notified of this activity.
   - **C.** No customer interaction is required.
   - **D.** Customers must explicitly consent to this activity.

4. An IT department is building an AI system and will begin training it with labeled data. What kind of training is this?
   - **A.** Supervised learning
   - **B.** Unsupervised learning
   - **C.** Semi-supervised learning
   - **D.** Reinforcement learning

5. A regional bank is building an AI system to help analyze loan applications and make approval and interest rate recommendations. What kind of harm must the bank ensure is not realized?
   - **A.** Financial
   - **B.** Social cohesion
   - **C.** Privacy
   - **D.** Bias

6. A regional bank is building an AI system to help analyze loan applications and make approval and interest rate recommendations. What safeguard should the bank implement to ensure fairness?
   - **A.** Event detection
   - **B.** Supervised learning
   - **C.** Human in the loop
   - **D.** Hallucination detection

7. A social media platform uses an AI algorithm to select items to appear in users' data feeds. What potential harm could result if the AI algorithm is biased?
   - **A.** Job displacement
   - **B.** Manipulation of public opinion
   - **C.** Erosion of human autonomy
   - **D.** Privacy

8. An organization has implemented a neural network AI system to review job applications and make interview recommendations. An employment candidate recently applied for a position, was rejected by the AI system, and requests an explanation for the rejection. What potential challenge is the organization about to face?
   - **A.** Security incident
   - **B.** Privacy violation
   - **C.** Difficulty in providing a reason for the rejection
   - **D.** Economic inequality

9. The highest priority in the design of the AI system controlling an autonomous vehicle should be:
   - **A.** Privacy
   - **B.** Safety
   - **C.** Fairness
   - **D.** Security

10. The highest priority in the design of an AI system visually identifying returning customers in a retail store is:
    - **A.** Security
    - **B.** Fairness
    - **C.** Safety
    - **D.** Privacy



Chapter 2

Organizational Readiness

THIS CHAPTER COVERS AIGP DOMAIN I, "UNDERSTANDING THE FOUNDATIONS OF AI GOVERNANCE," SUBDOMAIN I.B, "ESTABLISH AND COMMUNICATE ORGANIZATIONAL EXPECTATIONS FOR AI GOVERNANCE." THE FOLLOWING TOPICS ARE COVERED:

- ✔ AI governance roles and responsibilities
- ✔ AI governance training
- ✔ Various approaches to AI governance
- ✔ Descriptions of AI developers, deployers, and users

THE AI GOVERNANCE PROFESSIONAL (AIGP) PERFORMANCE INDICATORS IN THIS SUBDOMAIN INCLUDE:

- ✔ Define roles and responsibilities for AI governance stakeholders.
- ✔ Establish cross-functional collaboration in the AI governance program (for example, for efficacy and diversity of expertise and perspective).
- ✔ Create and deliver a training and awareness program to all stakeholders on AI terminology, strategy and governance.
- ✔ Differentiate approaches to AI governance based upon company size, maturity, industry, products and services, objectives and risk tolerance.
- ✔ Identify differences among AI developers, deployers and users from a governance perspective (for example, with respect to responsibilities, opportunities and needs).



AI governance cannot succeed without clear expectations, defined roles, and a shared understanding across the organization. This chapter examines the fundamental aspects of operationalizing AI governance, including assigning responsibilities, promoting cross-functional collaboration, enhancing stakeholder awareness, tailoring governance to fit the organization's context, and addressing the distinct needs of developers, deployers, and users. Whether your organization is just beginning to implement AI or managing dozens of models in production, these principles will help ensure governance is effective, scalable, and aligned with both risk and opportunity. The goal is to make responsible AI not just a policy but a consistent practice.

Roles and Responsibilities for AI Governance Stakeholders

AI systems introduce unprecedented complexity, risk, and societal impact to the world. As organizations move from experimentation to operationalizing AI at scale, strong governance becomes a prerequisite, not just a competitive advantage or compliance tool, but a moral and strategic imperative. Central to that governance is the clear definition of roles and responsibilities. Without clarity on "who owns what," organizations face blurred accountability, regulatory pressures, and ineffective oversight.

An effective AI governance function is quickly being considered a part of overall corporate due care. An organization facing legal issues related to its AI systems will face greater consequences if it lacks an AI governance program.

This section offers a practical and detailed guide to assigning roles and responsibilities to key stakeholders in AI governance. It breaks down organizational functions, offers lifecycle-based responsibility mapping, and includes real-world examples and implementation strategies. The goal is to empower organizations to develop governance models that are both robust and adaptable.

Without clearly defined roles, AI governance fails before it begins. However, when responsibilities are formalized, understood, and enforced:

- Risks are better managed.
- Ethical commitments are upheld.
- Innovation proceeds with confidence and integrity.

This section demonstrates that AI governance is a team effort, requiring commitment from stakeholders across technical, legal, operational, and executive levels. By aligning the right people with the right responsibilities across the AI lifecycle, organizations can build systems that are not only smart but also safe, fair, ethical, and accountable.

The Strategic Importance of Defined Governance Roles

Roles and responsibilities serve as the backbone of AI governance. When expectations are documented and understood:

- Accountability can be assigned and enforced.
- Collaboration between teams becomes more efficient.
- Gaps and overlaps in risk ownership are reduced.
- AI systems are better aligned with business, legal, and ethical standards.

AI governance should not be centralized in a single team. Like cybersecurity or data privacy, it is a shared responsibility spanning disciplines:

- **Business, data, and product owners** must align AI with corporate objectives, user needs, and market impact.
- **Technical teams** must ensure resilience, robustness, and transparency.
- **Legal and compliance experts** must navigate regulations and enforce risk controls.
- **Executives** must champion and fund governance priorities.

In short, governance is a distributed function that succeeds only when its roles are delineated.

Stakeholder Categories and Their Governance Roles

AI governance roles are present throughout the organization. While titles may vary, responsibilities tend to cluster around common functions and tasks. This section provides an overview of the primary categories, accompanied by expanded details.

Executive Leadership

Typical roles include:

- Chief Executive Officer (CEO)
- Chief Operating Officer (COO)
- Chief Risk Officer (CRO)

- Chief Data or AI Officer (CDAO)
- Chief Compliance Officer (CCO)
- Chief Information Security Officer (CISO)
- Chief Privacy Officer (CPO)

Governance responsibilities of these roles include:

- Set the tone from the top for responsible AI development and deployment.
- Ensure the organization's strategic use of AI aligns with values and mission.
- Approve risk thresholds and AI applications in high-stakes domains (for example, healthcare, criminal justice).
- Allocate funding and resources for governance tools, training, and staffing.
- Hold teams accountable for compliance with governance protocols.

For example, a CEO who publicly endorses an internal "AI Ethics Charter" and makes adherence to it part of annual leadership KPIs reinforces that governance is not optional but embedded.

AI Governance Board or Council

A typical AI governance board, council, or steering committee would include cross-functional leaders from data science, cybersecurity, privacy, compliance, legal, ethics, risk, and business units.

Governance responsibilities include:

- Develop and maintain the organization's AI governance framework.
- Conduct risk assessments and reviews for sensitive AI use cases.
- Oversee implementation of audit and impact assessments (for example, AIA, DPIA).
- Serve as a final reviewer for algorithmic risk escalations.
- Monitor and report progress on responsible AI goals to the board of directors.

For instance, a multinational retailer creates an "AI Steering Committee" that assesses whether planned AI projects meet pre-established ethical and regulatory criteria before deployment.

Legal, Ethics, and Compliance Teams

Legal counsel governance responsibilities include:

- Identify and interpret existing regulations (for example, GDPR, EU AI Act, CCPA) and assess applicability to the organization and its AI systems.
- Advise on legal liability, intellectual property, and third-party risk.
- Draft contracts and SLAs that include AI-specific requirements.
- Participate in impact assessments and incident response investigations.

Ethics officer or responsible innovation lead governance responsibilities include:

- Translate organizational objectives and values into actionable principles for selecting and designing AI systems.
- Lead ethical impact assessments, especially for systems that affect rights or freedoms.
- Host "ethics labs" or roundtables for contentious use cases.
- Develop a moral narrative that can inform decision-making in complex and ambiguous cases.

Compliance officer governance responsibilities include:

- Embed governance controls into operational workflows and product delivery pipelines.
- Ensure adherence to internal policies and external standards (for example, ISO/IEC 42001).
- Conduct periodic reviews and internal audits on AI systems and processes.
- Train employees on AI compliance risks and best practices for mitigation.

The Role of a Chief AI Ethics Officer (CAIEO)

A growing number of organizations are establishing CAIEO roles to oversee the ethical governance of their AI systems. Their mandate typically includes promoting public accountability, implementing transparency strategies, and balancing profit goals and public interest.

Technical and Engineering Teams

Technical and engineering roles include data scientists and engineers, software developers and integrators, and IT engineering and operations. Aside from leaders, personnel on these teams will not shape governance; instead, they will be required to comply with governance-driven policies and procedures.

Data scientist and AI engineer responsibilities include:

- Select, train, validate, and document AI models.
- Proactively identify and address issues of model bias, drift, or unintended behaviors.
- Apply explainability techniques (for example, SHAP, LIME) to support transparency and auditability.
- Document all aspects of the AI system and data design.

Data engineer responsibilities include:

- Ensure data quality, lineage, and security across the AI system lifecycle.
- Implement pipelines that support data minimization and privacy controls.
- Tag and catalog datasets used in model training and evaluation.

AI ops team responsibilities include:

- Operationalize AI deployment through established SDLC/AI lifecycle processes.
- Implement monitoring for data drift, model performance, and runtime anomalies.
- Automate rollback or alert procedures in response to critical thresholds.

As a best practice, consider introducing "model governance checkpoints" into development lifecycles. These gates check for documentation, fairness assessments, and performance benchmarks before allowing deployment.

Cybersecurity and Risk Management Teams

Responsibilities for cybersecurity and risk management teams include:

- Integrate AI-related risks into the *Enterprise Risk Management (ERM)* portfolio.
- Maintain a risk register specifically for AI systems or implement tags in the risk register to isolate and analyze AI-specific risks.
- Work with other organization leaders to identify and evaluate financial, reputational, and legal exposures related to AI.
- Conduct scenario planning and tabletop exercises for AI-related events, incidents, and failures.
- Define mitigation strategies based on likelihood and severity matrices.
- Include all AI systems in vulnerability management, threat intelligence, device hardening, event monitoring, penetration testing, and all other standard processes.

For example, a bank's risk team categorizes facial recognition software as "high-risk" due to the potential for bias and data protection breaches. As a result, it mandates human review and ongoing impact monitoring.

Product and Business Unit Leaders

Responsibilities for product managers and business unit leaders include:

- Sponsor AI use cases that meet business objectives and customer needs.
- Define success metrics in line with organizational values and risk appetite.
- Provide domain-specific input on data relevance, operational impact, and regulatory nuances.
- Participate in model validation to ensure outputs align with real-world expectations.
- Serve as liaisons between technical teams and end users.

For instance, in a logistics company, the routing optimization system is co-owned by a product manager, who ensures the AI aligns with delivery KPIs, and a compliance officer, who ensures fairness in driver assignment.

End Users and Frontline Staff

Responsibilities for end users and staff include:

- Interact with AI systems in operational contexts (for example, content moderators, clinicians, loan officers) and according to corporate policies.
- Monitor for inaccurate, unsafe, or biased outputs.
- Provide feedback for model refinement and retraining.
- Escalate anomalies or concerns through governance reporting channels.

Enable User Empowerment

Consider designing interfaces that allow users to flag AI recommendations, request human override, or obtain explanations of outcomes. These interfaces promote transparency, empower end users, and embed governance into everyday workflows.

External Parties

Several categories of external parties play stakeholder roles in an organization's AI systems.

Vendors' and suppliers' responsibilities include:

- Must meet governance requirements stipulated in procurement and contracting processes.
- Provide model documentation, risk assessments, and performance metrics.
- Support ongoing monitoring and transparency obligations.

Regulators' and auditors' responsibilities include:

- Review organizational compliance with relevant AI regulations and standards.
- Evaluate AI systems for fairness, accountability, and societal impact.

Civil society, affected communities, and the public at large can contribute by:

- Offering valuable perspectives on how AI impacts rights, equity, and inclusion.
- Participating in participatory design workshops, focus groups, or community impact reviews.

Assigning Responsibilities Across the AI Lifecycle

Different roles have varying responsibilities depending on the phase of an AI system's lifecycle. Table 2.1 provides a high-level lifecycle-responsibility map.

TABLE 2.1 AI Responsibilities Across the AI Lifecycle

Phase	Primary Roles	Key Responsibilities
Ideation & Design	Execs, Product Owners, Ethics Leads	Define acceptable use cases, assess ethical risk, allocate funding
Data Acquisition	Data Engineers, Legal, Compliance	Ensure lawful collection, assess data bias, enforce privacy controls
Model Development	Data Scientists, Risk Teams, Ethics Leads	Train and validate models, document assumptions, evaluate bias
Testing & Validation	QA Engineers, Domain Experts, Compliance	Conduct robustness and security tests, validate accuracy and reliability, approve for deployment
Deployment	AI Ops, IT, Business Leads	Monitor model behavior, implement alerts, enforce SLAs
Monitoring	Risk Officers, Users, Compliance	Audit for drift, gather user feedback, update documentation
Decommissioning	Legal, IT, Governance Board	Retire outdated models, preserve data logs, update records

AI governance must be embedded "left to right" across the lifecycle. Don't treat it as an afterthought or appendage. Remember that the role of governance is to ensure specific outcomes, meaning that governance functions must be present in every stage of the AI system lifecycle.

Implementation Tools and Techniques

Each governance role should be documented and managed according to sound document management practices. Standardized templates for an organization's AI governance program should define:

- Mission statement
- Decision rights and veto power
- Escalation procedures
- Performance metrics (for example, model audit completion rates)

Assign governance tasks across stakeholders using *RACI*—Responsible, Accountable, Consulted, and Informed. Table 2.2 depicts a typical RACI matrix.

TABLE 2.2 Example RACI Matrix for Governance and Operational Roles

Task	Dev Lead	Risk Officer	Legal	Product Owner
Model training	R	C	I	C
Fairness evaluation	C	R	C	C
Regulatory disclosure	I	C	R	C
Post-deployment monitoring	R	A	I	C

Cross-functional Collaboration in the AI Governance Program

AI governance does not operate in a silo; instead, it is a cross-functional endeavor that requires coordination across departments, disciplines, and even external stakeholders. The complexity of AI systems, their social implications, and their operational dependencies mean that no single team or expert group can (or should) govern AI effectively in isolation. Collaboration across the organization is a strategic necessity.

This section explores how to establish and maintain cross-functional collaboration as a pillar of your AI governance program. It covers why it matters, who should be involved, how to structure collaboration effectively, and how to sustain it over time. You'll also find examples and tools to help operationalize these principles within your organization.

Cross-functional collaboration isn't a checkbox: it's the connective tissue that ensures AI systems remain aligned with corporate culture, human values, business goals, regulations, and societal expectations.

Why Cross-functional Collaboration Is Critical

Due to their complexity and the new ways in which they consume data, produce data, and impact business outcomes, AI systems influence:

- Legal and regulatory exposure (for example, discrimination, data privacy violations)
- Reputational risk (for example, unethical uses, opaque decision-making)
- Technical robustness (for example, model bias, drift, adversarial attacks)
- Business operations (for example, automation efficiency, scalability)
- Human impact (for example, ethics, fairness, dignity, job displacement)

No single team—whether legal, data science, IT, cybersecurity, risk management, or HR—can fully assess or mitigate all these risks. Cross-functional collaboration ensures that diverse lenses are applied to AI oversight.

When AI governance is centralized under a single corporate function, it often lacks business relevance or technical depth. Shared ownership across the organization ensures:

- Higher compliance with governance processes
- More context-aware decision-making
- Greater organizational buy-in
- A culture that prioritizes responsible innovation
- A lower risk of post-implementation issues and incidents

Why Can't We All Get Along?

Many organizations experience friction between departments, such as IT and cybersecurity, which often seem to be at odds. Organizations that want to succeed with AI must find ways to set aside their differences and join forces to achieve the best possible AI-enabled outcomes.

Principles of Effective Cross-functional Governance

To establish durable and effective collaboration, organizations must embed structural, cultural, and procedural enablers into their relationships and processes. Diversity of expertise is one such structural enabler, by including stakeholders with varied backgrounds, including:

- **Technical roles:** Data scientists, AI engineers, IT architects
- **Legal/compliance roles:** Privacy counsel, cybersecurity, regulatory experts
- **Ethics roles:** Chief compliance officers, responsible AI leads
- **Business leaders:** Product owners, process managers
- **Operational staff:** HR, finance, customer service
- **End-user representatives:** Both internal and external users

Diverse Perspectives Improve Risk Detection

In one case, a language model was flagged for bias by a linguist, rather than a data scientist, who identified sociolinguistic stereotypes during testing. Multidisciplinary review teams widen the net of risk detection.

While collaboration requires flexibility, clear role definitions help prevent gridlock and turf wars. Define:

- Who owns which decisions (for example, final deployment approval)
- Who provides input (for example, ethics, legal, cybersecurity)
- Who is responsible for documentation
- Who monitors post-deployment outcomes

Consider using *RACI* matrices or *decision-rights* frameworks to reduce ambiguity. Keep comprehensive business records, including meeting proceedings and decisions that can be reviewed later.

Organizations should not treat collaboration as episodic (for example, only during risk reviews). Instead, embed it into the AI lifecycle as an iterative function:

- Early design reviews
- Data acquisition planning
- Model validation stages
- Deployment readiness checks
- Incident response simulations

Cross-functional checkpoints and decision gates throughout the AI systems development process create safer, more aligned outcomes.

Building the Collaborative Structure

If you ask ten people what is meant by collaboration, you'll likely get a few different answers. Here are a few ideas that can be considered to get people from across the organization to work together on AI-related issues.

Working Groups and Advisory Boards

A standing, cross-functional AI governance working group can serve as a hub for ongoing coordination and collaboration. Consider membership for:

- Representatives from legal, risk, and compliance
- Technical leads from AI/ML or engineering
- Product/business representatives
- CCO or ethics advisors
- Cybersecurity and privacy representatives
- End-user or community liaisons (if applicable)

Responsibilities:

- Review high-risk use cases
- Coordinate risk assessments (AIA, DPIA, model cards)

- Refine governance policies
- Escalate concerns to executive-level AI boards
- Track accountability and training metrics

For instance, a financial institution forms a "Responsible AI Council" composed of 12 rotating members from various departments. The council meets monthly to assess models in development and annually to suggest updates to the governance charter. Optionally, such a council could be empowered to make changes to the governance charter, rather than merely making suggestions.

Establish Cross-functional Use Case Reviews

Establish review protocols for any AI system that exceeds specific risk thresholds or is classified in higher-risk categories. The goals of these reviews include:

- Assess ethical and legal risk
- Validate stakeholder representation
- Test model performance across demographic groups
- Confirm mitigation plans

Required participants include:

- Data science lead
- Risk/compliance officer
- Business owner
- Ethics reviewer
- End-user advocate

When should a review be triggered? Consider a use case review when:

- Automated decisions affect individuals' rights
- Data includes sensitive personal or biometric data
- The system impacts hiring, lending, or medical treatment
- Model complexity reduces explainability
- Expanded regulations increase regulatory scrutiny

Collaboration Templates and Artifacts

Organizations can standardize their documentation and decision-making records with shared artifacts:

- **Model fact sheets or model cards:** Include purpose, limitations, and evaluation metrics
- **Risk assessment forms:** Structured walkthroughs of risk categories (bias, drift, misuse)
- **Impact assessments:** Templates for *Data Protection Impact Assessments (DPIAs)* and *Algorithmic Impact Assessments (AIAs)*

- **Ethics checklists:** Covering dignity, fairness, explainability, interpretability, and human oversight

Shared documents reduce misunderstandings and serve as a record of collaboration. Just as AI systems should have a high degree of transparency, so too should AI governance.

Fostering a Collaborative Culture

Even with governance structures in place, collaboration can falter without accountability and cultural reinforcement.

Technical and non-technical teams often speak different "languages." Organizations can overcome this with:

- Joint workshops to build shared vocabulary (for example, what do "ethical" and "fairness" mean in context?)
- Shadowing programs between departments
- Rotational governance roles (for example, a legal advisor embedded with the data team for one quarter)

For example, a tech firm hosts an annual "AI Ethics Hackathon" where engineers, designers, lawyers, and social scientists collaborate to develop solutions to known AI challenges and pitfalls.

Governance is often treated as "extra" work. To reward cross-team contributions, shift this attitude by:

- Including governance participation in performance reviews
- Creating recognition programs for collaboration
- Assigning credit for successful cross-functional governance efforts

Governance as a KPI

A company tracks "AI project risk mitigated before deployment" as a *key performance indicator (KPI)* for both technical and legal teams. This tracking creates shared incentive.

Collaboration often fails when certain voices dominate the conversation. Create psychological safety within the whirlwind of power dynamics by:

- Rotating meeting facilitators
- Creating anonymous input channels
- Empowering junior staff to flag ethical concerns
- Requiring diverse panel participation in decision-making

Ethical Foresight Tip

People closest to the technology often see risks early, so be sure to give them opportunities to speak up. Encourage input from those with less institutional power but more valuable insights.

External and Cross-organizational Collaboration

In the context of information technology (which includes AI), organizations are no longer islands; instead, they work closely with vendors and third-party service providers. When using third-party AI systems:

- Require participation in your governance reviews
- Review their model documentation and risk assessments
- Align on shared principles (for example, no use of sensitive attributes for prediction)
- Establish joint incident response protocols

For example, a retail chain utilizes a third-party facial recognition vendor and requires joint testing for demographic bias prior to deployment.

In the context of regulations, organizations need to engage proactively with:

- **Regulators** to align their practices with evolving laws (for example, the EU AI Act)
- **Industry consortia** (for example, IEEE, ISO, NIST, ENISA) to contribute to standard-setting
- **Academic institutions** for critical perspectives and testing methodologies

For instance, an insurance firm partners with a local university ethics center to co-design fairness benchmarks for underwriting models.

In high-impact applications (for example, law enforcement, social services), community collaboration matters. Organizations can collaborate by:

- Conducting focus groups and stakeholder interviews
- Inviting civil society groups to review use cases
- Creating participatory design workshops
- Publishing impact statements and soliciting feedback

Remember, cross-functional doesn't just mean across departments. It means listening to those affected by AI systems, especially when decisions have a significant impact on areas such as housing, policing, or credit.

Sustaining Collaboration Over Time

Initial enthusiasm often fades when the honeymoon ends. Sustained AI governance collaboration requires:

- **Governance charters:** Codify roles, meeting cadence, and escalation paths
- **Training programs:** Jointly train cross-functional teams on AI risks, biases, and legal updates
- **Shared metrics:** Track how collaboration improves risk mitigation, transparency, and compliance
- **Retrospectives:** Hold postmortems on governance failures or near misses to improve

Governance doesn't end when the AI system is deployed. Instead, it's just getting started.

For example, a healthcare AI team hosts quarterly governance retrospectives, including clinical staff, to assess any unintended consequences of deployed models.

Training and Awareness Program on AI Terminology, Strategy, and Governance

Well-designed AI governance frameworks are likely to fail without organizational understanding and buy-in. From boardrooms to server rooms, stakeholders need a shared vocabulary and a common understanding of what AI is, how it aligns with the organization's strategic goals, and how it should be governed to ensure success. This alignment begins with education.

An effective training and awareness program is the linchpin that transforms governance policy from a document into day-to-day organizational behavior. Training and awareness are not add-ons to AI governance—they are the mechanism by which policy becomes practice. By equipping stakeholders with the knowledge, vocabulary, and pathways to act responsibly, organizations reduce risk and increase alignment. It builds fluency, reinforces accountability, and enables teams to identify risks before they cascade into failures.

This section explores how to design and implement a successful AI governance training and awareness initiative. It walks through audience segmentation, curriculum development, delivery strategies, and success metrics.

Leverage What You Already Have

As discussed throughout this book, you'll want to leverage existing processes, resources, and techniques as you build AI governance. Here, you'll want to utilize resources and methods used to deliver AI governance training, such as your *learning management system (LMS)*, recordkeeping, policies, and executive messaging.

Why Training and Awareness Are Foundational

AI isn't a single technology—it's a suite of methods, tools, and philosophies that impact personnel and processes in new ways. Among stakeholders, AI literacy will range from neophyte to expert. For instance, legal teams may not understand overfitting, developers may not grasp regulatory constraints, and executives may not be fluent in governance requirements.

Training bridges these gaps by:

- Defining guardrails and system development structure
- Creating shared language
- Aligning priorities across roles
- Reducing risk through early detection
- Empowering teams to act confidently within governance constraints

Without Awareness and Accountability, Policy = Theater

An organization may have ethical guidelines or risk protocols, but if employees don't understand them—or don't know they exist—they are unlikely to follow them.

The Scope of the Training Program

Not every stakeholder needs to become a data scientist, nor every data scientist a risk manager; however, every role must understand AI sufficiently to make sound, ethical, and compliant decisions. Core objectives should include:

- **AI terminology:** Key concepts (for example, machine learning, neural networks, algorithmic bias, model drift, explainability)
- **AI strategy:** How AI supports the organization's mission, priorities, and risk posture, in the form of principles or succinct examples
- **Governance principles:** Policies, roles, review processes, documentation requirements
- **Risk awareness:** Ethical, legal, operational, and reputational risks associated with AI
- **Escalation channels:** How to raise concerns, flag anomalies, or seek guidance

Training should be tailored to the needs of different stakeholder groups. Consider creating audience "personas" and aligning learning goals accordingly. Table 2.3 provides some examples that you can consider using as a starting point.

Curriculum Structure

This section provides an example curriculum structure that can be tailored to a specific organization's needs. The types of considerations that may impact this structure include organizational size, market sector, regulatory landscape, risk appetite, and corporate culture.

Example Structure and Content Areas

Table 2.4 contains an example structure for various topics in AI governance training. Based on your organization's business sector, regulatory landscape, and other relevant factors, you'll need to tailor this approach to meet your organization's specific needs.

TABLE 2.3 Mapping Audiences to Training Needs and Topics

Audience	Needs	Example Topics
Executives & Board Members	Strategic alignment, legal exposure, enterprise-level risk	AI strategy, regulatory trends, reputational risks
Developers & Data Scientists	Technical rigor, fairness tools, standards	Model documentation, bias mitigation, reproducibility, standard vendors and tools
Legal, Risk, and Compliance	Regulatory fluency, audit readiness	AI-related laws, impact assessments, vendor risks, customer expectations and requirements
Business/ Product Teams	Value alignment, impact awareness	Use case vetting, success criteria, user impact
Frontline Users	Operational understanding	System behavior, human-in-the-loop workflows, recourse options
IT and Cybersecurity	Controls, daily operations, incident management	System and event monitoring, event and incident response, control owner responsibilities
All Staff	Basic literacy, ethical culture	Definitions, examples, escalation paths

Example Module Titles

Here are some suggestions for naming training modules. While it's easy to overthink the names of training modules and courses, it's also important for the names of modules to fit their subject matter. Adding a little intrigue is helpful, too, as it can spark the curiosity of the training audience.

- "AI vs. Traditional Software: Why AI-specific Governance Is Needed"
- "What Every Manager Needs to Know About Algorithmic Bias"
- "Inside a Model Card: Transparency in Practice"
- "Risk Escalation in the AI Lifecycle: When to Raise Your Hand"
- "Governance in Action: A Cross-functional Use Case Review"

Delivering the Training

The methods for delivering AI governance training are as varied as the audience, in terms of the roles and their learning styles.

TABLE 2.4 Mapping Audiences to Training Needs and Topics

Training Topic	Subject Coverage
AI fundamentals (for all)	What are AI, ML, and automation How does AI differ from traditional software Common myths and misconceptions
Strategic alignment (for managers, execs)	Why the organization is investing in AI How AI will enable and further corporate objectives Governance's role in scaling AI safely Business value and reputational risks
Governance framework (for all)	Overview of internal policies and roles What documentation is required How models are reviewed and approved
Ethics and risk (for all)	Principles: fairness, accountability, transparency Examples of ethical pitfalls and mitigations Incident response process
Role-based application (customized tracks)	Case studies relevant to each department Step-by-step guides (for example, "How to conduct a fairness review")

TABLE 2.5 Learning Styles and Their Strengths

Format	Strengths	Examples
Live workshops	Interactive, team building	Cross-functional case simulations
e-Learning modules	Scalable, asynchronous	Onboarding courses, policy refreshers
Job aids and playbooks	Role-specific, accessible	Ethics checklists, escalation flowcharts
Peer learning	Context-rich, social	Lunch-and-learns, community of practice
Microlearning	Time-efficient, engaging	Five-minute videos, quick quizzes

Blended Learning

It's essential to recognize that not everyone learns in the same way; instead, individuals have distinct learning styles. It's wise to utilize a combination of delivery formats to cater to diverse learners and reinforce the material. Table 2.5 includes common learning styles.



Training Tip

People generally learn more through practice. Pair live sessions with hands-on exercises (for example, red teaming a use case, writing a model factsheet) to make abstract principles tangible.

Mandatory vs. Elective Content

Why provide only compulsory training modules when some personnel have a hunger to learn more? You'll want to establish core mandatory modules (for example, for all employees annually) and elective modules for specialized roles. Elective tracks could include:

- "Emerging AI Capabilities" for all
- "AI Compliance and the Law" for legal teams
- "Operationalizing Governance in AI Ops" for engineers
- "Human-centered AI Design Concepts" for product teams

Regional and Cultural Considerations

Multinational and multicultural organizations may need to consider adapting their AI governance training to reflect:

- Local regulations (for example, CCPA in California, PIPEDA in Canada, GDPR in Europe)
- Language preferences
- Cultural interpretations of fairness, bias, privacy, and risk

Maintaining Training Records

Tracking attendance and competency quiz scores is a crucial practice for corporate training across all areas of focus. Organizations with a healthy accountability culture may track attendance and follow up with stragglers who don't attend. Tracking quiz scores provides additional evidence of not only attendance but comprehension, reducing or eliminating the ability for attendees to claim they didn't understand the training.

Some regulations, like the EU AI Act, explicitly require documentation of employee training in AI-related roles. Meanwhile, customers, investors, and the public are increasingly evaluating companies based on their ability to act responsibly with AI.

Building Awareness Beyond Formal Training

Organizations should treat AI governance like a brand—it needs visibility and reinforcement. Ideas include:

- Monthly newsletter on "Responsible AI in Practice"
- Internal blog featuring interviews with AI governance champions

- Governance "tip of the week" in Teams channels
- AI governance trivia contests
- "Spot the bias" challenges
- Escape room-style simulations of governance failures

For instance, a global logistics company publishes a quarterly "Safety Digest" that occasionally includes AI-related content, such as model incidents, regulatory updates, and success stories from internal audits.

Culture Starts at the Top

Executive engagement on the topic of AI governance helps signal that AI governance is essential to the organization's success, rather than mere compliance theater.

Tracking and Measuring Success

Metrics-driven organizations will not stop at tracking completion rates but will also aspire to measure impact. Some useful KPIs include:

- Percent of staff completing training within 90 days
- Pre/post assessments of AI literacy
- Percent of AI projects with completed model documentation of risk escalations initiated by trained staff
- User satisfaction with governance tools and channels

In the spirit of continuous improvement, organizations should establish formal feedback loops. Examples include:

- Learner feedback on clarity and relevance
- Suggestions for new module topics
- Departmental uptake and concerns

For instance, an organization notices repeated confusion around the concept of "explainability." In response, the AI governance team creates a workshop on model interpretation, providing examples of how the organization addresses this topic.

Common Pitfalls and How to Avoid Them

Wise planners understand that it's essential to learn from others' mistakes. Organizations that don't want to have to start over with AI governance training would prefer to do things right the first time. Table 2.6 includes some examples.

TABLE 2.6 Avoiding Training Pitfalls

Pitfall	Mitigation
One-size-fits-all content	Segment training by role and context
Overly technical modules	Use plain language and visuals
Lack of executive participation	Mandate C-level attendance in key sessions
No follow-through	Integrate training outcomes into performance goals and audit processes
Training as a one-time event	Create ongoing learning tracks and refresher cycles

Sustaining the Program Over Time

AI governance training is not a once-and-done event at the start of an organization's work with AI. Instead, organizations should develop a living AI governance knowledge environment that includes:

- A central portal with tools, templates, and policies
- On-demand libraries of past workshops
- Internal forums for discussion and Q&A

Training content needs to be periodically refreshed, particularly when regulatory changes occur, when emerging AI capabilities become available, and after AI-related incidents occur inside the organization and elsewhere.

Organizations that are serious about staff and leadership development can consider linking AI governance training to professional development plans, performance metrics, and career progression ladders.

For example, one organization's HR policy includes "AI Governance Fluency" as a core competency for promotion to team leadership.

Tailoring AI Governance to Organizational Context

AI governance is not one-size-fits-all. If that were the case, we wouldn't need this book or the AIGP certification—we would just grab a framework and make it our own. Just as no two organizations build, use, or depend on AI in the same way, their approaches to governing AI must reflect their unique context. A multinational bank deploying predictive

risk models will need a fundamentally different governance program than a tech startup optimizing logistics.

This section outlines how to differentiate AI governance strategies based on six critical factors:

- Company size
- Organizational maturity
- Industry
- Products and services
- Strategic objectives
- Risk tolerance

Whether you're helping a small company get started or evaluating the posture of a global enterprise, understanding these dimensions will guide the design of appropriate, scalable, and sustainable AI governance.

AI governance must be context-aware. A startup shouldn't adopt the same controls as a bank; a logistics firm doesn't need the same processes as a government agency. To be effective, governance must be *proportionate, relevant*, and *adaptive*.

This section illustrates how governance strategies vary according to company size, maturity, industry, product risk, objectives, and risk appetite. Understanding and adjusting to these factors ensures that governance frameworks are not only compliant but also practical, achievable, sustainable, and business-aligned.

Company Size: Scaling Governance to Organizational Footprint

Smaller organizations operate far differently from larger ones. The processes required by larger organizations, which involve the actions of dozens, hundreds, or thousands of employees, would be onerous for smaller organizations. Similarly, the simplicity and elegance of small-organization processes would be chaotic if adopted by large organizations.

This section breaks down governance into three segments: small organizations with up to 200 employees, mid-sized organizations with 200–5,000 employees, and enterprises with 5,000 or more employees.

The contents of this section serve as guidelines. Like other aspects of AI governance, it's essential to implement policies, processes, and records in a manner consistent with existing practices.

Small Companies and Startups (1–200 Employees)

Common Traits:

- Limited legal and compliance resources
- Few, if any, formal governance structures
- Flat organizational structures

- Informal processes and high agility
- High employee empowerment
- High innovation-to-governance ratio

Governance Recommendations:

- **Start with core principles:** A concise AI ethics charter can provide guidance before a formal policy is established.
- **Appoint a governance champion:** Even without a full committee, designate one or two cross-functional leads.
- **Use lightweight tools:** Spreadsheets, shared checklists, and manual risk reviews are practical starting points.
- **Outsource when needed:** Legal reviews, audits, or red teaming may be contracted to external firms.

An example 30-person edtech startup embeds ethical risk questions into its product design meetings, guided by a single-page, responsible AI checklist.

Mid-sized Organizations (200–5,000 Employees)

Common Traits:

- Some formalized governance roles and processes
- Documented policies
- Early-stage internal compliance functions
- Active expansion of AI/ML across functions

Governance Recommendations:

- **Establish an AI governance council:** Include representatives from IT, legal, product, and data science.
- **Define risk-based review thresholds:** Require impact assessments for models with customer-facing or high-impact decisions.
- **Begin lifecycle integration:** Introduce documentation checkpoints during development and deployment.

For example, a 2,000-person retail chain requires an "AI Use Proposal" template for any new AI initiative, which is reviewed monthly by a steering group.

Large Enterprises (5,000+ Employees)

Common Traits:

- Extensive AI applications across multiple departments
- Separate IT department for each division or business unit

- Existing formal governance structures
- Complex regulatory and reputational risk profiles
- Defined employee empowerment
- Dedicated legal, risk, and data governance teams

Governance Recommendations:

- **Formalize multi-layer governance:** Use a federated model with centralized policy and distributed accountability.
- **Integrate governance with ERM, cybersecurity, and IT governance:** Align AI oversight with cybersecurity, procurement, and audit processes.
- **Invest in tooling and automation:** Deploy model monitoring platforms, audit trails, and access controls.
- **Report to executive leadership or the board:** Governance should have visibility at the highest levels of leadership.

For example, a global bank runs an internal "AI Regulatory Compliance Office" reporting to the CRO and reviews all models above a predefined risk score.

Organizational Maturity and AI Governance

Like organizational size, the maturity level of an organization will determine the level and rigor of governance formality that is likely to succeed, balancing the organization's ability to manage itself with its ability to identify and manage risk.

Larger organizations have maturity on their side but can be hindered by excessive bureaucracy. Smaller organizations have an advantage in terms of agility, but they may struggle to identify and manage risks effectively. The preceding statements assume that larger organizations have higher maturity than smaller ones, although this is not always the case. Some larger organizations still maintain a culture resembling that of startups, for instance. Like the other topics in this chapter, astute insight is required to develop the right level and rigor of governance that has the greatest potential for effectiveness, leading to success with AI projects.

The focus of existing maturity levels discussed here would most closely correspond with the maturity of an organization's *systems development lifecycle* (SDLC) or its IT or technology governance. These two governance models will not significantly differ from that of AI governance.

Maturity Models

Before delving into the details, let's first examine a typical maturity model. A *maturity model* is a scale of maturity levels, each accompanied by a brief description. For example, the maturity model for *COBIT* has six levels:

- **Level 0: Nonexistent:** There is a complete lack of recognizable processes.
- **Level 1: Initial/Ad Hoc:** Processes are unstructured and unpredictable. Success depends on individual effort, not repeatable practices.

- **Level 2: Repeatable:** Similar procedures may exist, but they are informal, not documented, and vary by person or situation.
- **Level 3: Defined Processes:** Processes are standardized, documented, and communicated. People follow a consistent, defined approach.
- **Level 4: Managed and Measurable:** Processes are monitored and measured. Management can identify deviations and take corrective action.
- **Level 5: Optimized:** Processes are continuously improved based on performance data and business goals. Automation and best practices are leveraged.

In the context of AI governance, an organization would want to determine the minimum level of maturity for processes managed through AI governance. An organization with existing processes can be assessed to determine its maturity level, in the form of a gap assessment, to determine whether processes exhibit at least the minimum maturity level or whether any should be improved.

Note on Maturity Level

Low maturity levels should not be considered bad, nor should high maturity levels be considered good. Finally, few organizations should aspire to bring their maturity level to level 5. Instead, organizations need to determine an appropriate maturity level, based on industry sector, risk appetite, and other factors. Moreover, not all processes need to be at the same level of maturity: it's perfectly acceptable for one process to be at level 2 while another is at level 4.

Low Maturity

Traits:

- Experimentation with pilot projects
- Little to no internal expertise
- Minimal policy infrastructure
- Lack of a formal development methodology

Governance Strategy:

- Focus on education and risk awareness
- Use external audits or ethics reviews to address internal gaps
- Document decisions even if informal (for future baselining)

Moderate Maturity

Traits:

- Multiple teams deploying models
- Informal standards beginning to take shape

- Increasing integration into business processes
- Development methodology with moderate rigor and maturity

Governance Strategy:

- Standardize terminology and model documentation
- Create centralized repositories for model cards and assessments
- Mandate use case vetting processes for customer-facing systems

High Maturity

Traits:

- Dozens to hundreds of models in production
- AI models with low latency and low downtime
- AI models handling thousands of concurrent requests
- Heavy AI dependence in revenue-generating systems
- Regulatory scrutiny and public exposure
- Effective, mature development methodology

Governance Strategy:

- Embed governance gates into AI ops pipelines
- Conduct regular third-party audits
- Create advanced risk models (for example, fairness metrics, explainability benchmarks)
- Use continuous monitoring and alerting infrastructure

Industry Sector-specific Governance Imperatives

One cannot construct an effective and compliant AI governance structure without first considering the industry context. Nearly all industries have existing regulations that must be considered to help organizations avoid *compliance risk*, which refers to the consequences of failing to comply with applicable laws and regulations.

Regulated Industries

Examples include finance, healthcare, insurance, energy, and telecommunications.

Governance Characteristics:

- High legal and regulatory expectations (for example, Basel III, HIPAA, GDPR)
- Extensive documentation and audit requirements
- Clear lines of accountability

Governance Best Practices:

- Use formal risk tiers for AI applications, such as those outlined in the EU AI Act
- Conduct model validation and stress testing
- Create traceability from data sources to predictions
- Maintain change control logs and versioning

For example, a health insurer implementing AI for claim denials must maintain explainability logs and demonstrate clinical fairness across demographic groups.

Business- and Consumer-facing Technology

Examples include e-commerce, social media, SaaS platforms, business applications, IT hardware, and consumer mobile apps.

Governance Characteristics:

- Rapid innovation cycles
- High public visibility
- Emerging risk from algorithmic influence and trust
- Privacy policy that includes AI uses of PII

Governance Best Practices:

- Prioritize user feedback mechanisms
- Mandate transparency features (for example, why this ad/content?)
- Conduct regular impact assessments on content ranking or personalization systems

Industrial and Manufacturing

Examples include logistics, automotive, chemical, and energy.

Governance Characteristics:

- Operational risks (for example, physical safety, environmental harm)
- High use of predictive maintenance, robotics, and optimization algorithms

Governance Best Practices:

- Partner AI with safety and reliability engineering teams
- Monitor model drift that could lead to equipment failure
- Require fail-safes and human override mechanisms

Public Sector and Nonprofits

Examples include government agencies, NGOs, and universities.

Governance Characteristics:

- Heightened accountability to the public
- Mission-oriented (rather than profit-driven)
- More constraints on procurement and innovation speed

Governance Best Practices:

- Publish open model impact statements
- Involve community stakeholders in design and review
- Document intended and unintended impacts on vulnerable populations

Products and Services: Governance Based on Use Case Impact

Not all AI systems carry equal risk. Regardless of industry sector, company size, or maturity, governance must scale in proportion to its impact.

Internal Use AI

Examples include expense report automation, HR chatbots, revenue forecasting, and marketing campaigns.

- Risk: Low
- Governance approach:
  - Use simplified documentation
  - Minimal review unless personal data is used

Decision Support AI

Examples include customer segmentation, predictive analytics, and customer support chatbots.

- Risk: Moderate
- Governance approach:
  - Validate assumptions and biases
  - Disclose AI involvement in dashboards or reports
  - Monitor accuracy over time

Automated Decision-making

Examples include job applicant screening, loan applications, medical diagnosis assistant, and tenant applications.

- Risk: High
- Governance approach:
  - Full lifecycle oversight
  - Bias audits and fairness testing
  - Explainability and appeals processes

For example, a telecom company uses an AI model to assess customer creditworthiness. Due to the potential for disparate impact, the model must undergo a legal and ethical review before launch and a probationary period of extensive human oversight.

Real-time/High-stakes AI

Examples include autonomous driving, AI robot surgeons, fraud prevention, and K-12 instruction.

- Risk: Very high
- Governance approach:
  - Red-team simulations and adversarial testing
  - Real-time model monitoring with auto-shutoff triggers
  - Human-in-the-loop validation

Risk Tolerance: Calibrating Governance Based on Appetite for Uncertainty

An organization's risk tolerance and risk appetite should directly influence its governance intensity. An organization with a high appetite and tolerance for risk is willing to take on more uncertainty than an organization with a low appetite and tolerance for risk.

Low Risk Tolerance

Examples include banking, financial services, healthcare, and energy.

Governance Traits:

- Pre-deployment risk quantification
- Conservative release gates
- Comprehensive documentation and signoffs

For example, a nuclear energy firm with low tolerance for model failure requires dual human approval for AI-guided operational changes.

Medium Risk Tolerance

Examples include large high-tech companies, consumer products, and retail.

Governance Traits:

- Tiered oversight depending on use case
- Partial automation with oversight
- Centralized ethics committee to adjudicate ambiguity

High Risk Tolerance

Examples include startup companies, social media companies, and research and development organizations.

Governance Traits:

- Faster iteration cycles with post-hoc reviews
- Distributed ownership of risk
- Ethical guardrails but fewer hard constraints



High risk tolerance should never mean *no governance*. Even risk-tolerant firms need escalation channels, monitoring, and accountability.

Business Alignment

Any governance program, including AI governance, should exhibit business alignment. *Business alignment* is the concept of developing policies, processes, and procedures that align with an organization's mission, goals, objectives, size, sector, and culture. All the topics discussed so far (company size, maturity, industry sector, context, and risk tolerance) are characteristics of business alignment. Understanding business alignment helps tailor governance scope.

This section considers additional characteristics of organizations. You should consider these characteristics as generalizations but not as the sole means for designing a governance program.

Innovation-driven Organizations

Examples include tech startups, software companies, and alternative energy companies.

Traits:

- Fast product cycles
- Emphasis on experimentation
- Relatively higher risk appetite

Governance Approach:

- Keep processes agile and feedback-driven
- Use pilot phases to test governance protocols
- Emphasize ethical foresight over strict compliance

Reputation-focused Organizations

Examples include healthcare providers, financial services, luxury goods, and hospitality brands.

Traits:

- High exposure to public perception
- Operating in trust-sensitive domains

Governance Approach:

- Make public-facing ethics commitments
- Use transparency dashboards
- Involve PR/legal in review boards

Cost-reduction or Efficiency-oriented Organizations

Examples include fast-food chains, airlines, big-box stores, and back-office services.

Traits:

- Emphasis on automation and operational AI
- Less focus on user-facing features

Governance Approach:

- Evaluate tradeoffs between accuracy and fairness
- Focus on accountability in automated decisions
- Document error tolerance thresholds and mitigations

Developers, Deployers, and Users in AI Governance

As AI is integrated into core business processes, products, and services, governance must evolve to account for the full community of stakeholders. Within any organization, three distinct groups often interact with AI systems: *developers, deployers*, and *users*. These roles overlap in some areas but differ significantly in terms of responsibilities, risks, and influence over AI behavior and outcomes.

Understanding the governance implications of each group is crucial for allocating accountability, designing effective safeguards, and ensuring that AI systems operate reliably, fairly, and ethically throughout their lifecycle. This section examines these differences in depth and provides a structured framework for allocating governance responsibilities across these three domains.

Definitions and Role Boundaries

AI Developers

An *AI developer* is an individual or team that designs, builds, trains, and evaluates AI models or systems. AI developers include:

- Data scientists
- Machine learning (ML) engineers
- Data engineers
- Research scientists
- Data privacy analysts
- Cybersecurity analysts

Focus Areas:

- Data curation and preprocessing
- Model selection and training
- Performance evaluation and tuning
- Documentation and testing

AI Deployers

An *AI deployer* is a stakeholder responsible for integrating AI models into production environments and operational workflows. AI deployers include:

- Software engineers
- AI ops professionals
- Product managers
- Infrastructure teams

Focus Areas:

- Model deployment and monitoring
- Infrastructure and scalability
- Integration into applications or services
- Post-deployment observability

AI Users

An *AI user* is an individual who interacts with AI systems during their operational use. AI users include:

- Internal users (for example, HR staff using a resume screening tool, or a developer directing an AI LLM to generate or examine source code)

- External users (for example, customers receiving AI-generated recommendations)
- Decision-makers relying on AI insights (for example, loan officers, clinicians, landlords)

Focus Areas:

- Interpretation of AI outputs
- Operational reliance on AI
- Reporting errors, inconsistencies, incidents, or harms
- Exercising rights to recourse or appeal

Responsibilities Across the AI Lifecycle

AI governance must account for the different touchpoints each role has in the AI lifecycle. In other words, the three roles described previously have specific responsibilities in the phases of the AI governance lifecycle.

Table 2.7 depicts the responsibilities of a typical AI developer. For example, a data scientist developing a credit scoring model is responsible for running fairness diagnostics and documenting known limitations, such as performance degradation on ethnic minority populations.

Table 2.8 depicts the responsibilities of a typical AI deployer. For example, an AI ops engineer is responsible for implementing automated alerts when an AI model's predictions deviate from expected accuracy and ensuring these alerts are escalated appropriately.

Table 2.9 depicts the responsibilities of a typical AI user. For instance, a recruiter using an AI resume filter must understand how to override the system when it excludes qualified candidates based on narrow criteria.

TABLE 2.7 AI Developer Responsibilities

Phase	Developer Tasks	Governance Responsibilities
Data collection	Curate and clean datasets	Check for representational bias, ensure consent/privacy
Model design	Select algorithms, tune hyperparameters	Document assumptions, ensure explainability
Evaluation	Measure accuracy, fairness, robustness	Perform bias testing, validate against benchmarks, assess security
Documentation	Record the development process	Create model cards, data sheets, risk disclosures

TABLE 2.8 AI Deployer Responsibilities

Phase	Deployer Tasks	Governance Responsibilities
Integration	Embed models into apps or tools	Ensure UI/UX supports explainability and user control
Scaling	Enable access and load management	Apply access controls and logging
Monitoring	Track real-time performance	Detect drift, alert on anomalies or failures
Maintenance	Manage model updates	Version control, rollback procedures, impact validation

TABLE 2.9 AI User Responsibilities

Phase	User Tasks	Governance Responsibilities
Interaction	Use an AI system in operations	Understand AI's role in decision-making, avoid blind trust
Feedback	Identify problems or inconsistencies	Report anomalies through proper channels
Oversight	Validate AI-assisted decisions	Exercise discretion, request human review if needed
Rights	Appeal decisions, seek transparency	Invoke recourse mechanisms when impacted

Opportunities and Leverage Points

Each AI governance role presents unique enforcement opportunities, if leveraged strategically.

Developers: Embed Governance into Design

Developers have control over key decisions that shape AI system behavior. Governance opportunities include:

- **Bias mitigation at the source:** Developers control training data and feature engineering
- **Transparency tooling:** Building explainability into the model (for example, using SHAP, LIME)

- **Ethical foresight:** Considering downstream consequences during experimentation
- **Reproducibility:** Using consistent pipelines, versioning, and documentation
- **Security and privacy:** Assessing data flows and designing solutions per privacy policies, security policies, and requirements

Encourage developers to use ethical design checklists and contribute to model fact sheets as a standard step in the development cycle.

Deployers: Operationalize Governance at Scale

Deployers translate developer intentions into real-world impact. Their governance contributions include:

- **Embedding monitoring infrastructure:** To detect performance and fairness drift.
- **Automating governance checkpoints:** To ensure models cannot be deployed without documentation or risk review.
- **Implementing rollback protocols:** To address emergent harms or system failures.

For example, a deployer configures a staging environment that automatically checks for missing model documentation before allowing promotion to production.

Users: Detect and Report Issues from the Front Lines

Users experience AI systems in context, which allows them to:

- **Spot unintended consequences:** For example, confusing recommendations, unfair treatment, or degraded accuracy.
- **Validate outputs with domain knowledge:** For example, clinical expertise to detect diagnostic errors.
- **Initiate recourse or appeals:** For example, alert governance teams when harms are suspected.

Empowering frontline users with training and reporting mechanisms enables them to become effective risk sentinels. AI governance training, discussed earlier in this chapter, tells them what they need to know, what to look for, and how to report issues.

Resource Needs and Expectations

AI developers, deployers, and users require training, tooling, and support that cater to each group's unique needs. Details follow.

AI Developer

Remember, this is an individual or team that designs, builds, trains, and evaluates AI models or systems.

Needs:

- Tools for bias detection, explainability, and robustness testing
- Governance guidelines tailored to AI development workflows
- Feedback from deployers and users to improve model design

Expectations:

- Clear success criteria that go beyond accuracy
- Time allocated for documentation and model audits
- Safe space to flag ethical concerns

AI Deployer

Remember, this is a stakeholder responsible for integrating AI models into production environments and operational workflows.

Needs:

- Access to validated, well-documented models
- Monitoring platforms that surface relevant metrics (drift, latency, usage)
- Guidance on integrating human oversight and user interfaces

Expectations:

- Shared responsibility with developers for post-deployment risk
- Resources to manage model updates and changes responsibly
- Participation in incident response plans

AI User

Remember, this is an individual who interacts with AI systems during their operational use.

Needs:

- Clear understanding of when and how AI is used
- Training on how to interpret outputs and exercise discretion
- Simple reporting channels for errors or harms

Expectations:

- Right to explanation, especially for impactful decisions
- Mechanisms for appeal, override, or human review
- Protection from retaliation when raising concerns

Governance Conflicts and Misalignments

Without alignment, the three AI governance roles may come into conflict.

Developers vs. Users

- **Tension:** Developers optimize for technical performance; users care about practical usability and clarity.
- **Governance solution:** Incorporate user feedback into development cycles; mandate usability testing.

Developers vs. Deployers

- **Tension:** Developers expect a smooth handoff; deployers often deal with incomplete documentation or opaque code.
- **Governance solution:** Establish shared AI ops pipelines with required metadata and risk summaries.

Users vs. Deployers

- **Tension:** Users may want more control and convenience than the interface allows, while deployers may restrict access for security or scalability reasons.
- **Governance solution:** Involve users in UI/UX testing and include options for human review and escalation.

Other conflicts may arise in the context of AI system development and operations, such as conflicts between IT and cybersecurity, or between IT and privacy. These and other conflicts in the context of systems development are not addressed in this book.

Role-based Governance Controls

Tailoring governance controls to each role ensures greater development and operational effectiveness.

Controls for Developers

- **Model cards:** Summarize model purpose, performance, and limitations.
- **Fairness testing protocols:** Assess disparate impact during training.
- **Peer code review for risk:** Governance-focused reviews beyond functionality.
- **Gate reviews:** Formal reviews require management approval before proceeding to subsequent design, development, and testing steps.
- **Ethics approval for novel use cases:** Pre-deployment review of ethical implications.

Controls for Deployers

- **Deployment checklists:** Include documentation, ownership, and escalation paths.
- **Access control logs:** Record who can trigger or modify models.

- **Real-time drift alerts:** Notify when model behavior changes significantly.
- **Gate reviews:** Formal reviews require management approval before proceeding to subsequent deployment steps.
- **Audit trails:** Preserve system behavior history for investigations.

Controls for Users

- **Transparency disclosures:** Show when and how AI is involved in decisions.
- **Override mechanisms:** Let users revert or challenge AI outputs.
- **Feedback channels:** Allow flagging of issues directly from the interface.
- **Recourse procedures:** Provide a formal appeal or human review process.

Role Coordination and Integration

Structuring cross-role communication encourages dialogue across roles. For example:

- **Joint retrospectives** after AI system launches
- **Cross-functional ethics committees** for high-risk use cases
- **Shared dashboards** showing model performance, risk metrics, and feedback loops

For instance, after launching a recommendation engine, the product team conducts a "Governance Postmortem" with data scientists, customer service agents, and UX designers to identify usability issues and potential ethical risks.

Organizations may want to consider integrated AI governance tools that support and involve all three roles. Examples include:

- Model registry with audit metadata
- Alert systems that route incidents to the right role
- Role-based access to governance documentation

Summary

Establishing and communicating organizational expectations for AI governance requires more than policy documents or executive declarations—it demands structured, role-based action and sustained cross-functional engagement. Organizations can move from abstract principles to concrete governance execution by defining responsibilities, fostering collaboration, identifying and educating stakeholders, adapting governance to context, and differentiating the roles that interact with AI systems.

The foundation of any governance program lies in clearly defined roles and responsibilities. Organizations must map accountability across executive leadership, AI governance committees, technical teams, legal and compliance departments, cybersecurity and privacy, business units, and end users. By aligning responsibilities with specific phases of the AI lifecycle—from data collection to model retirement—organizations can enhance oversight, mitigate risk, and cultivate ethical foresight. Accountability tools, such as RACI matrices, lifecycle role maps, and governance charters, help ensure clarity and reduce overlap or gaps in responsibility.

Effective AI governance is not the domain of a single function but instead is a team sport. Cross-functional collaboration is essential for capturing diverse expertise and perspectives. Legal teams understand regulatory nuance, engineers grasp system architecture, privacy analysts understand data usage, and ethics officers bring human impact into focus. Creating cross-functional governance committees, use case review boards, and collaborative documentation processes enables shared ownership and decision-making. Beyond formal structure, fostering a collaborative culture through peer learning, leadership modeling, and psychological safety ensures that governance is not just imposed but also internalized.

Building organizational capacity through training and awareness is another vital pillar of successful AI governance. AI literacy must extend beyond technical teams to include business managers, legal professionals, and operational staff. Each group requires tailored education on terminology, governance structures, strategic objectives, and risk awareness. Successful programs blend live training, asynchronous modules, job aids, and real-world case studies. A mature training strategy not only imparts knowledge but reinforces governance as a continuous, performance-linked practice. Campaigns, recognition programs, and gamified challenges can further embed awareness into daily behavior.

AI governance must also be tailored to an organization's context. A 20-person startup with one AI engineer and limited legal capacity cannot be governed in the same way as a multinational financial institution. Governance approaches must scale and flex based on company size, AI maturity, industry, product type, business objectives, culture, and risk tolerance. Highly regulated sectors, such as healthcare or finance, demand stricter controls, while innovation-driven industries may prioritize agile and lightweight oversight. Similarly, organizations using AI for internal efficiency may adopt more limited governance than those deploying AI in customer-facing, high-stakes uses. Contextual governance necessitates adaptive frameworks and tiered oversight mechanisms to strike a balance between risk and responsiveness.

Governance must account for the different roles that interact with AI systems: developers, deployers, and users. Developers control the building blocks of AI (data, algorithms, and evaluation) and thus carry responsibility for embedding transparency, fairness, and documentation. Deployers are responsible for safely integrating AI into systems, maintaining operational performance, and ensuring monitoring and rollback capacity. Often overlooked, users play a critical role in identifying frontline, emergent issues, validating system behavior in real-world settings, and initiating escalation or appeals when harm occurs. Governance must be role-sensitive, with training, tools, and expectations customized for each group.

Exam Essentials

Be able to define the roles of AI developers, deployers, and users. AI governance assigns different responsibilities based on how individuals interact with AI systems. Developers design and train models and are responsible for documentation, fairness testing, and technical transparency. Deployers integrate models into operational environments, manage infrastructure, and monitor performance. Users apply AI outputs in real-world contexts and are often the first to notice failures, bias, or unintended consequences. Understanding the distinctions among these roles is critical for ensuring appropriate safeguards are in place at every phase of the AI lifecycle.

Know why cross-functional collaboration is critical in AI governance. AI governance requires input from various functions, including legal, technical, ethical, operational, and business. Each stakeholder group identifies different risks and can detect failures that others may overlook. Cross-functional collaboration ensures that ethical issues, regulatory compliance, risk, and business value are considered together, not in silos. This collaboration often takes the form of governance committees, risk and compliance reviews, and multi-role project teams. You need to recognize when and how diverse roles should coordinate to govern responsibly.

Understand how governance strategies vary based on company size, risk tolerance, and other factors. Smaller or early-stage companies often rely on lightweight governance such as checklists, part-time ethics advisors, or external audits. Larger organizations—with more models in production, regulatory exposure, and public visibility—require formalized governance structures, dedicated risk teams, and ongoing compliance reporting. Additionally, a company's risk tolerance (for example, conservative vs. experimental) influences how strict or flexible governance controls should be. Be prepared to evaluate how governance approaches scale according to organizational context.

Be prepared to identify role-specific responsibilities across the AI lifecycle. Governance must be aligned with the various phases of an AI system's lifecycle, encompassing development, deployment, and monitoring. Developers are accountable for data integrity, model documentation, and fairness testing. Deployers are responsible for post-deployment observability, rollback plans, and integration controls. Users must be trained to detect problems and understand when to escalate concerns. The exam may present lifecycle scenarios and ask you to identify who is responsible at key checkpoints.

Recognize the components of an effective AI governance training program. AI governance training should be tailored to the responsibilities and literacy level of each stakeholder group. Executives need strategic alignment and risk awareness; developers require tools for bias testing and transparency; users must understand how to interpret or challenge AI outcomes. Strong programs combine live training, e-learning, policy walkthroughs, and case-based exercises. Expect questions that test your understanding of how to build scalable, role-appropriate awareness initiatives.

Review Questions

The following questions are designed to test your understanding of this chapter's material. For more information on how to obtain additional questions, see this book's Introduction. You can find the answers in Appendix A.

1. Which role is primarily responsible for selecting training data and conducting bias testing during model development?
   - **A.** AI deployer
   - **B.** End user
   - **C.** AI developer
   - **D.** Compliance officer
2. What is the most appropriate governance structure for a small startup with limited AI resources?
   - **A.** Fully automated audit pipelines
   - **B.** A board-level AI oversight committee
   - **C.** Lightweight governance with designated role champions
   - **D.** Centralized AI governance office with legal authority
3. Which of the following best illustrates cross-functional collaboration in AI governance?
   - **A.** Engineers writing code based on user feedback
   - **B.** A legal team monitoring AI regulatory updates
   - **C.** A data scientist leading all model deployment tasks
   - **D.** An ethics lead, product manager, and ML engineer co-reviewing an AI use case
4. Why is it important to tailor AI governance training to different stakeholder roles?
   - **A.** Because only technical staff require training
   - **B.** To increase automation in governance processes
   - **C.** Because each role interacts with AI differently and needs context-specific knowledge
   - **D.** To reduce the number of governance roles
5. What distinguishes an AI user from a developer or deployer in governance terms?
   - **A.** Users monitor system drift
   - **B.** Users design model architectures
   - **C.** Users interact with the AI in real-world scenarios and may flag harm
   - **D.** Users control the data pipeline

6. Which governance strategy is most appropriate for a highly regulated industry like healthcare?
   - **A.** Informal team agreements
   - **B.** Lightweight ethics principles
   - **C.** Formal reviews, documentation, and auditability
   - **D.** No oversight beyond the product team

7. In a RACI matrix, what does the "A" stand for?
   - **A.** Analyzed
   - **B.** Accountable
   - **C.** Audited
   - **D.** Authorized

8. Which of the following is a governance responsibility specific to AI deployers?
   - **A.** Performing legal impact assessments
   - **B.** Conducting data annotation
   - **C.** Integrating models into production and setting up monitoring
   - **D.** Filing incident reports with regulators

9. What is a primary governance benefit of cross-functional AI use case reviews?
   - **A.** Increased software development speed
   - **B.** Reduced need for end-user input
   - **C.** Holistic risk evaluation from multiple perspectives
   - **D.** Avoiding documentation of model design

10. Which of the following would most likely appear in a well-designed AI governance training program?
    - **A.** A primer on generative art
    - **B.** A tutorial on coding best practices
    - **C.** A case study showing governance failure and mitigation
    - **D.** A financial literacy quiz



Chapter

3

Updating Policies for AI

THIS CHAPTER COVERS AIGP DOMAIN I, "UNDERSTANDING THE FOUNDATIONS OF AI GOVERNANCE," SUBDOMAIN I.C, "ESTABLISH POLICIES AND PROCEDURES TO APPLY THROUGHOUT THE AI LIFECYCLE." THE FOLLOWING TOPICS ARE COVERED:

- ✔ Updates to policies needed for AI
- ✔ Updates to security and privacy policies
- ✔ Updates to third-party risk management policies

THE AI GOVERNANCE PROFESSIONAL (AIGP) PERFORMANCE INDICATORS IN THIS SUBDOMAIN INCLUDE:

- ✔ Create and implement policies to ensure oversight and accountability across all AI lifecycle stages (for example, use case assessment, risk management, ethics by design, data acquisition and use, model development, training and testing, deployment and monitoring, documentation and reporting and incident management).
- ✔ Evaluate and update existing data privacy and security policies for AI.
- ✔ Create and implement policies to manage third-party risk (for example, procurement, supply chain and human resources).



As organizations increasingly rely on AI to power decisions, automate operations, and drive innovation, the need for robust governance grows more urgent. This chapter focuses on creating and implementing policies that ensure oversight and accountability of AI throughout its entire lifecycle. From defining responsible use cases and managing risks to updating privacy and security frameworks and controlling third-party exposure, this chapter provides a practical foundation for embedding trust into AI systems. Whether you're deploying in-house models or sourcing external tools, effective governance requires more than compliance: it demands intentional design, clear roles, and enforceable policies at every stage. This chapter equips you to build those foundations with confidence.

IT Governance Is Assumed to Exist

This chapter discusses new and updated policies that are often necessary to address the unique risks introduced by AI systems. These policies address AI-specific considerations that are not typically found in traditional business applications. It is assumed that organizations already have an appropriate level of IT governance in place. The methods in this chapter assume that AI governance activities take place *in addition to* IT governance activities, not in place of them.

Oversight in the Age of Autonomous Decision-making

Most organizations' existing policy covers general information systems management concerns, but AI systems introduce new types of situations, such as autonomous decision-making and impacts on data privacy.

This section examines several aspects of organizational policy that may require revision to ensure proper oversight and accountability mechanisms at each phase of the AI lifecycle. While organizations often focus on late-stage controls, such as deployment monitoring or

post-incident reporting, effective governance must be lifecycle-aware. This means introducing enforceable, reviewable policies from the moment a use case is proposed to the moment a model is retired.

This section presents a lifecycle policy framework designed to:

- Anticipate and mitigate risks before harm occurs
- Enable auditability and regulatory compliance
- Foster public trust in AI-enabled systems

As AI capabilities continue to expand, the governance bar must rise in parallel. The future of trustworthy AI depends not only on what we build but also on how we govern it at every step.

The Lifecycle Policy Model

On the topic of AI systems impacting an organization's policies, let's start with the big picture. Any time an organization invests in a new information system, the resources consumed in the process warrant policy guardrails and development lifecycle checkpoints and gates. When the new information system is an AI system or an AI-enabled system, the risks and benefits escalate, further strengthening the need for policy to cover all angles.

Why Lifecycle Oversight Matters

The lifecycles of AI systems are inherently nonlinear and iterative. A flawed assumption made during data acquisition may surface as biased model behavior during deployment. A lack of documentation during development may hinder future audits or regulatory reviews.

As a case in point, in 2020, a major financial services firm faced backlash after its AI-driven credit scoring system offered significantly lower credit limits to women than men with comparable financial profiles. An investigation revealed that gendered assumptions had influenced the feature selection and training data, yet no policy required the team to document these choices or assess their ethical implications.

The lesson is clear: Governance must anticipate risk, not just react to it. The discovery, analysis, and decisions made when identifying AI-related risks should be integrated into "business as usual" procedures.

Core Features of a Lifecycle-oriented Policy

A comprehensive policy framework should:

- Define responsibilities and escalation paths for each lifecycle phase from cradle to grave.
- Be risk-calibrated (more rigorous for high-risk systems). Higher-risk use cases (for example, in healthcare, justice, and employment scoring) should trigger more stringent review, documentation, and oversight processes.
- Enable auditability and versioning of proceedings and decisions.

- Be embedded into technical workflows, not parallel to them.
- Include mechanisms for continuous monitoring and feedback loops, to track post-deployment performance, incident reporting, and model recalibration as real-world data contexts evolve.
- Address data governance explicitly, including data provenance, consent, minimization, retention policies, and compliance with relevant data protection regulations (for example, GDPR, HIPAA).

It's not enough to create policies—they must be operationalized through procedures, gates, templates, checklists, recordkeeping, and oversight roles that teams are trained to use.

Use Case Assessment: Aligning Purpose and Risk

The development of use cases is near the very beginning of AI system development. Use cases are brief descriptions of business contexts, along with the role of an AI system in facilitating or improving a business process in some way.

Should AI Be Used at All?

Organizations need to thoughtfully decide whether an AI-based solution is the right approach to any business or technical problem. In some organizations, IT departments (and others) are so eager to utilize their new tools that their judgment may be impaired, leading them to use their new technology to address every issue, problem, and opportunity they encounter. To quote the character Malcolm in the movie *Jurassic Park*, "… your scientists were so preoccupied with whether or not they *could*, they didn't stop to think if they *should*."

Before a single line of code is written, teams must evaluate whether AI is appropriate for the problem at hand. This assessment should consider:

- The societal context of the use case (for example, criminal justice, healthcare, hospitality).
- Whether the intended outcome could amplify inequities or cause harm.
- The availability and quality of data to support development.
- Whether less complex or more interpretable systems could meet the same goals. AI systems are often resource-intensive and complex. Sometimes, a straightforward automation or rule-based solution is sufficient. Not every problem requires AI.

Some organizations require a "justification memo" for all proposed AI use cases, which is then reviewed by an AI ethics or risk committee. This memo documents purpose, scope, stakeholder impact, and alignment with organizational values. In addition to this, before any AI system development begins, organizations should utilize other governance frameworks, such as an IT Steering Committee, to follow all required steps. This should include a feasibility study and risk assessment, to determine whether the project should be funded and, if so, when it should begin and what resources it will use.

Risk Rating Frameworks

Proposed AI systems should be assigned a classification level or risk rating (such as the classification model found in the EU AI Act). This classification must be done at the very onset of any new AI system development project, as it will drive the level and rigor of design development and review, risk assessments, documentation, and other key aspects. Absent a formal risk classification, use cases should be assigned a risk rating, such as low, medium, or high, based on criteria such as:

- Impact on individuals or protected groups
- Legal or regulatory exposure
- Degree of automation in decision-making

High-risk use cases might require third-party audits, external stakeholder input, or executive sign-off before proceeding.

Risk Management Tailored for AI

An organization's risk management framework likely needs to be adjusted to accommodate the use of AI systems. One key activity is the assignment of risk or classification level to every prospective AI system prior to its design and development. Frameworks, such as the EU AI Act, define risk levels with prescribed activities to be taken during development and ongoing operations.

Separate or Combined IT and AI Governance?

Organizations with effective IT governance in place may opt to add AI-specific steps to their IT governance and systems development lifecycle processes. Combining AI and IT governance into a single process that accommodates both is often more effective than maintaining separate processes.

AI systems introduce novel risks that traditional software risk management does not fully address:

- **Epistemic uncertainty:** An AI system may be confidently wrong due to data gaps or improper training.
- **Automation bias:** Human operators defer to model outputs even when flawed.
- **Emergent behavior:** The unexpected or unplanned outcomes exceeding an AI system's intentions, especially in multi-agent or generative systems.

In the era of AI, data integrity is something we will have to focus on much more than we have in the past. A policy framework must ensure that risk identification happens continuously, not just during planning and development.

For example, a retail AI recommendation system was found to reinforce gender stereotypes by showing household products to women and electronics to men. Though technically performant, it perpetuated cultural biases, revealing a blind spot in risk identification.

Policies should include scheduled and trigger-based risk reviews, such as:

- After major data or model updates
- When monitoring flags drift or degraded performance
- In response to public feedback or legal changes

Reviews should be documented, with clear risk owners assigned. Follow-through should be required to address all identified risks.

Ethics by Design: From Principle to Practice

Data subjects often have a pretty clear idea of right and wrong, including an organization's acquisition and use of data about them. Organizations considering the use of AI systems must develop formal ethical guidelines that become a part of the standard AI system development lifecycle.

Operationalizing AI Ethics

Abstract values, such as fairness and dignity, must be translated into practical safeguards and actionable steps. This includes:

- Ethical impact assessments tied to each model's use case
- Mechanisms to surface "moral crumple zones," where responsibility is unfairly shifted to low-authority individuals (for example, customer service agents explaining AI decisions they can't control)
- Stakeholder engagement (especially affected communities) in design choices

Like a risk analysis or a privacy impact analysis, an ethics analysis of a proposed AI system should be consistent, repeatable, and performed by trained personnel. Such an assessment could consist of a few simple ethics questions, such as:

- Could this system exacerbate social inequities?
- Are vulnerable populations affected disproportionately?
- Can users contest or appeal decisions?

A more formal approach could consist of a lengthier questionnaire or even an online tool for organizations conducting many such assessments.

Building Ethics into Technical Design

AI system designers and developers should be equipped with tools, such as:

- Fairness toolkits (for example, Aequitas, AI Fairness 360)
- Explainability libraries (for example, SHAP, LIME)
- Bias detection dashboards during model validation

Policies should mandate the use of such tools and require outputs to be documented in business records.

Data Acquisition and Use

AI systems consume large quantities of their data during training and operational use. Organizations considering the use of AI systems that will process their data should already have effective data governance in place. Lacking that, the development of data governance policies, procedures, and tooling should be considered a high-priority prerequisite for AI.

Data Lineage and Documentation

Every AI project should include the development and maintenance of a *data provenance* (also known as *data lineage*) document, capturing:

- Source of each dataset
- Licensing or usage terms
- Known limitations (for example, underrepresentation, class imbalance)

For example, an HR AI trained solely on past hiring data may encode historical biases or discrimination. A lineage review would surface this and trigger a policy-mandated mitigation step, like rebalancing or excluding biased features.

Data Lineage Should Already Be in Place

Data lineage should not be required only for AI projects, but for every IT system project, whether in-house, outsourced, or online, such as SaaS.

Consent, Privacy, and Synthetic Data

An organization's data management and/or privacy policies should also:

- Require informed consent where appropriate and required by law
- Address the use of third-party data brokers
- Mandate privacy-preserving techniques (for example, *differential privacy*) in sensitive domains

Organizations using *synthetic data* must also establish policy safeguards to ensure realism without compromising the confidentiality of personal data.

Model Development: Guardrails for Creation

Organizations need to develop a standard approach to the development of AI models and systems, so that most or all such systems have consistencies that make supporting them more

straightforward. This concept is similar to the adoption of standard programming languages to avoid an excessive number of unique languages selected at the whim of developers.

Enforcing Reproducibility and Accountability

AI model development should not be an artisanal process but rather the result of an engineering discipline. In addition to those required for the systems development lifecycle (SDLC), policies should require:

- A short list of approved AI tools, systems, and platforms, similar to SDLC programming language standards
- Version control systems (for example, Subversion, DVC, MLflow) with managed check-in and check-out rules
- Peer-reviewed or manager-approved pull requests for model changes
- Model cards explaining purpose, limitations, and testing results
- Traceability of training data sources and preprocessing pipelines
- Automated testing for performance, bias, and robustness as part of the CI/CD pipeline
- Approved communication interfaces and protocols (for example, Model Context Protocol—MCP) to standardize model interaction with APIs, services, or agents

Some firms enforce a "two-person rule," where no model may advance to testing unless it has been reviewed and signed off by another qualified AI practitioner. But a more formal approach would be a gate process with formal review and approval.

Safe and Interpretable Model Architectures

An organization's AI development policies should steer teams away from unnecessarily opaque models, especially in high-risk use cases. That includes:

- Enforcing the use of interpretable models unless performance tradeoffs are documented and approved.
- Requiring justification and formal approval for complex architectures, such as deep neural networks.

For example, a rule-based system may be more appropriate for eligibility decisions than a black box ensemble model.

Training and Testing: Preparing for the Real World

The success of an AI platform rests firmly on the success of properly planned training and testing. Lacking effective training, an AI system is unlikely to deliver the expected results.

Because this book focuses on AI governance rather than AI engineering, it does not delve into the details of what constitutes effective training and testing, in part because there are too many variables at play. Instead, as a governance professional, you or someone in your

organization needs to construct business rules within AI governance to define requirements for model training and testing. I do, however, provide the categories and principles of training and testing; first, of model training:

- Training data is accurate and complete
- Privacy rights of data subjects are preserved (past, present, and future)

Note that this discussion focuses on AI model and system training, rather than individual training, which is addressed in Chapter 2.

Here are the categories of testing:

- Accuracy
- Fairness
- Bias
- Explainability
- Performance
- Security

Rigorous Validation Standards

Policies for training and testing should be developed to ensure that AI models and systems perform as required and are also fair, reliable, and resilient. Policies should define:

- Performance benchmarks tailored to use case goals
- Subgroup analysis to detect disparate impact, such as gender- or racial-based bias
- Controls for data leakage and overfitting

All testing, including test plans, test results, and any override approvals (for example, management's approval for go-live if a system's performance doesn't meet requirements) should be documented as a part of the system's permanent business record.

For example, a fraud detection model must not only maintain accuracy but also minimize false positives for legitimate users. Policies may require scenario testing with real-world transaction flows, as well as the level of tolerance for false positives.

Red Teaming and Adversarial Testing

Like any new information system (AI or not), a new AI system should undergo security testing to assess its resilience against various types of attacks. The types of tests should include:

- Adversarial input testing
- Prompt injection scenarios
- Simulations of user misuse
- Convention cyber attacks

Policy can designate a cybersecurity "red team" role tasked with stress-testing the system and logging outcomes for governance review.

Test Results Are a Vital Gate

As discussed in this section, test plans and test results should be documented and archived. AI governance should include a checkpoint at the conclusion of testing, so that management can understand test results, ask questions about them, and make a go or no-go decision on whether the AI systems development project may move forward.

Deployment and Monitoring: Guarding the Gate

When we arrive at the point of deployment, AI governance policy should require at least one opportunity for management to review all the results of analysis and testing prior to management providing a formal approval for the AI system to be moved into production.

Pre-deployment Controls

Before deployment, policies should require:

- Ethics and compliance review checkpoint
- Risk assessment and mitigation review
- Complete documentation package submitted
- Monitoring plan approved by a relevant governance body

In a multi-tiered risk model, the level of rigor for these steps may vary depending on the risk level and context.

For example, a healthcare AI system must demonstrate not only model efficacy but patient communication readiness (for example, human-in-the-loop protocols and privacy).

Post-deployment Monitoring

After an AI system has entered production service, AI governance policies must require:

- Real-time performance tracking against baseline
- Alerts for model drift, out-of-distribution data, and fairness metrics
- Human oversight, especially for consequential decisions

Monitoring dashboards should be accessible to non-technical stakeholders, including compliance officers and perhaps regulators. Due to this diverse audience, monitoring should encompass business-related metrics and reporting, rather than merely focusing on technical performance.

Low-maturity Organizations: Beware

Organizations with low process maturity, particularly in systems development and IT service management, will face significantly higher risks of problems with their AI systems. If this is the case in your organization, I encourage you to identify someone who can discuss this issue openly and clearly with senior management.

Documentation and Reporting: Building Institutional Memory

You've seen it mentioned throughout this section, and here it is again. Any technical governance program, including for AI, must include thorough documentation. This documentation helps those working on the AI system later on understand how the system was initially designed, tested, built, and operated. You can reverse-engineer some aspects of a system but not its requirements, test results, and business decisions; only a set of business records can fill that need.

Living Documentation

The management and maintenance of AI systems should be documented so that all parties involved better understand their design, design limitations, operating parameters, performance expectations, and more. AI systems must be auditable. Policies should require:

- Model documentation is updated after each significant change
- Decision logs for ethical and legal review points
- "Explainability snapshots" for deployed models

Also, AI systems must be auditable. Again, this means that the system's purpose, requirements, design, training, testing, and operations must all be documented, just like any other information system.

Internal and External Reporting

Depending on jurisdiction, system risk, and context, policies should also mandate:

- Regular internal reporting to governance committees
- External transparency reports (anonymized as needed)
- Disclosures to users or the public, especially for systems affecting rights

Incident Management: Planning for Failure

Security incident response (IR) is not new—it's been a mainstay of standard practice since the 1990s, and even earlier for some organizations. Like other disciplines in IT, such as cybersecurity and privacy, formal and documented incident response procedures need to be extended to all AI systems in use.

Organizations with a working security incident response program in place will find that improvements to the program will represent relatively small increments of change.

Defining and Escalating Incidents

Not all AI system errors are incidents. AI governance and security incident response policies must clarify:

- What counts as an AI incident (for example, unauthorized use, biased decisions, system hallucination, data leakage, drift)
- Reporting timelines and thresholds
- Communication templates for regulators, customers, or the public

One or more of these policy changes will necessitate enhanced or additional monitoring. After all, it's difficult to declare that an incident is occurring (or has occurred) if there is no detection capability in place. Hence, it may be necessary to utilize or develop a monitoring capability to enable this detection.

For example, if a chatbot begins generating inappropriate responses, policy may require disabling the model, notifying affected users, and triggering a complete root cause analysis.

AI Playbooks Are Needed

In security incident response, the standard practice is to document a broad, overall security incident response plan that includes roles and responsibilities, severity levels, management reporting, and notification to internal and external parties. Standard practices also include the development of numerous "playbooks," which are detailed, step-by-step instructions to follow when specific types of incidents occur.

One or more playbooks will likely have to be written to cover the types of incidents that are unique to AI systems. Like other IR playbooks, new AI-system-related playbooks must be tested, and personnel must be trained.

Root Cause and Remediation

Standard incident response policies and procedures often require some level of post-incident review. The purpose of such a review is to assess how effectively the organization's detection and response actions were executed, and whether any adjustments to policies, system design, or incident response procedures are warranted. Such changes help the organization avoid similar incidents in the future and/or respond to them more effectively.

After an incident:

- A multidisciplinary team conducts a postmortem
- Governance gaps are identified
- Policies and training are updated accordingly

Policies and the organization's culture should emphasize blameless retrospectives that encourage surfacing systemic issues, not scapegoating individuals.

Evaluate and Update Existing Data Privacy and Security Policies for AI

Traditional data management, privacy, and security frameworks were not designed with AI in mind. They focus on controlling access, maintaining confidentiality, and preventing breaches—critical but insufficient in the context of AI systems that learn, generalize, and infer from data. With AI, data is not just a static asset—it is the fuel for decision-making models, often in high-stakes environments.

As such, organizations must evaluate and revise existing data management, privacy, and security policies to reflect AI-specific challenges. This includes addressing AI's ability to infer sensitive information, the risks of *reidentification* in *anonymized* and *pseudonymized* datasets, and the implications of synthetic data and data augmentation. This section offers a structured approach to assessing and updating privacy and security policies to make them AI-ready.

Evaluating and updating existing data privacy and security policies is no longer optional—it is essential for:

- Protecting individual rights in the face of powerful AI systems
- Meeting legal obligations under evolving AI regulation
- Ensuring data subject and stakeholder trust in AI-driven processes

The organizations that act now to future-proof their governance frameworks will be best positioned to innovate responsibly in the age of intelligent systems.

Why AI Disrupts Traditional Data Governance

AI systems use data in ways rarely found before. Hence, organizations implementing AI governance need to review and revise their data governance.

AI Is Data-hungry by Design

AI and ML models, especially LLMs and deep neural networks, require vast and diverse datasets. This hunger for data creates incentives to:

- Aggregate data from disparate sources
- Purchase or scrape data from third parties
- Repurpose data collected for unrelated purposes

Traditional privacy policies often fail to account for these AI-specific behaviors and may lack controls for:

- *Purpose limitation* in AI systems
- Ongoing access to data post-training
- Third-party training data governance
- *Data retention* practices

Inference and Reidentification Risks

AI models can infer sensitive traits, such as race, political views, or sexual orientation, even if such labels are not explicitly present. Moreover, models trained on supposedly anonymized or pseudonymized data have been shown to leak identifying information, especially when exposed to external inputs or prompted in an adversarial manner.

For example, in 2021, researchers demonstrated that GPT-style language models could memorize and regurgitate sensitive data (for example, SSNs, emails) from their training corpus, even if the data had been lightly obfuscated. This weakness challenges the assumptions of traditional privacy controls.

Privacy Policy Gaps and AI-specific Threats

AI is transforming the way that PII is utilized, and the current limitations of AI pose challenges to the "right to be forgotten" and other privacy needs. Nowhere else in IT is AI so impactful as in privacy.

Privacy governance is a prerequisite for the development of AI governance. Without privacy governance, organizations are more likely to run afoul of privacy laws or societal expectations.

Common Policy Shortfalls

Most organizations' existing data privacy and security policies:

- Do not specify how long training data is retained post-model development
- Lack controls for *model inversion* or *membership inference attacks*
- Assume data minimization applies only to storage, not to what models retain
- Do not include data subject rights in model outputs (for example, right to explanation, opt-out, deletion)

As a result, privacy and security risks may propagate silently from data to models to predictions.

AI-specific Privacy Threats

Notable AI-specific threats to privacy include:

- **Model inversion:** Reconstructing input data by probing the model.
- **Membership inference:** Attempts to determine whether a specific record was a part of the AI system's training set.

- **Model poisoning:** Attackers maliciously inject false data into training pipelines.
- **Function creep:** Expanding the use of data beyond stated original purposes through AI reuse.
- **Unintended memorization:** LLMs may retain and reproduce fragments of sensitive training data, even without direct prompts.

These threats demand a new policy posture: one that sees privacy not as a one-time state but as a continuous, systemic consideration across data and models.

Key Policy Areas to Evaluate and Update

This section examines some areas where organizations may need to update their privacy policies.

Purpose Limitation in Dynamic Pipelines

AI development often involves reusing data across experiments, fine-tuning, and *ensemble modeling*. Privacy policies (including, in some instances, public-facing notices of privacy practices—NOPP) must:

- Define permissible reuse conditions
- Require new consent if original use cases change significantly
- Establish logging and audit trails for each dataset's usage across model versions



Policy Tip

Introduce a "use case registry" that tags each dataset with approved purposes and flags violations during pipeline construction.

Enhanced Anonymization and Pseudonymization Standards

Traditional anonymization and pseudonymization techniques may not be effective against AI reidentification attacks. Updated policy requirements should include:

- Quantifiable risk metrics (for example, k-anonymity, l-diversity)
- Differential privacy for sensitive domains
- Mandatory re-anonymization when combining datasets

Strengthening Anonymization

Consider synthetic data generation combined with differential privacy for public datasets and routinely audit model behavior for memorization patterns.

Lifecycle-based Data Retention Controls

Organizations need to amend their privacy policies to distinguish between:

- Raw input data
- Transformed/feature data
- Labeled datasets
- Model artifacts

Each layer requires distinct retention and deletion policies. Considerations should include:

- Whether training data should be deleted after model convergence
- Whether synthetic or transformed versions of data are governed by the same rules
- How "data deletion" requests apply to learned model weights

Consider implementing a "forgetting budget"—a threshold on how much data a model can memorize, triggering retraining when exceeded. Since it is impractical to retrain an AI model when a single data subject's record has been deleted, this threshold approach may be more pragmatic (and be sure to disclose this practice in the NOPP).

Data Subject Rights in AI Contexts

GDPR, CPRA, and other privacy regulations have defined data subject rights that apply to all manual and automated processes, including those with AI. Let's look at some of these issues.

Right to Explanation and Access

Updated privacy policies must clarify how data subjects can:

- Request access to the data used to train models that affect them
- Receive a meaningful explanation of AI-driven decisions

Understand when an AI system is involved in decision-making. This guidance aligns with legal regimes like the GDPR (Articles 13–15, 22) and emerging AI-specific laws (for example, the EU AI Act).

For instance, a user denied a mortgage must be informed if an AI model was used or influenced the decision, the key features that contributed to it, and how they can exercise their rights to challenge or appeal the decision.

Right to Be Forgotten (Data Deletion)

Based on how many AI models function with training data, the right to be forgotten may pose a particular challenge. Privacy policies should define whether:

- Models must be retrained when personal data is deleted
- Deletion applies to downstream systems, fine-tuned derivatives, or API-accessible models
- A process exists for data subjects to verify that deletion has occurred

This is especially important for foundation models trained on large, uncurated web datasets, given the high risk that harmful, unlawful, or inaccurate content may be incorporated into the AI model.

Security Policy Updates for AI Systems

Organizations need to review and update their information security policy to account for the use of AI. AI-powered systems expand the *attack surface* in new ways not present in traditional information systems.

Expanding the Threat Model

In addition to the universe of traditional vulnerabilities, AI systems in particular are vulnerable to:

- **Adversarial attacks** that manipulate inputs to cause harmful outputs
- **Data poisoning** during training
- **Model theft** via API access
- **Prompt injection** and system jailbreaks in LLMs

Traditional security policies often omit these AI-specific threat vectors. Organizations need to revise their security policies to:

- Include AI in threat modeling exercises
- Define monitoring responsibilities for AI-specific logs, telemetry, and behaviors
- Enforce rate-limiting, access control, and model segmentation

Securing the AI Supply Chain

AI systems rely on open-source code, pretrained models, and third-party APIs. Updated security policies must address:

- *Software Bill of Materials (SBOM)* for AI pipelines
- Vetting and licensing review of third-party models
- Continuous vulnerability scanning for AI components

For example, a company using a public LLM must ensure that prompt data cannot be *exfiltrated* and that the LLM itself was sourced from a trusted repository.

Third-party risk management is discussed in more detail later in this chapter.

Practical Steps for Policy Revision

Regarding the revision of security and privacy policies, here is a standard approach. Before beginning this section, keep in mind that the advice included here assumes that the organization already has a data classification policy, together with an inventory of the data

used in its information systems. Organizations lacking this practice will be at a distinct disadvantage in terms of the management and monitoring of data used to train their AI systems.

Conduct a Gap Assessment

Organizations should start with a privacy and security policy audit focused on AI exposure:

- Are data subject rights extended to model outputs?
- Are there controls in place for the reuse and retention of training data?
- Are AI threats included in incident response playbooks?

Build Cross-functional Review Teams

The review of security and privacy policies should not be siloed in IT, cybersecurity, or compliance. Instead, reviews should include:

- Data privacy officers
- Security architects
- AI/ML engineers
- Legal and ethics advisors

These reviews ensure that updates reflect both the technical reality of AI and the regulatory landscape.

Policy Compliance Artifacts

Published security policies are not an indication that those policies are followed or even known to an organization's personnel. Instead, there should be tangible evidence that policies are being followed. Written policies are not enough—they must be supported by business records that substantiate compliance with policies. In the case of technical policies, confirming compliance may involve examining an organization's architecture or the configuration settings of systems and devices.

One sign of higher organizational maturity regarding policies is a worksheet or chart that describes the location or nature of artifacts that demonstrate compliance.

Policies to Manage Third-party Risk

As organizations accelerate their adoption of AI technologies, many components of the AI lifecycle are increasingly outsourced, procured, or delegated to external partners. From pretrained models and cloud infrastructure to crowdsourced annotators and algorithmic APIs, third parties now play a central role in the development and operation of AI systems. As AI capabilities proliferate, governance must follow the data and the model, wherever they go.

This expanded ecosystem creates unique governance challenges. A single third-party failure, such as a data breach, a biased model, or irresponsible vendor practices, can compromise the integrity of an entire vendor system and that of its affected customers. Managing this risk requires robust, AI-specific third-party risk policies, rather than merely extending traditional vendor checklists.

This section describes a comprehensive approach to creating and implementing policies that govern third-party AI risk across procurement, supply chain management, and workforce interactions.

Why AI Multiplies Third-party Risk

Third-party risk management was hard before the broad adoption of AI. The emergence of AI introduces new risks that further complicate third-party risk management practices.

Beyond the Traditional Vendor Model

In a typical enterprise, *third-party risk management (TPRM)* focuses on legal, financial, and cyber exposure: ensuring the vendor is solvent, insured, and compliant. When implemented by vendors and service providers, AI introduces new technical and ethical forms of risk, such as:

- Black box algorithms that lack explainability
- Datasets of unknown origin or labeling quality
- LLM APIs with unpredictable outputs or safety filters

These risks often slip past standard procurement controls because they arise from the model's behavior, not the vendor's documentation.

For example, a healthcare provider integrates a third-party AI-powered diagnostic model into its patient portal. The model performs well in testing but exhibits racial bias in deployment due to the use of unrepresentative training data. The vendor had not disclosed this issue, and the provider had no policy requiring fairness validation.

As a result, enterprises must expand TPRM frameworks to include AI-specific risk assessment criteria, focused on model behavior, data lineage, transparency, and auditability.

Hidden Dependencies and Sub-tier Risks

Modern AI systems often use nested layers of services, libraries, and data sources. A procurement team may contract with one vendor, unaware that:

- The vendor relies on subcontracted data labeling teams overseas.
- The model was trained on a dataset scraped from the Internet or on copyrighted data without authorization.
- Code libraries used in deployment include outdated or vulnerable dependencies.
- One or more of the vendor's own third parties may also use AI-powered capabilities affecting the organization.

Third-party risk must extend beyond direct vendors to the entire AI supply chain. The practical challenge for TPRM managers is this: how many levels down should the TPRM program probe to identify risks with third parties, fourth parties, and beyond?

Core Policy Objectives for Third-party AI Risk Management

An effective governance framework for third-party AI engagement should:

- Identify and categorize third-party touchpoints in the AI lifecycle
- Impose due diligence requirements that include AI-specific risks
- Define contractual obligations around data use, fairness, security, accountability, and recourse
- Establish monitoring and review mechanisms to ensure ongoing compliance

These objectives must be supported by enforceable policies that map to contract terms, as well as operational workflows in procurement, compliance, security, and engineering.

Service Provider Vulnerability Management

Few service providers will permit their customers to perform vulnerability scans or penetration tests on service providers' services. Instead, customers must rely on external audits and other validations of service providers, which are often conducted by qualified cybersecurity firms. Service providers then make the reports of these audits and validations available to customers, usually under non-disclosure terms.

Procurement Policy Controls for AI-enabled Solutions

Organizations that take a more methodical and mature approach to procurement can incorporate AI-specific steps into their vendor intake and purchasing procedures. It's essential to determine whether an organization's vendors and service providers utilize AI and how they employ it, as some applications may alter the organization's risk profile.

AI Risk Flagging at Procurement Intake

Organizations should modify their procurement intake process to flag when:

- A product includes AI, machine learning, LLMs, or automated decision-making
- Data from the organization will be used to train models
- The vendor processes personal, biometric, or sensitive information

Any procurement action flagged as "AI-relevant" should trigger a supplemental review by legal, risk, and technical stakeholders.

Examples of AI-relevant services include resume screening tools, customer service chatbots, and fraud detection engines.

AI-specific Due Diligence

Procurement policies should require vendors to provide:

- Documentation of model training data sources and licensing
- Explanation of model logic or decision criteria (or a justification for lack thereof)
- Results of internal or third-party audits (for example, for bias, security, robustness)
- Evidence of compliance with applicable regulations (for example, CPRA, GDPR, EU AI Act)

Organizations may decide to use standardized AI supplier questionnaires that map responses to risk levels and approval requirements.

Vendor Risk Tiers

Classifying vendors into risk tiers is a standard practice in TPRM, because not all vendors pose equal risk. Policies should classify vendors into risk tiers based on:

- Function criticality (for example, impacts users or financial decisions)
- Data sensitivity (for example, PII, health data)
- Level of automation (for example, fully autonomous vs. human in the loop)
- Regulatory context

Higher-risk tiers should trigger:

- Enhanced review (for example, fairness testing)
- Contractual safeguards (for example, incident response and indemnity clauses)
- Continuous monitoring requirements

Not All Third Parties Disclose Their AI Capabilities

Not all third parties disclose the presence of AI capabilities in their ecosystems, although this is not a black-and-white issue. A vendor might use AI in its marketing and outreach, but not in its product development, while another vendor may permit its developers to use AI for code generation or quality testing. The bottom line—it's not enough to ask *whether* AI is used in a vendor organization; it's vital to understand *how* it's utilized.

Contracting and Legal Safeguards

Organizations must incorporate AI-related clauses and terms into their contract templates to manage risk at the corporate level. Some vendors may use their own AI-related terminology, making it necessary for legal counsel to review a vendor's contract language carefully.

AI-specific Contractual Clauses

Standard vendor contracts are often silent on AI-specific risks. Organizations should update their standard contract templates to include these provisions:

- **Disclosure requirements:** Vendors must disclose uses of AI capabilities in their organizations or by their third parties.
- **Data use limits:** Vendors are prohibited from using customer data for retraining without explicit consent.
- **Fairness and bias:** Contracts must require vendors to test for and remediate disparate impact.
- **Security practices:** Contracts must define requirements for model access control and incident reporting.
- **Continuous monitoring:** A vendor must implement continuous monitoring to detect various types of problems in its AI system.
- **Audit rights:** Contracts must grant the customer the right to audit or inspect high-risk vendor systems.
- **Termination:** Contracts must define AI-related conditions for which the customer organization may want to terminate the agreement, whether for cause or not.

Here is an example clause on model training: *Vendor shall disclose the data sources and preprocessing techniques used in model training and certify that no personal data was obtained without a lawful basis.* Some organizations will use an expanded clause with more stipulations.

Liability and Incident Handling

Security incidents and breaches occasionally occur in vendor and service provider organizations. Indeed, such organizations may be targeted if an attacker wants to attack their customers. Updated policies should also require:

- **Clear definitions of responsibility** in the event of AI errors or harms
- **Time-bound incident notifications** (for example, within 24 hours)
- **Model deactivation rights** if risks exceed acceptable thresholds

Terms like this help ensure that downstream accountability is enforceable, rather than merely aspirational.

AI Supply Chain Governance

Since you are reading this book, you are probably sold on the idea that AI governance is vital to an organization's successful use of AI capabilities. Since most information technology (hardware, software, applications, and tools) is developed by other organizations, AI governance needs to be extended to supply chains and be incorporated into organizations' third-party risk management processes.

Mapping the AI Supply Chain

Organizations must create visibility into their AI supply chains to provide continuous awareness of the presence and state of all AI models and systems. This includes:

- Models built in-house using third-party data or tools
- Open-source libraries used in AI workflows
- Pretrained foundation models fine-tuned for internal use

Consider requiring vendors to submit an *AI system bill of materials (AI-SBOM)* listing all upstream inputs (code, data, platforms, and so forth) used in model development.

Open-source AI Risk Policies

Many AI teams rely on *open-source* models, datasets, and toolkits. Policies regarding must address:

- **Licensing clarity:** Ensure all open-source assets are properly licensed.
- **Security vetting:** Scan packages for known vulnerabilities.
- **Bias review:** Validate data quality and representation in open datasets.

Organizations using open-source resources in critical or strategic systems should consider establishing an open-source governance function to assess and approve high-impact tools.

Open-source Is Less Secure Than We Thought

For years, there was a broad understanding that open-source software was more secure because the source code was available, allowing many individuals to examine it for defects. However, a 2022 paper titled "Tragedy of the Digital Commons" by Chinmayi Sharma refuted this long-held belief.

Other Third-party AI Issues

There are a few more use cases in which outsourcing may impact an organization's risk profile. Two such cases are discussed here.

Contracting with Human Annotators and AI Workers

The development of AI-related capabilities often relies on *crowdsourcing* for tasks such as *data labeling*, linguistic experts for fine-tuning, and external consultants for reviews, assessments, and audits. In this regard, organizations should consider policies that require:

- Vetting of vendor practices for ethical labor standards
- Data confidentiality agreements with all annotators
- Clear instructions and QA protocols for human labelers

For example, a company uses annotators to flag toxic content on a social media site.

Employee Use of Generative AI Tools

A different dimension of outsourced AI involves employees themselves, who may introduce third-party risk by:

- Uploading sensitive data into tools like ChatGPT or Gemini
- Copy-pasting code from LLMs
- Using unauthorized plugins or APIs

Organizations should amend their *Acceptable Use Policies (AUPs)* to define:

- Approved AI-related tools and use cases
- Prohibited data types for input (a data classification policy defines data types)
- Logging and review mechanisms for AI usage

Organizations typically conduct training programs to inform workers of the intent behind security policies, acceptable use policies, and other relevant policies, as well as how these policies are enforced.

Monitoring and Reviewing Third-party AI Risk

Third-party risk is not static. Instead, risks related to third parties' use of AI can drastically change, impacting an organization's risk profile. Proactive work is required to detect these changes in risk.

Ongoing Vendor Oversight

An organization's third-party risk management policies should mandate periodic reviews of:

- Model performance (for example, drift, bias)
- Compliance with contract terms
- Changes to vendor business practices or ownership

High-risk vendors should be placed on a continuous monitoring program, including periodic check-ins and annual recertification.

Triggers for Reassessment

New events should prompt an immediate policy-driven reassessment of third-party risk. Triggers occurring within any third party may include:

- Additions or changes in the services provided by the third party
- Security incidents and breaches
- Public controversy involving the vendor (for example, model misuse)
- Technical incidents (for example, *hallucinations* or outages)
- Regulatory changes affecting the model's domain

In the absence of continuous monitoring, organizations are likely to miss the events that would trigger reassessments of vendors and their AI systems.

Summary

This chapter examined the crucial role of policy in establishing oversight and accountability throughout the entire AI lifecycle. As AI becomes more powerful and widespread, organizations must ensure that their governance mechanisms are not limited to reactive measures or narrow technical audits. Instead, effective AI governance requires an integrated policy framework that begins at the earliest stages of system design and continues through deployment, monitoring, and incident response.

AI policy requires a lifecycle approach, highlighting the importance of creating oversight touchpoints at every major system development phase, not unlike that of traditional SDLC. From use case assessment and ethical impact reviews to training, testing, and post-deployment monitoring, each stage presents unique risks. A single misstep, such as biased training data or untested edge cases, can lead to unintended harm, regulatory violations, or public backlash. Policies must define roles, establish review checkpoints, outline documentation standards, and establish escalation paths to ensure that decisions are both intentional and traceable. Ethics by design, data governance, model accountability, and incident response all play essential roles in this continuous oversight strategy.

Organizations need to evaluate and update data privacy and security policies through an AI-specific lens. Traditional frameworks are not built to handle the complexities introduced by AI models that can infer, memorize, or repurpose data in new and unpredictable ways. Organizations must reframe data governance to account for the extended life of data in model artifacts, the risk of reidentification through model probing, and the challenges of granting data subject rights (such as deletion and explanation) once data is baked into a model. This calls for enhanced anonymization standards, lifecycle-based retention controls, and more rigorous application of privacy principles in both structured data and AI-driven outputs. Security policies must also evolve to address AI-specific threats, such as adversarial attacks, model inversion, and compromised supply chains.

Organizations need to expand their governance focus to include third-party risk, an increasingly urgent issue as organizations outsource AI development, data labeling, infrastructure, and tools. Traditional vendor risk management practices fall short when applied to AI systems. Organizations must develop new procurement policies that flag AI-enabled systems, demand transparency about model behavior and data sourcing, and enforce contractual terms covering fairness, explainability, and compliance. AI supply chains, including open-source models and data, must be mapped and vetted, while workforce policies must address both external annotators and internal employees using generative tools. Ongoing monitoring, tiered risk classification, and reassessment triggers ensure that third-party risks are not just identified during onboarding but managed continuously over time.

This chapter reinforced the message that AI governance cannot be siloed, reactive, or static. Instead, it must be dynamic, anticipatory, and deeply embedded in the people, processes, and technologies that support AI systems. Policies serve not only to reduce legal and reputational risk but also to build public trust and operational resilience in an era defined by intelligent systems.

Exam Essentials

Be able to explain how AI governance policies must span the entire AI lifecycle, from use case selection to incident response. Governance isn't a one-time activity. The exam will test your understanding that oversight must be embedded at every stage—beginning with ethical use case vetting, through data sourcing, model design, deployment, and ongoing monitoring. Each phase has unique risks and requires distinct policy attention.

Understand how privacy and security policies must address model-specific risks, such as inference attacks and data memorization. Unlike traditional systems, AI models can unintentionally leak personal data or be probed to reveal training records. You'll need to recognize where traditional controls fall short and how concepts like differential privacy, membership inference testing, and lifecycle data retention apply in AI contexts.

Be aware of procurement and third-party management practices that must include AI-specific reviews, documentation, and risk controls. Exam scenarios may involve vendors providing AI-enabled services. You'll be expected to identify appropriate due diligence steps, such as requesting model transparency, understanding the origins of training data, and incorporating AI-specific contractual clauses regarding fairness and explainability.

Understand traditional data governance principles, such as purpose limitation and data minimization, that must be reinterpreted in the context of AI. AI challenges static notions of data use. The test may assess your ability to reason about issues such as model reuse, synthetic data, or reidentification risks. Understanding how foundational privacy concepts evolve in the age of machine learning is key.

Be able to explain how contractual language with AI vendors must cover bias, explainability, data usage, and auditability. Many third-party risks can only be mitigated through strong contract terms. You may face questions about what should be included in service agreements or how to structure accountability when using external AI models or services.

Review Questions

The following questions are designed to test your understanding of this chapter's material. For more information on how to obtain additional questions, see this book's Introduction. You can find the answers in Appendix A.

1. Which of the following best describes a key principle of lifecycle-based AI governance?
   - **A.** Apply oversight only at the deployment stage.
   - **B.** Delegate AI governance solely to the data science team.
   - **C.** Embed accountability mechanisms at each phase of the AI development process.
   - **D.** Use governance only for models trained on sensitive data.
2. What is the primary risk associated with using anonymized training data in AI models?
   - **A.** The data cannot be used under GDPR.
   - **B.** Anonymized data increases training time.
   - **C.** AI models can potentially reidentify individuals through inference.
   - **D.** Anonymized data always improves model fairness.
3. Which of the following is a valid policy control for managing third-party AI risk?
   - **A.** Relying on vendor self-attestations without further review
   - **B.** Requiring model documentation and training data provenance
   - **C.** Limiting partnerships to vendors with ISO 42001 certification only
   - **D.** Prohibiting all third-party AI tools entirely
4. A policy that mandates periodic monitoring for model drift is best aligned with which lifecycle stage?
   - **A.** Data acquisition
   - **B.** Deployment and monitoring
   - **C.** Training and testing
   - **D.** Use case approval
5. Which of the following is a recommended contractual clause when procuring an AI-enabled service?
   - **A.** The vendor must disclose all proprietary source code.
   - **B.** The vendor is exempt from bias-related complaints.
   - **C.** The vendor must allow audit access and disclose data sources.
   - **D.** The vendor will maintain confidentiality for three months only.

6. What is one reason traditional privacy policies may be insufficient for AI systems?
   - **A.** They do not address the retention of inference outputs.
   - **B.** They assume all data is structured.
   - **C.** They focus only on system availability.
   - **D.** They prohibit all personal data processing.

7. What is the purpose of assigning risk tiers to AI vendors?
   - **A.** To limit the number of contracts issued
   - **B.** To determine marketing budgets for the AI solution
   - **C.** To apply appropriate governance controls based on risk level
   - **D.** To allow procurement to bypass risk reviews for large vendors

8. Which policy control helps enforce purpose limitation in AI pipelines?
   - **A.** Allowing unlimited data reuse for internal projects
   - **B.** Using a use case registry to tag approved applications of data
   - **C.** Prohibiting the use of cloud storage
   - **D.** Allowing only unstructured data in training

9. Which of the following would most effectively reduce bias in AI model development?
   - **A.** Encrypting all datasets before training
   - **B.** Testing model outputs against diverse population subgroups
   - **C.** Using only synthetic data in training
   - **D.** Optimizing models solely for accuracy

10. What does "model inversion" refer to in the context of AI privacy risk?
    - **A.** Reversing a model's optimization objective
    - **B.** Recovering training data from model outputs
    - **C.** Changing labels after training is complete
    - **D.** A method of retraining a model on augmented data

PART II

Legal and Regulatory Obligations





Chapter

4

Privacy and Data Protection Law

THIS CHAPTER COVERS AIGP DOMAIN II, "UNDERSTANDING HOW LAWS, STANDARDS, AND FRAMEWORKS APPLY TO AI," SUBDOMAIN II.A, "UNDERSTAND HOW EXISTING DATA PRIVACY LAWS APPLY TO AI." THE FOLLOWING TOPICS ARE COVERED:

- ✔ Privacy requirements
- ✔ Operational privacy obligations
- ✔ Special data categories

THE AI GOVERNANCE PROFESSIONAL (AIGP) PERFORMANCE INDICATORS IN THESE SUBDOMAINS INCLUDE:

- ✔ Understand how notice, choice, consent, and purpose limitation requirements apply to AI.
- ✔ Understand how data minimization and privacy by design requirements apply to AI.
- ✔ Understand how obligations on data controllers apply to AI (for example, regarding privacy impact assessments, use of third-party processors, cross-border data transfers, data subject rights, incident management, breach notification and recordkeeping).
- ✔ Understand the requirements that apply to sensitive or special categories of data (for example, biometrics).



Artificial intelligence technologies must operate within the boundaries of existing data privacy laws and regulations, although they were not designed with today's AI-specific realities in mind. While AI introduces new complexities, such as opaque decision-making, large-scale data inference, and expanding use cases, the underlying legal principles remain grounded in frameworks like the GDPR (General Data Protection Regulation), CPRA (California Consumer Privacy Act), Brazil's LGPD (Lei de General de Protección de Datos), and others. In this chapter, we explore how time-tested privacy obligations apply in the age of AI, with emphasis on notice, consent, purpose limitation, data minimization, and privacy by design. It outlines the legal responsibilities of data controllers throughout the AI lifecycle, from risk assessments and third-party management to cross-border transfers and breach response. The chapter also delves into the strict requirements surrounding special category data such as biometrics, health, and political beliefs. Bridging legal text with practical AI governance, this chapter equips readers to recognize how compliance duties remain firm, even as technologies continue to evolve.

Notice, Choice, Consent, and Purpose Limitation in AI

Notice, choice, consent, and purpose limitation are pillars of modern privacy practices and regulations. Any AI system that uses personal data must not only comply with applicable laws but embed them into its very design.

This section applies foundational principles from GDPR Articles 5, 6, 12–14, and 21 to the context of AI-driven processing.

Notice

A foundational element of data protection, *notice* refers to informing individuals about how their personal data will be collected, used, stored, and shared. In AI, notice must also disclose how automated decision-making may affect individuals, especially when profiling or inferencing is involved.

Regulations such as the GDPR (Articles 13 and 14) and CPRA mandate that individuals receive timely and clear information, ideally before data collection occurs, including:

- The purposes of processing
- Categories of personal data
- Legal basis for processing
- Data retention periods
- Recipients of the data
- Whether decisions will be made solely by automated means

Unique Challenges in AI

AI systems introduce several complexities that undermine traditional notice practices:

- **Opacity of algorithms:** AI systems often operate as "black boxes," making it difficult to explain how data is processed or how decisions are reached.
- **Inference:** AI can create sensitive data (e.g., political views or health status) from non-sensitive inputs, which isn't explicitly covered in most existing laws.
- **Dynamic behavior:** AI and ML models evolve over time, potentially changing how data is used after initial collection.
- **Multiple data sources:** AI systems often combine data from third parties, sensors, or open web sources, making it more challenging to trace provenance and notify data subjects.
- **Automated decision-making:** Existing provisions (like GDPR Article 22) are narrow and don't fully address AI's nuanced impacts on bias, fairness, and discrimination.
- **Accountability gaps:** Most laws assume a clear "controller" and "processor," but AI supply chains (training data providers, model developers and deployers, API operators, and so forth) make this very confusing.

For example, a U.S. city deploys AI-powered facial recognition to identify persons of interest in real time. Citizens walking through public areas are never explicitly notified of this practice. Although the system is intended for public safety, individuals are often unaware that their biometrics are being collected and processed, which constitutes a direct failure of the notice requirement. Note that this form of AI is prohibited in the EU; this is discussed in Chapter 6.

Key Elements of an Effective AI Privacy Notice

The elements of an effective and defensible privacy notice include:

- Plain language and layered formats
- Explanation of AI's role in decision-making

- Any presence of human oversight in AI decision-making
- Ability to access or contest decisions
- Ability to opt out of automated decision-making
- Clarification of data subject rights
- Justification of algorithmic fairness and necessity

Choice and Consent

Privacy laws, such as the GDPR, require that data subjects be given the option to choose whether an organization uses their personal information and to provide consent if they agree to its use.

Defining Consent and Its Variants

Consent is the individual's voluntary agreement to the processing of their data. Under the GDPR, valid consent must be:

- Freely given
- Specific
- Informed
- Unambiguous
- Revocable
- Recorded

In the context of AI, the principle of meaningful choice becomes increasingly critical, particularly where AI has a significant impact on rights, such as in employment, insurance, credit scoring, or criminal justice. Consent mechanisms should be tailored to the specific context and risk level associated with the processing of personal data.

Consent can take several forms:

- ***Explicit consent*:** Required for sensitive data or high-risk AI use (for example, health-related AI)
- ***Implicit consent*:** Often insufficient for most AI applications
- **Opt-in vs. opt-out:** GDPR generally requires opt-in; some jurisdictions permit opt-out under limited conditions

Obstacles to Meaningful Consent in AI

Obtaining consent as required by law is not always straightforward, particularly when designing business processes and information systems. Obstacles include:

- **Power imbalance:** Users often lack genuine bargaining power, especially when AI is integrated into essential services.
- **Lack of understanding:** Complex AI systems reduce user comprehension of what they're consenting to.
- **Opacity:** Some organizations may not fully disclose all uses of personal information, either because these uses are not well-documented or because the organizations do not want to disclose them.
- **Dark patterns:** User interface designs that trick, coerce, or confuse consumers into giving consent without real deliberation.

For instance, a wellness app collects biometric data to provide personalized recommendations. During onboarding, users are presented with a pre-checked box stating, "I consent to all processing for personalization." There is no granular control or clear explanation of data use. This fails to meet standards for informed or specific consent under GDPR. These challenges highlight why "meaningful choice" in AI contexts requires more than just a checkbox. It demands transparency, accessible explanations, and consent flows designed to respect user autonomy.

Dark Patterns and AI: A Compliance Risk

Dark patterns refer to deceptive interface designs that manipulate users into making unintended or unwise choices. In AI, they can manifest as follows:

- Forced consent through bundled services
- Obscured or hidden opt-out buttons
- Biased default settings
- Illusion of choice, presenting multiple options that all lead to the same outcome

Regulators in the European Union and the United States are increasingly cracking down on these practices.

Purpose Limitation

Purpose limitation means that personal data must be collected for *specified, explicit, and legitimate purposes* and not further processed in a way incompatible with those purposes. The GDPR defines purpose limitation in Article 5(1)(b).

AI systems often push the envelope on this principle due to their reliance on massive, multipurpose datasets and exploratory modeling. A model trained to optimize product recommendations may be repurposed later to detect fraudulent behavior, a secondary use not disclosed to the data subject at the time of collection.

Purpose Creep and Reuse

Many organizations lack the discipline to document and track the stated purposes for collecting personal data about data subjects. Even before the advent of AI, organizations would collect PII for one purpose and then reuse that PII for other purposes. *Purpose creep* refers to the gradual expansion of data use beyond its original intended purpose. It is especially common in AI and ML environments where data is reused across:

- Product features
- Research and development
- Monetization strategies
- Unrelated predictive tasks

While some secondary uses may be compatible with the original purpose (for example, fraud detection in financial services), many are not, particularly when they involve sensitive data or high-impact decisions.

AI-specific Purpose Challenges

Some of the AI-specific challenges related to purpose limitation are as follows:

- **Training vs. inference:** Data collected for training may be reused indefinitely across model updates.
- **Third-party models:** Organizations may purchase or license models trained on datasets that are unknown or undisclosed.
- **Transparency gaps:** Users are rarely informed about the number of downstream purposes a model serves.

In one case, a credit card company collects transaction data to detect and prevent fraud. Later, it uses the same dataset to build a model that infers customer income levels and segments users for marketing campaigns—a use not disclosed in the original privacy notice, which raises concerns about purpose limitation.

Purpose Limitation in AI and ML Systems

To support purpose limitation, organizations should:

- Align dataset documentation with intended use
- Separate datasets for distinct model tasks
- Conduct compatibility assessments for secondary uses
- Update privacy notices with each material change

Data Minimization and Privacy by Design in AI

Data minimization and privacy by design are foundational concepts in modern privacy, and they must be applied to all processes and systems, including those that utilize AI and ML capabilities. These concepts are rooted in GDPR Article 5(1)(c) and Article 25, which require controllers to implement privacy-enhancing measures by default and design.

Data Minimization

The concept of data minimization requires that organizations collect only the personal data that is "adequate, relevant, and limited to what is necessary" for stated intended purposes (GDPR Article 5(1)(c)). The principle is closely linked to the concepts of necessity, proportionality, and accountability.

In AI systems, data minimization becomes complex due to:

- The perceived (and real) need for large training datasets
- The exploratory and experimental nature of model development
- The potential for overcollection in anticipation of future use

While AI thrives on data volume, unbounded data collection increases privacy risks, legal and regulatory exposure, and technical debt. Minimization strategies must therefore be deliberately engineered into the AI lifecycle.

Data Minimization Across the AI Lifecycle

These are some of the ways that data minimization can be implemented:

- **Data collection:** Limit the scope of personal data. Avoid collecting sensitive data unless strictly necessary for the task. Do not collect data based on the speculation that it may be needed later.
- **Preprocessing and feature selection:** Drop features that do not contribute to model performance. Mask or redact sensitive fields.
- **Model training:** Utilize techniques such as *federated learning* (a model training approach where multiple entities train a model while keeping data decentralized) or *differential privacy* to train models without requiring raw personal data.
- **Inference and deployment:** Limit model input to the minimal feature set required to deliver functionality.

For example, a law enforcement agency uses AI-powered predictive policing to identify high-risk areas. Initially, the team collects names, addresses, race, income, and prior convictions. Following a data minimization review, the system is redesigned to utilize only

anonymized crime rates and timestamps aggregated at the neighborhood level, thereby removing direct identifiers and reducing profiling risks.

Synthetic Data and Minimization

Synthetic data is artificially generated data that reflects the statistical properties of real datasets without exposing actual personal information. When used correctly, it can:

- Reduce reliance on sensitive data, including PII
- Support model training and testing
- Improve privacy and compliance posture

However, synthetic data is not always privacy-preserving by default. If it's generated from overfitted models or small populations, it may still pose a risk of re-identification.

Balancing Utility and Minimization

AI development teams often argue that more data yields better models. While this can be true in early stages, diminishing returns set in, and the marginal value of additional personal data may not justify the increased risk. Privacy-preserving machine learning (PPML) disciplines are evolving to challenge the "more is better" mindset by emphasizing *smart data*, not *big data*.

Data minimization can be achieved through two main approaches:

- **Collection discipline:** Reducing the scope, detail, and volume of personal data collected
- **Technical safeguards:** Reducing identifiability and exposure of data that has been collected

Data minimization techniques include:

- Data masking and tokenization
- Pseudonymization
- On-device processing
- Privacy-preserving analytics (for example, *homomorphic encryption*, a form of encryption in which computations may be performed on encrypted data without having to decrypt it)

Privacy by Design (PbD)

Privacy by design is an architectural and organizational approach that embeds privacy protections into systems, processes, and technologies from the outset, rather than adding them on later. Originally formulated by Dr. Ann Cavoukian in the 1990s, the concept gained

legal force in GDPR Article 25 and is also referenced in frameworks like ISO/IEC 27701 and NIST AI RMF.

The goal of privacy by design in AI is to proactively engineer AI and ML systems that minimize risks to individuals, while preserving model utility and business goals.

Seven Foundational Principles of Privacy by Design

The foundational principles of privacy by design are:

- Proactive, not reactive; preventive, not remedial
- Privacy as the default setting
- Privacy embedded into design
- Full functionality: positive-sum, not zero-sum
- End-to-end security: full lifecycle protection
- Visibility and transparency
- Respect for user privacy: keep it user-centric

These principles must be interpreted within the unique context of AI development, deployment, and monitoring.

Implementing Privacy by Design in AI Development

Privacy by design can be implemented in various stages of development, including the following.

In the problem framing stage:

- Conduct a *Data Protection Impact Assessment (DPIA)* to identify privacy-related risks prior to system design
- Assess necessity and proportionality
- Determine whether the AI capability could be achieved with less intrusive methods

In the data and model development stage:

- Use privacy-preserving training methods
- Apply bias detection and model interpretability tools
- Validate fairness and explainability

In the system deployment and maintenance stage:

- Limit retention of raw inputs
- Enable user contestability and feedback
- Continuously monitor drift and compliance degradation

For example, a startup organization develops a natural language processing model to assist in legal document review. Instead of collecting entire case files from law firms, the team implements on-premise processing where documents never leave the client's

environment. The model is trained on deidentified data extracted with legal oversight. Logging is minimized, and audit trails are encrypted for enhanced security. These measures exemplify the concept of privacy by design in action.

Embedding Privacy by Design in MLOps Pipelines

Privacy by design can be embedded into modern AI development pipelines through:

- **CI/CD integration**: Embed privacy scans in automated build processes.
- **Privacy linting**: Check for dangerous data uses or code paths.
- **Model cards**: Document privacy characteristics and risk assessments.
- **Drift detection**: Monitor when real-world data diverges from training data, potentially creating new risks.

MLOps teams should treat privacy as a key performance metric alongside accuracy and latency.

Aligning Privacy by Design with Organizational Roles

To operationalize privacy by design in AI, cross-functional coordination is essential and may include work from these groups:

- **Privacy teams** must work with data scientists and MLOps engineers.
- **Product managers** should champion privacy-respecting features.
- **Legal and compliance teams** must oversee risk assessments and documentation.
- **Security teams** must implement access controls, encryption, and monitoring to protect personal data in use, in transit, and at rest.
- **Executives** must foster a culture of responsible AI development.

Embedding privacy into AI system development is not just a technical issue but a strategic cultural commitment requiring investment, training, and leadership support.

Practical Tools and Frameworks

A growing number of frameworks support the integration of privacy by design in AI system development:

- **ISO/IEC 23894:** AI risk management
- **NIST AI RMF:** Core functions include "map, measure, manage, govern"
- **ENISA Guidelines:** Focus on AI-specific attack surfaces and mitigations
- **AI-specific DPIAs:** Include templates developed by ICO (UK), CNIL (France), and others

Governance professionals should ensure that corporate policy stipulates that AI projects consistently utilize these tools and that outputs (including risk scores, mitigation plans, and residual risks) are logged for audit readiness.

Practical Implications and Governance

Up to this point, this chapter has covered well-known privacy principles in the context of AI systems. Now, it examines how to apply these principles in practice. This section reflects the accountability obligation under GDPR Article 5(2), reinforced by Article 24, which outlines the responsibilities of controllers.

AI is not exempt from privacy laws, and the most responsible organizations recognize that privacy is not a constraint on innovation but a catalyst for building trustworthy, high-quality systems.

Compliance is not a checkbox activity. It is a dynamic, cross-functional process that must evolve with AI systems. Governance professionals must possess the fluency to interpret legal mandates and the agility to adapt them to machine learning pipelines, model deployment workflows, and real-time decision-making environments.

As AI systems reshape how personal data is processed and decisions are made, applying classical privacy principles remains essential and increasingly complex. Notice, choice, consent, and purpose limitation serve as foundational rights that uphold individual autonomy and trust. Data minimization and privacy by design ensure these rights are respected throughout the technical architecture and operational practice of AI.

Operationalizing Privacy Principles in AI

Translating the concepts of notice, choice, consent, purpose limitation, data minimization, and privacy by design into actionable practices requires robust governance structures. These principles must be operationalized across the entire AI development and deployment lifecycle, from ideation to model retirement.

This requires:

- Embedding privacy checkpoints in AI workflows
- Aligning legal, technical, and business teams
- Establishing accountability and oversight mechanisms

Governance Touchpoints Across the AI Lifecycle

Table 4.1 includes key activities performed at various stages of AI system development.

Governance professionals should ensure that privacy is not treated as a one-time activity but as a continuous obligation.

TABLE 4.1 Privacy Activities in the AI System Development Lifecycle

AI Lifecycle Stage	Privacy Activities
Problem definition	Conduct DPIAs, define legitimate purpose, assess necessity
Data acquisition	Validate lawful basis, apply data minimization, assess third-party sources
Model development	Perform bias and fairness audits, embed privacy-preserving ML techniques, apply pseudonymization/anonymization
Model evaluation	Confirm outputs align with original purpose, validate explainability and transparency
Deployment	Enable opt-outs, provide contestability and notices, apply access controls, encrypt data in transit and at rest
Monitoring and updates	Monitor drift, ensure purposes remain consistent, reevaluate risks

Privacy Roles and Responsibilities

As a part of an AI governance program, privacy-related responsibilities should be formally assigned to various positions, including:

- **Data protection officer (DPO):** Ensures legal compliance and advises on DPIAs, risk management, and subject rights.
- **Chief AI officer/AI governance lead:** Drives ethical development and use of AI, ensuring alignment with internal policies and external regulations.
- **MLOps and engineering teams:** Implement privacy-respecting infrastructure, logging, monitoring, and model lifecycle management to ensure data integrity and security.
- **Product and UX designers:** Ensure that consent mechanisms and user interfaces are intuitive, fair, and free of dark patterns.
- **Security teams:** Implement technical safeguards such as encryption, pseudonymization, intrusion detection, and incident response to protect AI system data.
- **Legal and compliance teams:** Translate legal requirements into practical controls, conduct audits, and manage breach response.

Additional considerations, such as applicable regulations, industry sector, and organizational culture, may necessitate additional responsibilities.

Documentation and Auditability

Regulators increasingly expect AI systems to be *accountable* and capable of demonstrating compliance through documentation and evidence. Key accountability artifacts include:

- Data flow diagrams
- Model cards and system cards
- Records of processing activities
- Logs of consent and purpose linkage
- Risk assessment outputs (DPIAs, LIA balancing tests)
- Validation and testing reports

Having a centralized AI risk register or documentation hub can significantly improve audit readiness and support transparency obligations.

AI regulations are still emerging. The absence of AI-related regulation in a given industry sector or locale may be short-lived. Organizations should proceed cautiously, assuming that AI regulations may soon be enacted.

AI Governance Playbook: Key Policies to Implement

When developing an AI governance program, some of the policies to develop or revise include:

- **AI use policy:** Defines acceptable and prohibited AI uses
- **Privacy impact policy:** Requires DPIAs for high-risk AI systems
- **Data governance policy:** Covers lineage, quality, classification, and minimization
- **Model governance policy:** Describes versioning, retraining, and validation
- **Third-party risk policy:** Governs vendor AI tools and data sources
- **Incident response policy for AI systems:** Integrates AI-specific breach, model failure, or misuse scenarios into the organization's security incident response plan
- **Change management policy for AI models:** Defines review, approval, and rollback procedures when updating AI models

Scaling Governance

Governance processes can be challenging to scale for larger organizations. Some ideas for scaling AI governance include:

- **AI risk committees** or review boards
- **Tiered governance** based on AI risk levels (for example, high-risk AI vs. low-impact automation)
- **Tooling integration** with data catalogs, MLOps, and governance platforms

- **Training programs** to build awareness of AI privacy risks across the organization
- **AI centers of excellence** (COE) to encourage collaboration and learning across teams

AI governance must be tailored to context but always grounded in the principles of accountability, transparency, and fairness.

Data Controller Obligations in the AI Context

Data controllers play a significant role in the processing of personal information. This section explores the details of data controller obligations.

The obligations in this section derive from multiple GDPR articles, including Articles 5, 24–30, and 44–49, which govern the duties of controllers, processor relationships, and cross-border data transfers.

AI systems introduce novel risks to personal data, and data controllers bear the brunt of these risks. A data controller, as defined in GDPR Article 4(7), is the entity that "determines the purposes and means of processing." Whether deploying AI in healthcare, finance, advertising, or recruitment, organizations making these decisions carry primary responsibility for ensuring compliance with data protection laws.

AI doesn't alter the core legal privacy framework, but it does complicate it. Because AI often involves probabilistic modeling, opaque inference, and the reuse of large-scale data, the controller's obligations require adaptation. From conducting DPIAs to honoring data subject rights and managing incidents, controllers must reframe their approach to compliance through the lens of algorithmic systems.

This section provides a comprehensive overview of the legal and operational responsibilities that apply to controllers deploying or commissioning AI systems. Each obligation is explored in the context of practical AI risks and illustrated with examples and governance guidance.

Privacy Impact Assessments and Risk Management

Risk assessments are the cornerstone of any privacy program, enabling the proactive discovery of privacy-related risks that a new process or system may introduce. When AI or ML is a part of a business process, a risk assessment is even more important.

The Role of DPIAs in AI Development

A Data Protection Impact Assessment (DPIA) is a structured process used to identify, assess, and mitigate risks to individuals arising from the processing of high-risk data. According to GDPR Article 35, controllers are required to conduct a DPIA when processing "is likely to result in a high risk to the rights and freedoms of natural persons."

AI almost always triggers that high-risk threshold, especially when it involves:

- Automated decision-making with legal or similarly significant effects
- Large-scale processing of sensitive categories of data
- Monitoring of publicly accessible areas
- Systematic profiling

AI's unpredictability, scale, and potential for hallucinations and bias elevate the risk profile beyond traditional IT systems, making DPIAs not only legally required but strategically essential.

Like other aspects of AI governance, the rigor and depth of a DPIA can be predetermined through the initial risk tier assigned to the new or revised system. A DPIA for a high-risk AI system would contain more detailed information than a DPIA for a low-risk system.

DPIA Lifecycle in AI Projects

A DPIA should be conducted early in the AI development lifecycle, ideally at the design stage, and then updated throughout the project. The assessment typically includes:

- A description of the processing activity and its purpose
- A description of the training data to be used (at the field level)
- A description of the data to be used during system operations (at the field level)
- A list of applicable cybersecurity, privacy, and AI-related laws and regulations
- An evaluation of necessity and proportionality
- A risk analysis from the data subject's perspective
- Measures to address or mitigate risks
- Consultation with stakeholders (including DPOs or regulators, if required)

For instance, a retailer plans to launch a customer segmentation AI system that infers income levels from transaction and location data to tailor marketing. The DPIA reveals a risk of socioeconomic discrimination and potential exclusion. Mitigation includes excluding location features, increasing model transparency, and ensuring opt-out options.

When Must a DPIA Be Conducted? (GDPR Triggers for AI)

Under GDPR, DPIAs are mandatory if the AI system involves:

- Automated decision-making with legal/similar effect (for example, job applications, credit scoring)
- Profiling or predictive analytics at scale
- Sensitive data such as biometrics, health data, or religious beliefs

- Vulnerable populations (for example, children, the elderly)
- Systematic monitoring of behavior (for example, via wearables or apps)

Even where not explicitly required, regulators encourage DPIAs as a best practice for all AI applications that process personal data. An initial DPIA may also serve as a system benchmark that can be reviewed later.

Common Pitfalls in AI DPIAs

Some organizations may struggle to adapt traditional DPIAs to AI systems due to:

- Lack of clarity on system boundaries (for example, where the model ends and business logic begins)
- Inadequate risk framing (focusing on technical risks instead of human impacts)
- Failure to revisit DPIAs post-deployment (especially in continuously learning models)

Best practices for AI DPIAs include:

- Using AI-specific DPIA templates (for example, from CNIL or ICO)
- Involving interdisciplinary teams (privacy, engineering, ethics, UX)
- Documenting model logic, data sources, and residual risk justifications

Aligning with Emerging Frameworks

Rather than designing AI DPIAs from scratch, organizations should examine other frameworks that include DPIA-style assessments, including:

- **NIST AI RMF:** Recommends mapping risks early and iteratively
- **OECD AI principles:** Emphasize stakeholder consultation and transparency
- **EU AI Act:** Introduces mandatory conformity assessments for high-risk AI systems

Controllers should anticipate these developments and integrate multi-framework risk evaluations where applicable.

Using Third-party Processors in AI Projects

It seems that everything in the information technology world is being outsourced to vendors and service providers. In the context of privacy and AI systems, many organizations are opting to outsource some (or many) aspects of data processing to external service providers.

Controller vs. Processor in the AI Context

Under GDPR Article 4(8), a processor is any party "which processes personal data *on behalf of* the controller." In AI projects, third-party processors are often engaged for:

- Cloud storage or compute environments
- Model training or MLOps services
- Pretrained or foundation models from vendors
- Data classification, labeling, and annotation services
- AI-as-a-Service platforms

Controllers must ensure that processors act *only* on their documented instructions and implement appropriate technical and organizational safeguards to protect their data from misuse and compromise.

The complexity and opacity of AI supply chains can blur the controller-processor boundary. For example, an AI vendor offering "black box" models with limited configuration may function more like a joint controller than a processor.

Contractual Requirements

Controllers must have a binding contract (commonly a Data Processing Agreement or DPA) with each processor. According to GDPR Article 28, the contract must specify:

- Subject matter, duration, nature, and purpose of processing
- Types of personal data and data subjects
- Processor obligations (for example, security, confidentiality, subprocessing)
- Assistance with rights requests and DPIAs
- Return or deletion of data after processing ends
- Audit and inspection rights

For example, a financial institution contracts with a machine learning vendor to analyze call center recordings for quality assurance. The DPA includes a clause allowing the vendor to use the data to improve its own models, thereby undermining the controller's instructions and potentially violating purpose limitation and data transfer rules, especially if the vendor is training unrelated services.

Due Diligence Checklist for AI Vendors

An organization's procurement or third-party risk management function should utilize a detailed checklist or procedure that includes the following steps for AI vendor due diligence:

- Does the vendor use subcontractors for model development?
- Can the vendor articulate the provenance and composition of training data?
- Is the model explainable or auditable?
- What logging and security mechanisms are in place?

- Can the vendor support data subject requests?
- Does the vendor agree to process only on instruction and not reuse data?
- Does the vendor exercise due care in cybersecurity and privacy?
- Will the vendor agree to an audit in the event of a breach?

Due diligence should include technical and legal vetting, not just a review of terms of service.

Processor Drift and Compliance Gaps

AI systems are often dynamic, meaning that a processor's behavior may change post-deployment. Examples of changes include:

- Continuous training with customer data
- New feature rollouts that alter data collection and/or processing
- Model updates that introduce profiling

Controllers should monitor processors throughout the contract lifecycle and require:

- Notification of a change in the processor's third-party processors
- Change notification procedures
- Security and functionality audit logs
- Access to performance and fairness metrics

Strong governance is crucial in preventing "processor drift," where delegated tasks become uncontrolled and unmanaged processing.

Cross-border Data Transfers

A key component in privacy laws involves the location of processing and storage of personal data. Cross-border data transfers are common, and they must be documented and evaluated for compliance.

Global Data Flows in AI Systems

AI ecosystems are inherently international. Typical cross-border activities include:

- Serving customers who live overseas
- Training models in offshore cloud environments
- Accessing pretrained models from non-local vendors
- Using global telemetry or IoT data streams
- Outsourcing data labeling to lower-cost regions

Under GDPR, personal data may only be transferred outside the EEA if appropriate safeguards are in place. This includes:

- Adequacy decisions (for example, Japan, South Korea, UK)
- *Standard Contractual Clauses (SCCs)*
- *Binding Corporate Rules (BCRs)*
- Derogations for specific situations (for example, explicit consent, legal claims)

AI projects often overlook that *access* from outside the EEA (even if no data is stored) counts as a transfer.

AI-specific Data Transfer Challenges

Several challenges make AI data transfers especially risky:

- **Ongoing access vs. one-time transfers:** Models may pull training data periodically or continuously from global sources.
- **Inference risks:** Even if raw data stays local, inference outputs can reveal sensitive information.
- **Data localization laws:** Countries such as China and India are enforcing AI-relevant data localization laws that may conflict with GDPR transfer rules.

In one instance, a voice assistant company based in the European Union uses a United States–based transcription service to process audio commands. The SCCs are in place, but subsequent legal discovery in the United States exposes personal data of EU citizens, raising issues under Schrems II and EU data transfer rules.

AI Inference as a Transfer?

Recent legal interpretations suggest that:

- *Remote access* to data = transfer
- *Real-time model inference* from foreign servers = potential transfer
- *Use of third-party AI APIs* (for example, image recognition) = transfer if personal data is involved

Controllers must carefully assess AI services' architecture to determine whether cross-border risk exists—even when data "stays in place."

Transfer Impact Assessments (TIAs)

Since the Schrems II ruling invalidated the EU–U.S. Privacy Shield, organizations must conduct a *Transfer Impact Assessment (TIA)* when relying on SCCs or BCRs. A TIA should consider:

- Legal environment of the destination country
- Applicable privacy laws

- Nature of the data and processing
- Existence of government surveillance or access
- Supplementary measures (for example, encryption, pseudonymization)

For AI systems, TIAs may also need to address:

- Remote model training using EEA data
- Vendor access to telemetry or crash logs
- Use of pretrained models built on non-EEA data
- Cross-border API calls or inference requests that transmit personal data to non-EEA servers
- Third-party hosting or orchestration services that run AI models on infrastructure outside the EEA

Controllers must document TIAs and be prepared to justify their decisions to supervisory authorities.

Data Subject Rights and AI

Under GDPR and similar frameworks (for example, CPRA, LGPD), data subjects have a suite of rights, including:

- Right of access (GDPR Article 15)
- Right to rectification (GDPR Article 16, CPRA 1798.106, LGPD Article 18(III))
- Right to erasure ("right to be forgotten," GDPR Article 17, CPRA 1798.105, LGPD Article 18(IV))
- Right to opt out of automated decision-making (CPRA 1798.121)
- Right to restriction of processing (GDPR Article 18)
- Right to data portability (GDPR Article 20, LGPD Article 18(V))
- Right to object (GDPR Article 21)
- Right not to be subject to solely automated decisions with significant effects (GDPR Article 22)

Controllers must ensure these rights remain actionable, even when decisions are aided by complex AI systems.

AI-specific Implementation Challenges

In AI systems, these rights raise technical and operational challenges:

- **Access and explainability:** It's difficult to explain outputs from black box models, much less provide intelligible explanations to data subjects.
- **Rectification and erasure:** How do you correct or delete data embedded in model weights?

- **Portability:** AI outputs are often model-dependent and not easily portable in any meaningful way.
- **Objection to profiling:** Profiling is intrinsic to many AI systems, requiring robust opt-out mechanisms.
- **Restriction of processing:** Halting the use of personal data in active models can be operationally challenging, especially for continuous learning systems.

For example, a ride-sharing app utilizes an AI model to deactivate drivers based on their predicted risk of fraud. A driver receives an automated notice and requests details. The company cites "proprietary algorithms" without providing a meaningful explanation, thereby violating GDPR Articles 15 and 22.

GDPR Article 22: Human-in-the-loop or Window Dressing?

Article 22 prohibits decisions "based solely on automated processing" that significantly affect individuals, unless:

- It is necessary for a contract
- It is authorized by law
- It is based on explicit consent

Even with human review, regulators require *genuine oversight*—not rubber-stamping. Controllers must ensure that human reviewers can override AI decisions and understand the basis for these interventions.

Practical Solutions

Organizations facing data subject rights challenges have a few available remedies. Some of these remedies are closely tied to the design of an AI system and cannot be easily retrofitted.

- **Design for rights:** Design and build explainability and logging into models from the outset.
- **Model documentation:** Maintain metadata about inputs, versions, logic, and impacts.
- **Layered explanations:** Combine high-level summaries with technical details where needed.
- **Anonymization and pseudonymization:** Train AI models with anonymized or pseudonymized data whenever possible.
- **Deletion strategies:** Utilize retraining pipelines that exclude deleted records or limit personal data in training datasets.

Respecting legal and human rights in AI is achievable, but it must be intentional and incorporated into the AI system's initial design. AI has no moral intuition; it only follows

the rules and objectives defined by humans. It doesn't inherently know what is "right," so anchoring the concept of "rights" to legally defined and universally recognized rights (for example, human rights frameworks and privacy laws) makes it objective, enforceable, and consistent across contexts.

Incident Management and Breach Notification

Formal, documented incident management capabilities are a core element of any cybersecurity and privacy program. This section addresses AI-related additions and considerations for incident and breach management. For a comprehensive review of cybersecurity and privacy incident management, refer to my study guides on the CISM (Certified Information Security Manager), CRISC (Certified in Risk and Information Systems Control), CIPM (Certified Information Privacy Manager), and CDPSE (Certified Data Privacy Solutions Engineer) certifications.

Reframing Breaches in AI Systems

AI systems introduce novel types of security and privacy breaches, including:

- **Training data leakage:** For example, via inversion or membership inference attacks, or gradient extraction attacks
- **Output leakage:** For example, LLMs reproducing training data verbatim or revealing sensitive content through prompt injection
- **Bias amplification:** Producing harmful and unfair decisions at scale
- **Model drift or poisoning:** Undermining decision quality through adversarial inputs

Traditional security incidents (for example, credential theft, malware) still apply, but controllers must expand breach detection to cover AI-specific vectors.

GDPR Breach Notification Requirements

According to GDPR Article 33, controllers must notify the supervisory authority within 72 hours of becoming aware of a personal data breach. Also, if there is a high risk to data subjects' rights and freedoms, they must be informed without undue delay. Such notifications must include:

- Nature and scope of the breach
- Categories and number of data subjects affected
- Likely consequences
- Measures taken or proposed to mitigate harm

Failure to comply can result in fines of up to €10 million or 2% of the company's global turnover, and potentially higher penalties if combined with other infractions.

For instance, a research lab publishes an AI demo trained on scraped online health forums. During operation, the model outputs sensitive user posts nearly verbatim.

Once identified, the controller must assess whether this constitutes a breach and whether affected users can be identified and notified.

AI Incident Playbook Essentials

Every controller deploying AI should maintain an AI-aware incident response plan and playbooks. Include:

- AI model inventory, data provenance, and data lineage
- Breach impact assessment criteria for inference outputs
- Contact points for vendors or subprocessors
- Legal escalation routes for DPIAs and Article 33 reviews
- Training for response teams on AI-specific vectors

The faster your team can distinguish *model bugs* from *personal data breaches*, the better your compliance posture.

AI Incident Detection and Preparedness

AI complicates incident detection in several ways:

- **Silent failures:** Model outputs may degrade without obvious signs.
- **Lack of observability:** Poor logging or opaque pipelines hinder investigation.
- **Unstructured data:** Breaches may affect documents, media, or embeddings not monitored by standard DLP tools.
- **Third-party dependency:** Security incidents in upstream model providers, APIs, or hosting platforms may leak personal data without triggering internal alerts.

To address these risks, controllers should:

- Implement logging and anomaly detection at the model and data layers.
- Use model cards and system cards that document data sources and risks.
- Establish data breach drills that include AI scenarios.
- Perform periodic risk assessments, including third-party risk assessments, to identify additional risks.

Recordkeeping and Accountability

The strength and integrity of any governance program rely on effective documentation, thorough recordkeeping, and robust accountability.

Controllers Must Demonstrate Compliance

The GDPR's accountability principle (Article 5(2)) requires that controllers not only comply with data protection principles but also *demonstrate* that compliance. This demonstration includes maintaining detailed internal records (for each AI system in use) of:

- Processing purposes
- Categories of data subjects and personal data
- Recipients and transfers
- Retention periods
- Security and safeguarding measures

For AI systems, this recordkeeping obligation becomes significantly more complex, especially when models are dynamic, updated regularly, or embedded in other products and services.

Organizations employing risk tiers for AI systems can require more rigorous recordkeeping on high-risk AI systems, with less stringent requirements on lower-risk systems.

What to Document in AI Projects

To fulfill accountability obligations, controllers should maintain artifacts that describe:

- **Purpose definition:** Why the AI was developed and what decisions it supports
- **Training data provenance:** Where the data came from, how it was cleaned, and its legal basis
- **Model architecture and logic:** What type of model is used and how it functions
- **Evaluation results:** Accuracy, fairness, bias, robustness, and explainability scores
- **Risk assessments:** DPIAs, TIAs, and justification for any residual risk
- **Version history and change logs:** For model updates or policy shifts
- **Reviews and approvals:** Who participated in system development gate reviews, and who approved system development projects to proceed

For instance, a bank builds a credit-scoring model that it updates quarterly. It maintains an AI registry containing each version's training data snapshot, algorithmic assumptions, bias mitigation techniques, and output testing. When audited by regulators, the documentation supports claims of fairness and proportionality.

What Belongs in Your AI Compliance Binder

Organizations building their AI governance recordkeeping requirements can include these requirements:

- Model cards (summaries of AI models, risks, intended uses)
- System cards (documentation of end-to-end AI systems)

- Records of processing activities (GDPR Article 30)
- Records of events, incidents, and breaches
- Consent logs, opt-outs, and objection handling
- DPIAs, TIAs, and consultation records
- Documentation of human review processes for GDPR Article 22 decisions

These assets are essential not only for regulators but also for internal audits, ethics boards, and future product decisions.

Tools and Frameworks for Governance

Many organizations streamline recordkeeping of their governance programs with tools and frameworks, including:

- **Governance platforms:** Such as OneTrust, Navex, Collibra, or BigID
- **MLOps integrations:** Include privacy logging in CI/CD pipelines
- **Open documentation tools:** Like Model Cards++ or datasheets for datasets
- **AI risk registers:** Track known risks, mitigations, and review cycles

Embedding documentation in the development process, rather than treating it as a compliance afterthought, supports continuous governance.

Understanding the Requirements That Apply to Sensitive or Special Categories of Data

In addition to standard PII, there may be additional requirements for specific categories of personal data, including healthcare, biometrics, sexual orientation, and religious affiliation. This section addresses some of these special cases.

The processing of sensitive data in AI systems is governed by GDPR Article 9, along with related provisions in the CPRA, LGPD, and other global frameworks.

AI systems are hungry for data, and some of that data is more consequential than others. When AI systems process special categories of personal data, the stakes are higher, the risks are sharper, and the regulatory expectations are stricter.

Under laws like the EU's GDPR, California's CCPA/CPRA, Brazil's LGPD, and others, certain types of personal data are considered *inherently sensitive*. These include data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, genetic data, biometric data for identification purposes, health data, sex life, sexual orientation, and more.

In the context of AI, these data categories are often processed at scale and sometimes *inferred* rather than explicitly collected, elevating both the compliance burden and the ethical duty on developers, deployers, and data controllers.

This section examines the legal, operational, and governance requirements related to sensitive data in AI systems. It covers legal bases for processing, specific categories such as biometrics and health data, associated risks, technical safeguards, and rights-based obligations.

Legal and Regulatory Foundations

This section explores how GDPR and other regulations treat special classes of sensitive data.

GDPR Article 9 and the Prohibition Baseline

GDPR Article 9 sets the tone clearly when it states, "The processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs … shall be prohibited."

Note that this is the default position: the processing of special category data is forbidden unless one of a narrow list of exceptions applies. These exceptions include:

- Explicit consent
- Employment and social protection law
- Vital interests (for example, medical emergencies)
- Public interest in the area of public health
- Archiving, research, or statistical purposes (with safeguards)
- Establishment or defense of legal claims

When AI systems touch on these categories, even indirectly, the controller must ensure an appropriate legal basis, plus technical and organizational safeguards to protect the data.

Key International Equivalents

Beyond GDPR, other jurisdictions also impose heightened duties on sensitive data. A few examples are cited in Table 4.2.

TABLE 4.2 Selected Privacy Regulations Concerning Sensitive Data

Region	Law	Sensitive Data Definition	Key Provisions
California	CPRA	Biometric, health, race, religion, union, precise geolocation	Requires notice, opt-out, limited use
Brazil	LGPD	Biometric, health, racial or ethnic origin, religious/philosophical belief, political opinion, union membership	Separate legal basis, stricter safeguards
Canada	PIPEDA (interpreted via OPC)	Health, biometrics, ethnicity, religious beliefs	Consent must be meaningful; biometrics are high-risk
India	DPDP Act (2023)	Health, biometric, financial, genetic, sex life, political beliefs	Consent-based, with future AI-specific rules anticipated

What Qualifies as "Sensitive" Data?

While definitions vary, the following are commonly treated as sensitive across jurisdictions:

- Biometrics (for identification)
- Health and genetic data
- Racial or ethnic origin
- Religious or philosophical beliefs
- Political opinions and affiliations
- Trade union membership
- Sex life or sexual orientation
- Precise geolocation (in some jurisdictions)

Some laws also treat financial, government-issued ID, or children's data as sensitive, depending on context.

AI's Unique Risk Profile for Sensitive Data

AI can process sensitive data deliberately, incidentally, or inferentially:

- **Deliberate:** A health AI tool uses electronic medical records.
- **Incidental:** A CV-sorting tool collects applicants' political affiliations from past activism.
- **Inferential:** A facial analysis model predicts ethnicity or emotion without being explicitly told.

Each of these scenarios carries risks:

- Lack of a valid legal basis for collection and processing
- Discrimination and bias
- Lack of transparency
- Surveillance and *chilling effects* (when persons are discouraged from partaking in legal activities)

Controllers must understand these risks and take proactive measures, especially when deploying high-risk systems that may affect employment, healthcare, law enforcement, or access to essential services.

Effective data processing impact assessments (DPIAs) anticipate these types of risks.

Biometrics and Other High-risk Categories of Sensitive Data

Special categories of data, as defined in GDPR Article 9, referred to as "sensitive data," receive heightened protection due to their potential to cause significant harm or discrimination to data subjects if misused. These data types can expose individuals to surveillance, exclusion, harassment, or manipulation, particularly when processed by AI systems at scale.

AI systems don't just consume sensitive data; they can also generate it; for example, by inferring race, health conditions, or political views from innocuous inputs, such as browsing behavior, voice tone, or purchase history.

This section explores the major sensitive data categories most often implicated in AI development and deployment: biometrics, health and genetic data, religion and beliefs, political opinions, and demographic or labor-related attributes like race and union membership.

Biometrics

Biometric data refers to personal data resulting from specific technical processing related to the observable or measurable physical, physiological, or behavioral characteristics of a person that allows or confirms unique identification.

Common biometric data types include:

- Facial features
- Fingerprints and palm prints

- Iris and retina scans
- Voiceprints
- Gait and posture
- Typing rhythm and mouse movement

Biometric systems are commonly used for:

- Authentication (for example, logging into a system or unlocking a door)
- Identification (for example, matching a face against a watchlist)
- Surveillance and analytics (for example, estimating age or emotion in public spaces)

Biometric data is especially sensitive because it:

- Cannot easily be changed if compromised (you can get a new credit card number, but you cannot get new fingerprints)
- Is often collected passively or without consent
- May be repurposed for surveillance, profiling, or behavioral manipulation

For example, a shopping mall uses facial recognition to identify repeat shoplifters. The same system is quietly used to track customer movement patterns and trigger targeted ads. Customers are not notified or given a chance to opt out. This example involves biometric data protection, purpose limitation, and transparency obligations.

Are All Biometric Systems "High Risk"?

Not always. GDPR only applies heightened protection to biometric data *used to uniquely identify a person.* A biometric system used solely for categorization (for example, age bracket estimation without linking to identity) may not trigger Article 9. However, it may still create ethical and legal risks, depending on the context.

Healthcare and Genetic Data

AI is transforming healthcare, from predictive diagnostics and personalized medicine to operational optimization and patient triage. However, health and genetic data are among the most sensitive forms of personal information.

Key concerns include:

- Medical and employment discrimination
- Inference of conditions without consent
- Reidentification from anonymized datasets
- Use of patient data for unrelated research or training

Genetic data introduces additional complexity due to its familial and hereditary implications, potentially affecting individuals who never gave their consent.

An example telehealth platform uses natural language processing to screen patients for depression. It stores transcripts and uses them to fine-tune models, including sentiment analysis services offered to insurers. Patients were not informed that their interactions would be used in this manner, raising concerns about the use of health data and its potential secondary applications, such as employment and credit checks.

Privacy Tradeoffs in Digital Therapeutics

Digital therapeutics and health-focused AI tools offer clinical benefits but also create new vectors of exposure:

- Apps may leak location or device metadata
- Cloud storage may lack healthcare-grade security and privacy controls
- Data sharing with third parties (for example, advertisers or model trainers) may go unmonitored

Organizations should always treat "consumer health" data with the same care as regulated medical data.

Religious and Philosophical Beliefs

Belief systems, whether religious, philosophical, or ideological, are explicitly protected under GDPR. AI systems may encounter this data via:

- Survey or profile responses
- Social media
- Textual analysis (for example, sermons, prayers, or religious vocabulary)
- Inference based on behavior (for example, holiday travel, food purchases)

Processing this information without a strong legal basis can violate GDPR Article 9 and risk:

- Profiling or stereotyping
- Bias and discrimination
- Chilling effects on free expression
- Targeted advertising or manipulation

An example language model fine-tuned on public posts begins assigning users to "likely belief clusters." A streaming service uses this to recommend religious-themed content, and later to restrict or alter such recommendations. Users are often unaware that their spiritual identity has been inferred and acted upon.

Faith in the Dataset: Hidden Risks

Even if religious beliefs are not explicitly collected, they can be embedded in training data. Developers must consider:

- Bias in content moderation systems
- AI outputs that reflect or suppress belief-based expression
- Mislabeling or overgeneralization of faith-related attributes

Bias audits should include checks for religious stereotyping, discrimination, and marginalization.

Political Opinions and Affiliations

Political data is often inferred from user behavior, especially in advertising, social media, and content personalization. This category includes:

- Party membership
- Voting history (specific or inferred)
- Stated political views
- Engagement with political content

Misuse of political data in AI systems can:

- Manipulate opinion (for example, microtargeting voters)
- Suppress dissent
- Influence electoral outcomes
- Create profiling-based discrimination

An example AI system is used by a social media platform to assess users' political leanings for targeted advertising. The system learns to correlate news article clicks with specific ideologies and begins curating content accordingly. Political filter bubbles intensify, and some users are targeted with misinformation.

Political Data and Decision Automation

Automated decisions based on political views, such as ad eligibility, financial services, or employment screening, are likely prohibited under GDPR Article 22 unless explicit consent is provided and the requirements of Article 9 for special category data are also met.

In practice, this usually means obtaining *explicit consent* or having a clear legal mandate. Controllers must ensure:

- Such decisions are explainable and contestable
- Political data is never used without a clear legal basis
- Consent is freely given, specific, and informed

Race, Ethnicity, and Union Membership

AI systems frequently process race and ethnicity either directly (for example, through demographic forms) or inferentially (for example, from names, language, or location). This practice is particularly problematic in résumé screening, loan applications, housing platforms, and policing and surveillance.

Similarly, union membership can be inferred from email content, social media activity, or organizational behavior, often without user awareness or consent.

For instance, an organization utilizes an automated résumé filter that deprioritizes candidates from historically Black colleges due to historic hiring patterns in the training set. The employer is unaware of the bias, but the system's disparate impact on ethnic minorities could constitute indirect discrimination and unlawful processing of racial data.

Hidden Bias in AI Training Sets

Many training datasets contain biased historical decisions, even if sensitive attributes were excluded. Techniques such as fairness-aware learning, counterfactual evaluation, and sensitive attribute testing are crucial for identifying and mitigating this bias.

Legal Basis and Consent for Sensitive Data

As discussed earlier in this chapter, GDPR Article 9 forbids the processing of special category data by default. Controllers must identify one of several narrow exceptions to proceed legally.

These exceptions require careful justification and documentation. Unlike "ordinary" personal data (which may be processed under legitimate interest, performance of contract, and so forth), sensitive data demands a stronger foundation and higher accountability.

Explicit Consent: A High Bar

GDPR is clear on consent for the collection and use of sensitive data. Explicit consent must be:

- Freely given
- Specific

- Informed
- Unambiguous
- Affirmative (for example, no pre-checked boxes)
- Clearly documented

In AI, explicit consent is difficult to obtain and even more challenging to maintain due to:

- Indirect or inferential data collection
- Changes in model purpose over time
- Complexity of explanations required to make consent meaningful

For example, a wellness app asks for user consent to analyze heart rate and sleep data. The consent is buried in a lengthy privacy policy and doesn't explain downstream uses (for example, licensing data to third parties for emotion detection R&D). This practice does not meet the standard of explicit, informed consent for processing health data.

Consent vs. Contract in AI Systems

Controllers often confuse the performance of a contract (GDPR Article 6(1)(b)) with consent (Article 9(2)(a)). For sensitive data:

- Consent must be explicit and revocable
- Contract is not sufficient unless the sensitive data is *strictly necessary* to fulfill the contract
- Convenience or "nice-to-have" features do not qualify

Always assess whether the feature requiring sensitive data is essential, or merely optional and consent-dependent.

When Consent Is Not Enough

Even with explicit consent, processing sensitive data may be illegal or unethical if:

- The data subject is in a power-imbalanced relationship (for example, employer-employee)
- Consent is not freely given (for example, gating core services behind consent)
- The AI system creates significant effects that trigger Article 22 protections

In high-risk AI systems, regulators may require additional safeguards, such as DPIAs, transparency reports, and ongoing reviews.

Technical and Organizational Safeguards

Systems processing sensitive data without robust protections increase the likelihood of:

- Unauthorized access or breaches
- Identity theft
- Discrimination or bias
- Regulatory penalties
- Reputational damage

Controllers must implement technical and organizational measures (TOMs) that go beyond basic data protection controls. For sensitive data, these must reflect the heightened risks associated with it.

Core Technical Measures

The principal measures used to protect sensitive data include:

- Encryption at rest and in transit (especially for biometric and health data)
- Access controls and role-based permissions
- Anonymization or pseudonymization where feasible
- Data minimization and field-level masking
- Secure model training environments with audit logging
- Model explainability tools to detect discrimination

For example, a voice biometrics provider applies differential privacy to its training pipeline, ensuring that no single voiceprint can be reconstructed from the model's outputs. They also store voiceprints in an encrypted, access-controlled vault and rotate cryptographic keys every 90 days. These measures reduce the impact of potential compromise.

Federated Learning to Protect Health Data

Federated learning enables decentralized model training; personal data remains on users' devices, and only anonymized model updates are shared. This approach:

- Minimizes data centralization risks
- Helps meet data localization requirements
- Supports privacy in sensitive environments (for example, hospitals, mental health apps)

Caveat: Federated systems still require secure aggregation, validation, and rollback safeguards.

Organizational Safeguards

The principal processes and functions used to protect sensitive data include:

- DPIAs for all sensitive data processing
- Clear retention and deletion policies
- Training for engineers and data scientists on sensitivity and risk
- Review boards or ethics committees for high-risk AI
- Supplier vetting (for example, ensuring third-party biometric vendors meet GDPR standards)
- Formal incident and breach response procedures and training
- Formal risk management and periodic risk assessments

Organizations should also maintain audit trails of sensitive data processing decisions, including the rationale for legal bases, risk mitigation plans, and system performance monitoring.

Designing for Data Minimization

Organizations tend to over-collect personal data "just in case." For sensitive categories, this is dangerous and likely unlawful.

Best practices for data minimization include:

- Using synthetic data for early testing
- Feature reduction to exclude sensitive attributes not needed for model accuracy
- Conducting proportionality assessments to determine whether the model's purpose justifies the sensitivity level of inputs
- Performing periodic dataset audits to detect and remove unnecessary or sensitive data, including data that may be inferred during processing

Special Considerations in AI Applications

The collection and processing of sensitive data requires additional considerations.

Bias and Fairness in Sensitive Attributes

Sensitive data types often correlate with historically marginalized or protected groups. When AI models use or infer attributes such as race, religion, gender, or health status—even indirectly—they may replicate or exacerbate existing discrimination. Common pitfalls include:

- Training data bias (for example, historical hiring decisions that exclude certain groups)
- Proxy discrimination (for example, Postal codes standing in for race)

- Overfitting to sensitive features that skew predictions
- Unjustified exclusion of sensitive groups from model scope

Fairness assessments must include quantitative evaluations of how models perform across demographic groups and qualitative reviews of social context.

An example loan-approval AI system trained on historical lending data exhibits a significantly higher rejection rate for applicants inferred to be from specific minority backgrounds, even though race was not an input variable. The correlation was encoded in postal code and income features, triggering concerns over indirect discrimination.

Audit Questions for Sensitive Attribute Fairness

Here are some questions that can be used to ensure an AI model is free of bias:

- Is any sensitive attribute explicitly used in the model?
- Can non-sensitive features serve as proxies?
- How does model accuracy vary across protected groups?
- Is any demographic group underrepresented in the training data?
- Are fairness constraints applied and documented?
- Does the model's fairness hold over time as new data is introduced?

Regular audits, especially after model updates, are essential for AI involving sensitive categories.

Inference of Sensitive Data from Non-sensitive Inputs

AI systems are capable of *inferential privacy violations*, where models deduce sensitive characteristics from unrelated data. Examples include:

- Religious affiliation from language patterns
- Sexual orientation from social graph connections
- Health conditions from wearable sensor data
- Political leaning from browsing history

These inferences can trigger GDPR Article 9 protections even if the attribute was not explicitly collected.

An example dating app uses an LLM to suggest matches based on chat history and interaction patterns. Over time, it begins to infer users' mental health states and emotional volatility. No mental health data was collected; however, inferences could result in discriminatory outcomes or profile manipulation, potentially violating sensitive data rules.

Sensitive by Proxy: What to Watch For

When training and using an AI system, organizations should be on the lookout for:

- High correlation between features and protected categories
- Models trained on unbalanced or biased-open datasets
- Unlabeled sensitive attributes inferred during testing or deployment
- Vendors offering "demographic enrichment" services
- Post-deployment drift that changes model behavior or fairness across demographic groups

Controllers are responsible for downstream inferences—even when not explicitly designed.

Purpose Creep and Secondary Use

Sensitive data collected for one purpose (for example, health tracking) must not be reused for a new, incompatible purpose (for example, behavior prediction for insurance pricing) unless:

- The data subject explicitly consents
- The reuse falls under a clear legal exception
- The new purpose is demonstrably compatible with the original

AI systems often evolve rapidly, and organizations tend to use existing AI systems for new purposes, making purpose creep an ongoing risk.

The same company later repurposes an emotion-recognition model developed for gaming analytics to evaluate job candidates during interviews. The original dataset contains facial expressions captured in non-professional contexts, raising questions of purpose limitation and fairness.

DPIAs and Risk Management

DPIAs are mandatory under GDPR Article 35 when processing is "likely to result in a high risk to the rights and freedoms of natural persons." This is *almost always the case* with AI systems that process special categories of data.

Triggers of high risk include:

- Systematic monitoring of sensitive data
- Automated decision-making with legal or similar significant effects

- Processing of vulnerable populations
- Use of biometric or health data at scale

Controllers must document and justify their decision, whether they conduct a DPIA or not.

What to Include in a DPIA for Sensitive AI

A robust DPIA for AI involving sensitive data should include:

- Description of the AI system, including inputs, logic, and outputs
- Purpose of processing and justification for data sensitivity
- Risk assessment of potential harm to individuals
- Evaluation of the likelihood and severity of risks
- Safeguards, technical measures, and compliance controls
- Stakeholder input, including from the DPO, legal, and affected communities

For instance, a university introduces a voice-based emotion detection AI to monitor students' mental well-being. A DPIA identifies risks of misclassification, data misuse, and stigmatization. Mitigation includes opting in, local processing, deletion of raw data, and robust student communication.

Red-flag Questions for Sensitive AI Projects

- Does the system infer sensitive attributes?
- Are individuals aware that their sensitive data is being used?
- Can users opt out without losing access to core services?
- Are there downstream partners or vendors using the same data?
- Have the risks been reviewed by legal, ethics, and security teams?

If the answer to any of these is unclear or negative, a DPIA—and potentially regulator consultation—is required.

The Role of Risk Registers and Review Boards

High-risk AI systems involving sensitive data should be cataloged in an internal AI risk register, including:

- System description and owners
- Risk severity level based on impact and likelihood
- Dates of DPIAs and reassessments
- Regulatory exposure

- Incident history (if any)
- Mitigation status
- Links to supporting documentation for audit readiness

Some organizations also establish AI review boards or ethics panels to assess especially sensitive or controversial deployments, a governance best practice that supports both compliance and public trust.

Data Subject Rights and Sensitive Data

AI processing of sensitive data increases the importance and complexity of data subject rights, requiring stronger transparency, opt-out options, and safeguards against automated decisions.

Enhanced Rights Expectations

When AI systems process special categories of data, data subjects are entitled to heightened transparency, control, and redress. All core rights under data protection laws remain in force, but their application can become more complex and critical.

Key data subject rights should include:

- **Right to be informed:** Why is sensitive data being processed? How was it obtained?
- **Right of access:** What sensitive data is being held, and how is it being used?
- **Right to rectification:** Can users correct inferred or misclassified sensitive attributes?
- **Right to erasure:** Can users demand the deletion of biometric or health records? (There are specific challenges related to the removal of individual training records from an AI system that are discussed throughout this book.)
- **Right to object:** Can users opt out of profiling or automated processing based on sensitive data?
- **Right not to be subject to automated decisions:** With legal or similar significant effects (GDPR Article 22, CPRA 1798.121).

Sensitive Data Adds Risk to AI Rights Requests

Processing sensitive data often makes data subject requests (DSRs) more urgent, more resource-intensive, and more likely to lead to complaints if mishandled.

Challenges include:

- Providing intelligible explanations when decisions are made by opaque models
- Correcting inferred sensitive data that users never explicitly provided
- Retaining model performance while deleting training data tied to a specific user who wants to opt out
- Navigating conflicts between erasure requests and regulatory, contractual, or research retention requirements

For instance, a job seeker requests the deletion of data used by an AI hiring tool after being rejected. The company must determine whether the data is tied to model weights, whether retraining is feasible, and whether deleting the data would impact compliance with anti-discrimination documentation requirements.

How to Respond to a DSR Involving Sensitive Data

- Confirm identity with heightened caution (especially for biometric data)
- Provide a clear explanation of whether data was collected or inferred
- Identify any decisions or consequences linked to the data
- Explain the legal basis for processing and any third parties involved
- Log all actions taken and timeframes met

Controllers should prepare standard operating procedures (SOPs) for rights requests involving sensitive categories of information.

Opt-out Rights and Limitations in AI Contexts

Under both the GDPR and laws like the CPRA, individuals have the right to opt out of certain types of data processing, particularly when it involves profiling, targeted advertising, or automated decision-making.

In AI systems, opt-out mechanisms become both more important and more complex. Users must be able to decline processing for secondary purposes, object to profiling, or refuse to be subject to automated decisions, especially when such decisions may significantly impact them (for example, in hiring, lending, or healthcare contexts).

A compliant opt-out mechanism should be:

- Clearly presented and easy to exercise
- Granular, allowing individuals to decline specific AI-related uses (for example, targeted ads, automated screening)
- Effective, meaning the organization must act on the request and confirm suppression of the processing

For example, a job board uses an AI model to rank applicants based on inferred cultural fit. After opting out, one candidate is still scored by the system and filtered out, which constitutes a violation of the right to object under GDPR Article 21 and may require regulatory reporting if unresolved.

Opt-out rights are not absolute. Under GDPR, they may be overridden if:

- Processing is necessary for a contract or legal obligation
- There is a compelling, legitimate interest that overrides the individual's rights

However, when profiling is used for direct marketing, the right to opt out is *absolute* and must be honored without exception.

Controllers must document and enforce opt-out choices across all systems and vendors, including those related to model training and profiling contexts.

Rights in High-risk Automated Decision-making

When AI systems process sensitive data and significantly affect individuals (for example, rejection from services, employment, or credit), automated decision-making restrictions apply.

Under GDPR Article 22:

- Individuals have the right not to be subject to solely automated decisions.
- Exceptions (for example, explicit consent or contractual necessity) must include safeguards such as:
  - Meaningful human involvement
  - Right to contest the decision
  - Transparent explanation of logic and consequences

AI systems that use sensitive data to drive eligibility or pricing decisions (for example, health-based insurance premiums, biometric access denial) are especially likely to fall under these protections.

Summary

AI systems present unique challenges to longstanding data protection principles. As organizations leverage AI to analyze, predict, and make decisions about individuals, foundational privacy concepts like notice, choice, consent, and purpose limitation must be reinterpreted in ways that address the complexity and opacity of modern algorithms. Individuals must be informed about how their data is used, particularly when automated decision-making is involved. Yet, traditional privacy notices are often ill-equipped to explain the logic or implications of machine learning. Similarly, consent must be meaningful, informed, and explicit—a tall order when AI systems operate behind the scenes or rely on inferred data. Purpose limitation is frequently tested when models trained for one function are later repurposed for others, raising the risk of unauthorized secondary use. To mitigate these risks, data minimization and privacy-by-design principles must be integrated throughout the AI lifecycle, supported by techniques such as federated learning, differential privacy, and modular model architecture.

In parallel, AI controllers (those who determine the purposes and means of processing) bear extensive obligations under laws such as the GDPR. They must conduct DPIAs for high-risk processing, primarily when automated decisions affect people's rights. When engaging third-party processors, controllers must establish robust contracts and ensure

ongoing oversight to avoid "processor drift." Many AI ecosystems are global, triggering rules on cross-border data transfers and necessitating mechanisms such as Standard Contractual Clauses or Transfer Impact Assessments. Controllers are also responsible for honoring data subject rights, including access, deletion, and objection to profiling, even when models are complex and opaque. Incident management becomes more nuanced with the use of AI, especially when breaches involve model inversion, membership inference, or sensitive training data. To meet accountability requirements, organizations must maintain detailed documentation, including model cards, data maps, risk registers, and audit logs.

A critical area of focus is the use of sensitive or special categories of data in AI, including biometrics, health data, political opinions, religious beliefs, and racial or union affiliations. These data types are subject to elevated protection due to their potential to cause harm, facilitate discrimination, or enable surveillance. Controllers must identify a lawful basis for processing, often explicit consent, and apply additional safeguards. AI systems frequently infer sensitive attributes even when they weren't explicitly collected, a risk that complicates compliance and raises ethical concerns. DPIAs are often mandatory in these scenarios, and safeguards must go beyond encryption and access controls to include bias mitigation, fairness testing, and strong user transparency. Controllers must also ensure that individuals can exercise their rights concerning sensitive data, especially when automated decisions could significantly affect their lives. As AI capabilities grow, so too does the responsibility to uphold the dignity, autonomy, and privacy of those whose data powers these systems.

Exam Essentials

Processing of sensitive data is prohibited by default under GDPR unless a specific legal exception applies, such as explicit consent or substantial public interest. Special category data—including biometrics, health, religion, and political beliefs—cannot be processed unless one of a narrow set of exceptions under GDPR Article 9(2) applies. Explicit consent must be freely given, specific, informed, and documented. Where obtaining consent is not feasible, controllers must justify their processing under another recognized legal basis and apply additional safeguards.

Data controllers are accountable for the entire AI lifecycle, including DPIAs, lawful basis, data subject rights, vendor oversight, and compliance with cross-border transfer regulations. Controllers must not only ensure compliance but demonstrate it, including conducting DPIAs for high-risk AI systems, maintaining lawful processing grounds, fulfilling rights requests, managing processor relationships under GDPR Article 28, and using mechanisms like SCCs or TIAs for international transfers. These responsibilities cannot be delegated or ignored in AI contexts.

Purpose limitation, data minimization, and privacy by design are foundational to trustworthy AI governance and must be embedded from the outset. AI systems frequently challenge these principles due to their dynamic nature and appetite for data. Still, data should only

be collected for defined, lawful purposes and minimized to what is necessary. Privacy by design requires integrating privacy controls throughout development, not as an afterthought, including feature selection, model explainability, and user opt-outs.

AI systems can infer sensitive attributes (for example, race, religion, health) even without explicit collection, which still triggers legal obligations. Inferred sensitive data is subject to the same protections as explicitly collected data. Controllers must assess whether their systems are generating protected characteristics (intentionally or not) and apply appropriate legal bases and safeguards. Failing to do so risks unlawful profiling, discrimination, and regulatory action.

Automated decisions involving special category data may be restricted or require meaningful human involvement and additional safeguards. Under GDPR Article 22, individuals have the right not to be subject to decisions based solely on automated processing that produce legal or similarly significant effects. If such processing involves special category data, it is even more tightly restricted. Where permitted, these systems must provide explainability, human review, and redress mechanisms for affected individuals.

Review Questions

The following questions are designed to test your understanding of this chapter's material. For more information on how to obtain additional questions, see this book's Introduction. You can find the answers in Appendix A.

1. Under the GDPR, which legal basis is most often required for processing special category (sensitive) data?
   - **A.** Explicit consent
   - **B.** Performance of a contract
   - **C.** Legitimate consent
   - **D.** Implied consent
2. Which of the following best describes *purpose limitation* in the context of AI?
   - **A.** Collecting data from as many sources as possible
   - **B.** Allowing reuse of data for any purpose deemed profitable
   - **C.** Ensuring data is collected for a specific, explicit purpose and not reused incompatibly
   - **D.** Using third-party datasets without reviewing their original purpose
3. A Data Protection Impact Assessment (DPIA) is required when:
   - **A.** Any data is used for marketing
   - **B.** A low-risk AI system is deployed
   - **C.** A system is used internally by employees only
   - **D.** Processing involves special category data and could pose high risks to rights and freedoms
4. What role does a data controller play in the AI lifecycle?
   - **A.** Manages the physical server infrastructure
   - **B.** Develops algorithms for all third parties
   - **C.** Determines the purpose and means of data processing
   - **D.** Handles data breaches only
5. Which of the following would likely require heightened safeguards under privacy law?
   - **A.** User preferences on website themes
   - **B.** Facial recognition used for retail surveillance
   - **C.** Email addresses used for newsletter signups
   - **D.** A/B testing of homepage layout

6. What is a common risk of AI systems inferring sensitive data?
   - **A.** Low training cost
   - **B.** Reduced model explainability
   - **C.** Triggering legal obligations even if the data was not explicitly collected
   - **D.** Inference accuracy cannot be evaluated

7. What is a key element of privacy by design in AI systems?
   - **A.** Prioritizing model accuracy over user privacy
   - **B.** Embedding privacy protections throughout the system lifecycle
   - **C.** Outsourcing compliance to vendors
   - **D.** Avoiding DPIAs unless requested by a regulator

8. Which scenario best reflects the principle of data minimization?
   - **A.** Collecting all available user data for future analysis
   - **B.** Using anonymized transaction data limited to the current model's purpose
   - **C.** Archiving full user profiles for at least ten years
   - **D.** Combining datasets from unrelated apps to improve personalization

9. Which obligation must a controller meet when using third-party AI processors?
   - **A.** Sign a DPA with required contractual terms under Article 28 GDPR.
   - **B.** Include model logic in the user privacy notice.
   - **C.** Allow unlimited use of training data by the vendor.
   - **D.** Accept vendor liability for privacy compliance.

10. Under GDPR Article 22, what must be provided when a decision based solely on automated processing significantly affects an individual?
    - **A.** Access to the source code
    - **B.** Unlimited data portability
    - **C.** Meaningful human review and an option to contest the decision
    - **D.** Permanent deletion of the data



Chapter

5

Sectoral and Civil Laws

THIS CHAPTER COVERS AIGP DOMAIN II, "UNDERSTANDING HOW LAWS, STANDARDS AND FRAMEWORKS APPLY TO AI," SUBDOMAIN II.B, "UNDERSTAND HOW OTHER TYPES OF EXISTING LAWS APPLY TO AI." THE FOLLOWING TOPICS ARE COVERED:

- ✔ Intellectual property laws
- ✔ Non-discrimination laws
- ✔ Consumer protection laws
- ✔ Product liability laws

THE AI GOVERNANCE PROFESSIONAL (AIGP) PERFORMANCE INDICATORS IN THIS SUBDOMAIN INCLUDE:

- ✔ Understand how intellectual property laws apply to AI (for example, prohibiting or limiting use of data for AI training).
- ✔ Understand how non-discrimination laws apply to AI (for example, in the employment, credit, lending, housing and insurance contexts).
- ✔ Understand how consumer protection laws apply to AI (for example, prohibiting unfair and deceptive acts or practices).
- ✔ Understand how product liability laws apply to AI (for example, prohibiting design or manufacturing defects).
- ✔ Identify laws that apply to the AI model.



AI systems are increasingly subject to a variety of sector-specific and civil legal obligations, even in the absence of dedicated AI legislation. This chapter explores how existing laws, crafted initially for human actors and traditional technologies, are now being applied to AI development, deployment, and oversight. It examines the impact of intellectual property law on AI training and outputs, the role of anti-discrimination laws in high-stakes decision-making, consumer protection rules that govern transparency and fairness, and product liability standards for AI-enabled physical systems. By understanding these intersecting legal regimes, AI governance professionals can better anticipate compliance risks, support responsible AI system design, and ensure accountability across the AI lifecycle. Although these laws were not written with AI in mind, they are being amended and enforced in ways that have potentially far-reaching implications for developers, deployers, and users alike.

Intellectual Property Laws and AI

As AI systems become more capable of ingesting, interpreting, and replicating human expression, questions surrounding *intellectual property (IP)* have taken center stage. At the heart of the debate is a fundamental tension: AI systems require large volumes of data to learn effectively, yet much of that data may be protected by copyright, patent, trade secret, or other IP rights. This section explores how intellectual property laws apply to AI, particularly in the context of data ingestion and model training, and provides practical guidance for navigating the rapidly evolving legal landscape.

Foundations of Intellectual Property and AI

The legal regime of intellectual property encompasses a variety of protections intended to incentivize creativity, innovation, and disclosure. As AI systems increasingly ingest IP-protected content, existing legal protection doctrines are being tested in new ways. This section describes how traditional IP categories—copyright, patents, trade secrets, and database rights—apply to AI development and use.

Copyright Law: The Center of AI Training Controversies

Copyright law protects "original works of authorship" fixed in a tangible medium, including text, music, images, video, and software code. For generative AI systems that rely on scraping publicly available content from the web, this presents a legal and ethical minefield. Human content creators voice concerns about AI systems that threaten to disintermediate creators, depriving them of their ability to make a living through creative expression.

Key legal questions include:

- Is the use of copyrighted works for AI training a violation of the reproduction right?
- Does training constitute "transformative" use under the fair use doctrine?
- Are AI-generated outputs infringing if they closely resemble copyrighted works?

These questions are being tested in courts globally, often without clear precedent.

In 2023, Getty Images sued Stability AI for allegedly using millions of copyrighted images to train its generative model without permission or compensation. Getty's claim hinged on both copyright and trademark violations, arguing that the use of its images, even for training purposes, was not "transformative" and directly competed with its licensing business. (This case is still pending as of the writing of this book.)

Fair Use 101

The U.S. Copyright Act permits certain unauthorized uses of copyrighted material under the doctrine of *fair use*. Courts weigh four factors:

- Purpose and character of the use (for example, commercial vs. non-profit, transformative vs. duplicative)
- Nature of the copyrighted work
- Amount and substantiality of the portion used
- Effect on the market for the original copyrighted work

AI developers often claim that training is a fair use because models do not copy or replicate exact works but learn from analyzing patterns and relationships. However, this argument is being increasingly challenged in court.

Patent Law and AI-generated Inventions

Patent law protects new, useful, and non-obvious inventions. Inventors who want to monetize their inventions may choose to publish their invention's design with the U.S. Patent Office, which then grants the inventor exclusive rights to sell or license the invention

for a specific number of years (other countries operate similarly). While the focus of AI/IP discourse is typically on copyright and data, patent issues arise in two ways:

- Use of patented data or algorithms in training
- Patentability of AI-generated inventions

The latter has drawn significant attention. Should inventions devised by AI systems (without human inventors) be patentable? Similarly, should the prompts be used to create otherwise patentable works that are themselves patentable or protected by copyright?

The AI system DABUS was named as the sole inventor on patent applications filed in over a dozen jurisdictions. While South Africa granted the patent, most patent offices (including the USPTO and EPO) rejected the application, holding that only a natural person can be an inventor under current law.

Can AI Be an Inventor?

Globally, courts and patent offices have rejected the idea that AI can qualify as an inventor. The U.S. Federal Circuit held in Thaler v. Vidal (2022) that the Patent Act requires a human inventor, effectively shutting down DABUS-style applications for now.

Trade Secrets and AI Models

A *trade secret* is a form of intellectual property, *typically the design, methods, formulas, data, and know-how* of a product or process, that its owner chooses to conceal from the public, and whose secrecy provides economic value. Trade secret law protects confidential business information that derives economic value from not being generally known. In the context of AI, examples of trade secrets include:

- Proprietary training datasets
- Model architecture and hyperparameters
- Fine-tuning methodologies
- Inference optimization algorithms
- Data preprocessing and augmentation techniques
- Evaluation metrics and benchmarks
- Deployment infrastructure and scaling methods
- Data augmentation techniques

Because copyrights do not protect facts or ideas, and patents require public disclosure, many companies rely on trade secret protection for competitive AI advantages.

Challenges concerning the protection of trade secrets include:

- Ensuring the secrecy of datasets in collaborative environments
- Preventing reverse-engineering of outputs that may reveal training data and/or algorithms
- Balancing transparency for ethical review with trade secret preservation
- Supply chain exposure from third-party service providers handling sensitive datasets or models

While earlier versions of OpenAI's GPT were published openly, later versions like GPT-4 became proprietary. OpenAI cited safety and competitiveness as reasons for withholding training data and model details. This trade secret strategy also shields the company from claims (to an extent) by organizations that suspect IP infringement.

Database Rights and Regional Variability

In the European Union and select jurisdictions, *sui generis database rights* protect the substantial investment made in compiling data. These rights can restrict unauthorized extraction and reuse of data, even if the individual data points aren't themselves protected by copyright.

This becomes relevant when AI models are trained on structured datasets such as medical records, scientific databases, or financial archives. In the European Union and other places, the owners of these databases can forbid their use in AI system training.

The EU Database Directive

Under the EU's Database Directive (Directive 96/9/EC), rights holders can prohibit the extraction or reuse of substantial parts of a database. AI developers may need to obtain licenses, even if the content is non-copyrightable.

Use of Protected Works in AI Training

While foundational IP laws establish the baseline for what can and cannot be used, their application to AI training practices is not straightforward. This section dives into how IP law constrains (or permits) the use of protected works in AI training datasets, with a focus on current litigation, licensing approaches, and developer responsibilities.

The Training Data Dilemma

Training AI systems often requires massive datasets, typically scraped from websites, forums, news articles, art repositories, and codebases. Many of these sources contain copyrighted content or content governed by restrictive licenses.

For example, GitHub Copilot, originally powered by OpenAI's Codex, suggests code completions based on code from public GitHub repositories. In 2022, a class-action lawsuit alleged that Copilot reproduced open-source code without proper attribution, in violation

of open-source licenses such as the GNU General Public License (GPL). This lawsuit raised difficult questions:

- Does training on open-source code require compliance with its license terms?
- Is code completion a derivative work?
- Should developers be liable for using Copilot-suggested code?

Licensing Models for Training Data

To avoid IP disputes, some AI developers are turning to explicitly licensed or synthetic data sources. Licensing approaches include:

- Public domain content (for example, U.S. government data, Project Gutenberg)
- Creative Commons (CC) licenses, which may allow reuse with conditions
- Paid licensing agreements (for example, Shutterstock licensing its image library for training)
- Synthetic data generation, which reduces dependency on real-world datasets

Best Practice: Data Provenance Logs

Maintaining a record of data sources and license terms is increasingly essential. AI developers are adopting *data provenance logs* to track the origin, usage rights, and attribution requirements for each dataset, enabling a defensible posture in the event of IP claims. Data provenance is discussed in detail in Chapter 9.

AI Outputs and the Question of Infringement

While much of the IP conversation around AI has focused on training data, legal attention is increasingly shifting to what AI systems produce. As generative models create text, images, videos, code, and music, the line between inspiration and infringement becomes difficult to draw. This section explores whether AI-generated outputs can infringe existing IP, and who bears responsibility when they do.

When Is an AI Output Infringing?

Whether an AI output is infringing typically hinges on two factors. First, is the output similar to protected work (and to what extent), and second, what access to the original did the AI system have during training?

If a generative model reproduces copyrighted material, whether a verbatim paragraph, a near-identical photo, or a melody indistinguishable from a known song, it may be infringing, even if the developer did not intend it.

Tools like Suno and Udio allow users to generate music in various genres. Critics worry that outputs could mimic existing songs, leading to infringement similar to the *Blurred Lines* lawsuit, in which Marvin Gaye's family and Bridgeport Music argued that the song infringed on Gaye's *Got to Give It Up*. In that case, a jury found copyright infringement based on *feel* and *style* (not direct copying), raising concerns for AI music outputs that echo iconic tracks.

In another example, Denmark has proposed legislation to grant its citizens copyright protection of their face, voice, and body, as a potential legal remedy regarding the creation of *deepfakes*.

Output Attribution and Legal Accountability

A thornier issue is determining *who is liable* if an AI-generated output infringes IP rights. Options include:

- The developer of the model
- The entity that trained it on infringing data
- The end user who prompted and used the output
- The model itself (though not recognized legally as an agent or person in most jurisdictions)

Liability tends to fall on human or corporate parties based on theories of vicarious infringement or contributory liability.

Key Case: Andersen v. Stability AI (2023)

In this lawsuit, a group of artists alleged that AI image models like Stable Diffusion generated derivative works mimicking their art style. Plaintiffs argued that merely prompting "in the style of [artist name]" produced infringing content. The court has yet to rule decisively, but the case exemplifies how output attribution and artistic identity intersect.

Ownership and Protection of AI-generated Content

Beyond infringement risks, organizations must also determine whether and how AI-generated outputs can be protected under existing IP law. This includes questions about copyrightability, authorship, and the ability to commercialize outputs. There is not a great deal of legal precedent in this regard.

Copyrighting AI-generated Works

Under U.S. and most international copyright laws, only works created by human authors are eligible for protection. The U.S. Copyright Office has reaffirmed this principle in multiple guidance statements and court cases.

In 2023, the U.S. Copyright Office revoked protection for images in the graphic novel *Zarya of the Dawn* created with the AI tool Midjourney, reasoning that those images lacked human authorship. The textual narrative, written by a human, retained protection, but the images did not. This case illustrated the Copyright Office's "human authorship" policy.

Joint Authorship and AI as a Tool

Suppose AI is used as a *tool* rather than a creative agent. In such a case, courts may still grant protection, provided a human exercises sufficient control and creative judgment, leading some to distinguish between:

- **Tool-based use:** Human guides and edits the output (copyright eligible)
- **Autonomous generation:** Model creates with minimal human input (not eligible)

In commercial contexts, this distinction matters when registering copyrights, licensing AI-generated media, or asserting exclusivity over branded content.

Risk Mitigation and Compliance Strategies

Given the uncertainty regarding the evolving risks and uncertainties surrounding IP and AI, organizations must adopt proactive compliance strategies. This section includes key governance tactics to reduce exposure and establish defensible practices.

Establishing an IP-aware AI Governance Framework

The corporate governance framework(s) that address intellectual property should include:

- Due diligence for training data sources, to ensure the organization will not be subject to legal actions by data source owners
- Review of licensing obligations (primarily open-source or Creative Commons) for training data
- Controls over model output usage that may require reviews of outputs for legal sufficiency
- Recordkeeping for data provenance and model lineage
- Policies to codify the business rules for the above

These activities should be embedded in AI lifecycle processes, particularly model design and development, deployment, and audit stages.

Sample Checklist: AI IP Risk Review

- Were all training datasets legally acquired or licensed?
- Were open-source licenses respected (for example, GPL, MIT, CC-BY)?

- Does the model produce content that resembles known copyrighted works?
- Are users warned against the improper use of outputs?
- Is the organization prepared to respond to takedown or litigation?
- Does the organization have defensible policies and practices?

Embedding Contractual Protections

Because many AI systems are created by one entity and used by another, legal agreements between these entities need to address IP risks through the inclusion of clauses such as:

- Clear allocation of liability between vendor and customer
- Indemnification clauses for third-party IP claims
- Usage restrictions (for example, no use of inputs and/or outputs for commercial media)
- Representations and warranties regarding training data sources

These provisions are becoming common in AI-as-a-Service (AIaaS) contracts, especially for generative tools that output creative content.

In an example of contractual protections, Adobe's Firefly generative model explicitly disclaims use of copyrighted third-party data and limits user rights to commercialize outputs. Adobe provides limited indemnity, a trend that signals the increasing role of contracts in managing IP risk.

Investing in Explainability and Transparency

Models that can explain their outputs or trace their data lineage offer greater defensibility in IP-related legal disputes. Emerging solutions include *explainable AI (XAI)* methods, synthetic data auditing tools, and model watermarking or hashing to track data usage. Though not foolproof, these tools help organizations demonstrate diligence and may reduce the risk of "black box" infringement.

In the Author's Opinion

AI systems lacking explainability may find limited use in commercial environments where intellectual property is in play, whether IP is used in inputs or created as outputs. The claim of "we don't know how the AI created this" will likely not succeed as a part of a legal defense. Legal precedents may sideline opaque AI systems as a result.

Cross-border IP Conflicts and Jurisdictional Complexities

AI systems often operate across national borders: designed in one country, trained in another, deployed in another, and accessed globally. Intellectual property rights laws, however, remain

TABLE 5.1 Differences in IP Laws in Selected Jurisdictions

Region	Key Differences
United States	Fair use doctrine allows limited unauthorized use. No copyright granted in works generated solely by AI without human authorship.
European Union	More limited exceptions to copyright. Recognition of database rights. Moral rights stronger. Text and data mining exceptions exist but may require opt-outs by rights holders.
China	Explicit recognition of copyright in AI-generated content (under human guidance). State involvement in model compliance.
Canada/Australia	Approaching issues on a case-by-case basis. Some are open to broader exceptions for training.
Japan	Permits broad text and data mining for any purpose, including commercial use, provided it does not infringe privacy or security. No explicit recognition of AI authorship yet.

territorial: they are granted and enforced by individual jurisdictions, creating significant compliance and litigation challenges for multinational AI operations.

Varying Legal Standards Across Jurisdictions

Laws regarding what constitutes infringement, who may hold copyright, and how fair use or exceptions apply differ widely. Table 5.1 provides a sample.

AI system designers and developers must account for these differences not just in model deployment but in dataset acquisition, model access, and output distribution.

In response to legal threats in the European Union, some generative AI tools have implemented *geofencing* (to limit user access from specific countries and regions) and excluded certain datasets from region-specific deployments. This practice reflects the need to tailor AI services to comply with local legal standards.

Jurisdictional Enforcement and Forum Shopping

When disputes arise, such as an artist alleging their work was used in training, plaintiffs may research favorable jurisdictions, leading to *forum shopping*, where discovery processes and class actions offer strategic advantages.

Multinational organizations (which reasonably include organizations operating in a single country with users in many countries) must prepare for:

- Multi-jurisdictional IP claims.
- Conflicts of law: For example, where is the infringement deemed to have occurred?

- Extraterritorial enforcement: A liability risk even without physical presence if the AI system or its outputs are offered in a given market.
- Compelled data disclosures during discovery or regulatory review.

These legal challenges may serve to raise the economic barrier of entry for new AI vendors, based on the need for legal counsel with global intellectual property experience.

International Treaties on Intellectual Property

Treaties such as the Berne Convention of 1886 and the TRIPS Agreement of 1994 provide baseline IP protections across over 170 countries. While they harmonize some IP rights, they do not resolve core AI-related questions, such as:

- Is training on copyrighted data "fair use," "reproduction," or something else?
- Can machine-generated works be copyrighted?
- How do collective management organizations (CMOs) license training rights?

The gaps related to inputs to and outputs from AI systems call for new global instruments addressing AI-specific IP issues.

The Future of AI Regulations

As IP-related litigation grows, regulators are beginning to consider whether existing laws are sufficient or if new frameworks are needed to govern AI (this ship has sailed, in the author's opinion). This section explores early signals of legislative movement.

Proposed U.S. Legislation

Several members of the U.S. Congress have floated proposals to:

- Require AI developers to document data sources
- Prohibit unauthorized use (which, in some cases, must still be defined) of copyrighted content for training
- Establish registries or licensing frameworks for training datasets
- Block U.S. states from enacting their own AI regulations, leading to the kind of confusion experienced with many states' privacy laws

While no comprehensive federal law has been passed, increased scrutiny from the Copyright Office, FTC, and courts may drive future policymaking.

AI-related executive orders issued by the U.S. President have been focused on innovation and continued U.S. leadership in AI technology, but none so far have addressed intellectual property rights.

What to Watch: Future IP Cases That Could Reshape AI Law

- Disney & Universal v. Midjourney
- Getty Images v. Stability AI (United States and United Kingdom)
- Andersen v. Stability AI (United States)
- GitHub Copilot litigation (United States)
- European Parliament deliberations on AI-generated content disclosure

These cases could establish precedents on whether training on copyrighted data constitutes infringement, who is liable for outputs, and what disclosure obligations developers face.

European Union Initiatives

While the EU AI Act (covered in Chapter 6) does not directly resolve IP issues, it does introduce transparency obligations, such as documenting sources of training data and disclosing whether copyrighted training data was used.

Separately, the *EU Digital Services Act (DSA)* and *EU Copyright Directive* may influence AI platforms that distribute content. These regulations could require takedown mechanisms for infringing outputs, impose liability on platforms hosting generative tools, and encourage licensing frameworks for machine learning purposes. These would provide needed clarity, and hopefully not overly restrict AI system creativity and innovation.

Industry Standards and Self-regulation

In the absence of clear legal mandates, the AI industry is beginning to develop its own norms. Voluntary initiatives include:

- Model cards that disclose data sources and limitations
- *Data nutrition labels* outlining content origins and data provenance
- AI licensing marketplaces (for example, LAION's [Large-scale Artificial Intelligence Open Network, a German non-profit] open data models, Shutterstock's AI licensing portal)

Organizations that follow emerging best practices, especially on consent, attribution, and transparency, are more likely to withstand future legal scrutiny. Optimists believe that any industry that sufficiently and effectively polices itself may avoid harsh regulations, at least for a time.

IP Law as a Central Governance Challenge

Intellectual property law is no longer a niche concern for AI developers, but instead, it is a central governance challenge. As AI systems rely on massive, often uncurated datasets to

generate powerful outputs, the risks of infringement, ownership disputes, and cross-border litigation are growing. To get ahead of the curve, organizations building or deploying AI should:

- Document and vet training data sources carefully
- Understand global IP regulations and obligations
- Include IP risk protection language in legal agreements
- Monitor litigation and regulatory developments
- Include these activities in policies and practices required by AI governance

Doing so not only limits legal exposure but also fosters trust with users, creators, and regulators alike.

Generative AI and the Copyright Office's Position

The U.S. Copyright Office has issued multiple statements clarifying that:

- Works created solely by non-human authors are not protected by copyright law
- Applicants must disclose AI involvement in copyright registrations
- Omission of such disclosure may invalidate the registration

A 2023 policy update formalized these rules, urging creators to distinguish human and AI contributions in mixed-authorship works, a signal that IP registration practices will become more rigorous in the AI era.

Non-discrimination Laws and AI

As AI goes mainstream, organizations are employing AI systems in high-stakes decision-making: who gets a job interview, who is terminated in a corporate layoff, what interest rate a borrower receives, whether an insurance policy is approved, whether medical insurance will cover treatment, or which housing listings an applicant qualifies for. In each of these domains, longstanding civil rights and anti-discrimination laws impose legal obligations to ensure fairness and equal treatment. When AI is involved, the risk of perpetuating or exacerbating bias becomes a matter of legal and ethical compliance. This section explores how non-discrimination laws apply to AI systems, particularly in employment, credit, housing, lending, and insurance, and offers guidance on risk mitigation and regulatory trends.

Legal Foundations of Anti-discrimination in Algorithmic Systems

Before exploring specific industry sectors, it is important to understand the key legal principles that underpin anti-discrimination obligations in automated systems. These principles generally apply across multiple domains and form the backbone of enforcement actions and litigation.

Protected Classes and Prohibited Conduct

U.S. civil rights laws prohibit discrimination based on "protected characteristics," which commonly include these *protected classes*:

- Race
- Color
- National origin
- Sex (including gender identity and sexual orientation)
- Pregnancy
- Income
- Familial status
- Religion
- Disability
- Veteran status
- Genetic traits
- Age

Prohibited conduct typically falls into two categories:

- **Disparate treatment:** Intentional discrimination based on a protected class
- **Disparate impact:** Neutral policies that disproportionately harm protected groups

Disparate Impact 101

A policy or model that is neutral on its face but disproportionately affects a protected group can violate anti-discrimination laws, even if no intent to discriminate is shown. Plaintiffs must demonstrate:

- A significant adverse impact on a protected class
- Lack of business necessity or less discriminatory alternatives

Disparate impact theory is central to many algorithmic bias cases.

Algorithmic Bias as a Legal Risk

Bias in AI systems, whether intentional or not, may arise from:

- Biased training data (for example, historical decisions reflecting systemic discrimination)
- Proxy variables (for example, postal codes correlating with race or religion)
- Opaque models that mask bias
- Feedback loops that reinforce inequality
- Sampling bias (datasets that underrepresent or overrepresent certain groups or contexts)

When these issues manifest in hiring, lending, housing, or insurance, they may violate anti-discrimination statutes and expose organizations to lawsuits or regulatory penalties. When AI systems facilitate decision-making, organizations may find themselves defending them and their design and training.

Employment: AI in Hiring, Promotion, and Termination

AI systems are increasingly used in employment decisions, from résumé screening and video interviews to workforce analytics and performance management. Nevertheless, these tools must comply with labor and civil rights laws, particularly Title VII of the Civil Rights Act, the Americans with Disabilities Act (ADA), and the Age Discrimination in Employment Act (ADEA). Just because an AI system facilitates these processes, whether the AI system is opaque does not absolve any organization from liability when discrimination occurs.

Legal Requirements for Hiring Algorithms

Under Title VII of the Civil Rights Act, it is unlawful for employers to use selection procedures that cause a disparate impact on protected groups unless:

- The procedure is job-related and consistent with business necessity
- No less discriminatory alternative is available

This rule applies to both human- and machine-driven decisions.

In 2018, Amazon developed and later abandoned an internal résumé screening tool that penalized candidates from women's colleges or who used terms like "women's chess club." The model had been trained on résumés from a predominantly male applicant pool, introducing gender bias through historical data and practices. Although never deployed at scale, the tool serves as a cautionary tale for unexamined training data and proxy discrimination.

EEOC Guidance on AI in Employment

In May 2023, the Equal Employment Opportunity Commission (EEOC) issued guidance clarifying that employers are responsible for AI-based decisions, even if third-party vendors develop those AI tools. This guidance is consistent with the general principles of third-party

risk management. If any vendor or service provider's product or service results in an event or incident, the organization that selected the vendor is ultimately responsible for the incident.

Key EEOC Takeaways

- Employers must evaluate tools for disparate impact.
- Vendor disclaimers do not shield liability.
- Reasonable accommodations must be provided under the ADA, even for algorithmic assessments.
- Employers should audit and document algorithmic decision-making processes.

This guidance reflects growing consensus that "algorithmic outsourcing" of bias is not a defense.

Local Laws and Fair Hiring Ordinances

Beyond federal law, cities and states are enacting their own AI-related employment regulations. For instance, New York City Law 144 requires bias audits of automated employment decision tools, annual publication of audit results, and disclosures to candidates of algorithmic use. New York City's law is the first of its kind in the United States, but similar laws are being proposed in California, Illinois, and other jurisdictions. By the time you read this, there will likely be several more, particularly since Congress did not pass the "no state laws on AI" law.

Credit and Lending: Navigating FCRA and ECOA in the Age of AI

Credit decisions are prime territory for AI applications, from loan underwriting and fraud detection to credit scoring and dynamic pricing. However, these activities are tightly regulated in the United States by the Equal Credit Opportunity Act (ECOA) and the Fair Credit Reporting Act (FCRA), both of which apply regardless of whether a human or machine makes the decision.

Organizations considering using AI to improve decision-making in the credit business are arguably optimistic about improvements in their loan portfolios or their net income, and that those improvements will be worth the various types of risks that any IT project (never mind AI) entails.

Equal Credit Opportunity Act (ECOA)

The Equal Credit Opportunity Act (ECOA) of 1974 prohibits discrimination in credit transactions based on protected characteristics, including race, sex, marital status, religion,

national origin, age, or receipt of public assistance. The law applies to credit cards, mortgages, auto loans, small business lending, and personal lines of credit.

When AI systems are used to make or assist in these decisions, creditors must ensure:

- No disparate impact or use of prohibited proxies
- There are clear reasons for adverse actions (AI explainability)
- Fair access across protected groups

In 2022, the Consumer Financial Protection Bureau (CFPB) warned lenders that "black box" AI algorithms do not excuse them from ECOA obligations. Even if the model is opaque, lenders must be able to provide specific reasons for credit denials. This rule is another reason for ensuring that any AI model developed has solid explainability, as laws are likely to require it.

Fair Credit Reporting Act (FCRA)

The U.S. Fair Credit Reporting Act (FCRA) of 1970 governs the use of credit scores and third-party data in making credit decisions. When AI systems ingest or generate such data, they may qualify as consumer reporting agencies (CRAs—also known as credit bureaus) under FCRA.

The FCRA requires:

- Accuracy and integrity of consumer data
- Disclosure to consumers of decisions and data sources
- Mechanisms to dispute errors

The FCRA also applies to employment background checks. In many jurisdictions, a credit report can be used as a part of a background investigation that many organizations use as part of their screening process.

FCRA vs. ECOA

Aspect	FCRA	ECOA
Focus	Consumer data and reports	Discrimination in lending decisions
Triggers	Use of third-party consumer data	Protected class status
Key Requirement	Accurate, disputable reports	Equal access and justifiable criteria
AI Implication	Responsible data sourcing	Bias detection and justification logic

Together, these laws require both fair inputs and fair outputs in credit-related AI models.

Bias in Credit Scoring Models

Even when AI models do not include protected characteristics directly, they may rely on *correlated proxies* (for example, postal codes, education level, social network activity). These proxies can lead to unintended disparate impacts.

In 2019, multiple customers, including Apple co-founder Steve Wozniak, reported receiving significantly lower credit limits for Apple credit cards than their spouses despite similar financial profiles. The incident prompted an investigation by the New York Department of Financial Services (NYDFS) into whether Goldman Sachs's algorithm violated fair lending laws. Although no discriminatory intent was found, the case underscored the risk of algorithmic opacity and lack of explainability in credit decisions.

Housing: AI and the Fair Housing Act

AI is now used by landlords, property managers, and housing platforms to screen tenants, price rentals, recommend listings, and analyze neighborhood desirability. These systems must comply with the Fair Housing Act (FHA), which prohibits discrimination in the sale, rental, and financing of housing.

Protected Characteristics Under FHA

The Fair Housing Act (FHA) of 1968 is a part of the Civil Rights Act of 1968, and it protects against discrimination based on several characteristics, including:

- Race
- Color
- Religion
- Sex
- National origin
- Disability
- Familial status (for example, marital status, and having children)

These protections extend to AI-powered tenant screening platforms, property management systems, and rental and mortgage advertising algorithms.

Algorithmic Advertising and Targeting

Digital platforms often use AI to optimize housing ads based on user behavior or predicted interest (targeted marketing is nothing new—digital technology has just made it better). If this results in *differential exposure* (where an AI system causes users and groups to receive different information, even when the system treats all users equally in terms of decision logic) by protected class, it may violate the FHA, even when unintentional.

In 2019, the U.S. Department of Housing and Urban Development (HUD) charged Facebook with enabling housing advertisers to exclude users based on race, religion, gender,

and postal code. Facebook and HUD later settled, and Facebook implemented new rules for ad targeting in housing, credit, and employment.

What Counts as a Violation?

Under the U.S. Fair Housing Act, violations include:

- Discriminatory targeting (for example, excluding users of any specific ethnicity)
- Discriminatory impact (for example, algorithm limits access to specific groups)
- Failure to monitor third-party tools

All these violations can trigger liability and sanctions. Housing providers and platforms must audit algorithms for fairness across all protected groups.

Tenant Screening and Credit-based Bias

AI-powered tenant screening tools may use credit scores, criminal history, or eviction records, all of which can reflect systemic bias. If these criteria disproportionately exclude protected classes, landlords may face fair housing complaints.

Several lawsuits have challenged tenant screening companies for including inaccurate or outdated eviction data that disproportionately affects communities of color. Courts have ruled that screen methods, including those enabled by automated tools, must meet accuracy, relevance, and non-discrimination standards under both the FHA and the FCRA.

Insurance: Risk Assessment in a Regulated Landscape

The insurance industry has long relied on data analytics to assess risk and set prices. With the rise of AI, insurers now deploy models that analyze behavioral data, social media activity, IoT sensors, and even facial images—all to improve the accuracy of their actuarial data. While these capabilities promise personalized pricing, they do so at the risk of introducing unlawful discrimination.

Unlike credit and employment screening, there is no single federal anti-discrimination statute for insurance. Insurance regulation primarily occurs at the U.S. state level, with varying protections and oversight mechanisms.

Discrimination in Insurance Underwriting and Pricing

Insurance discrimination can occur through underwriting (through denials or coverage limitations) and pricing (by charging different premiums to similarly situated individuals). State laws typically prohibit unfair discrimination based on race, religion, gender, disability, national or ethnic origin, veteran status, and genetic information. Some states have expanded

protections to include sexual orientation, gender identity, and occupation. This author would not be surprised to see additional protection based on social media content.

Several states (California, Hawaii, and Massachusetts) ban or restrict the use of credit scores in auto insurance underwriting due to concerns of racial bias. AI models that use such proxies may violate these prohibitions, even if unintentionally.

Proxy Discrimination and Feature Selection

A common risk in insurance AI is the use of *proxy variables*: the factors that correlate with protected characteristics but aren't themselves directly prohibited.

Proxy Variables in Insurance

Variable	Potential Proxy For
Postal code	Race, socioeconomic status, religion
Education level	Income or nationality
Occupation	Gender or class status
Credit history	Race (via systemic inequality)

Insurers must justify model features based on actuarial relevance, not correlation alone. States like Colorado have required insurers to demonstrate that their AI models do not result in disparate outcomes.

State-level Developments in Insurance AI

As the primary regulators for insurance, U.S. states are introducing legislation or guidance on AI fairness in insurance. For instance, Colorado's SB21-169 (protecting consumers from unfair discrimination in insurance practices) became the first state law to require insurers to:

- Evaluate AI and big data systems for unfair discrimination
- Submit documentation of bias testing
- Implement governance policies to mitigate discriminatory impacts

This model may inspire similar laws in other jurisdictions, raising the bar for algorithmic accountability.

By the time this book goes to print, it is expected that more U.S. states will have enacted AI regulations.

Risk Mitigation and Compliance Strategies

In every business sector with discrimination regulations, organizations must build legal risk mitigation into their AI governance frameworks (nothing new here, as any governance framework must include risk management). This section outlines steps to comply with existing and anticipated non-discrimination laws while enabling AI-powered innovation.

Perform Bias Audits and Impact Assessments

Bias audits are essential to identify disparate impacts before deployment. Key elements of bias audits include:

- Statistical evaluation of training and test datasets
- Assessment of model performance across demographic subgroups
- Review of feature relevance and proxy risk
- Documentation of audit methodologies, audit results, mitigation decisions, and outcomes
- Ongoing post-deployment monitoring to detect bias drift

AI Model Bias Checklist

Some things to look for in an audit of an AI model include:

- Whether disparate impact metrics have been tested
- Whether protected class proxies were removed or justified
- If there is a business necessity for impactful features
- Whether model outputs are explainable and reviewable

Once again, we see a case building for organizations to employ *only* explainable and transparent AI.

Ensure Explainability and Transparency

Legal compliance often depends upon whether an organization can explain why an AI system made a decision, especially in adverse actions. Essential practices include:

- Use interpretable models or post-decision explanation tools
- Provide defendable, actionable reasons for denials or adverse outcomes
- Log decisions and model behavior over time
- Periodically review model behavior over time

Establish Human Oversight and Accountability

AI decisions, particularly those affecting individual subject rights, should not occur in a governance vacuum. Strong AI governance frameworks should include:

- Human-in-the-loop reviews for high-risk and high-value decisions
- Clear ownership of AI compliance within legal or compliance teams
- Proactive curation of emerging AI-related regulations, accompanied by compliance reviews
- Regular training on civil rights laws and AI ethics
- Coordination between technical, legal, and operational staff
- Periodic governance readouts to senior executives

Include Fairness Clauses in Vendor Contracts

Many organizations rely on one or more third parties to comprise their AI capabilities that make or assist in business decisions. Contractual protections are essential to define, share, and assign liability and ensure compliance.

In addition to existing third-party risk management practices concerning legal agreements, contract terms should require vendors to:

- Conduct and share the results of bias audits
- Disclose data sources and model documentation
- Agree to indemnify for regulatory violations
- Assist with explainability and adverse action notices

Many organizations' legal departments utilize legal agreement templates, which should be periodically updated to stay current with these and other contractual matters.

Fairness as a Legal Mandate, Not Just an Ethical Goal

Fairness in AI systems is not just a good idea: it's the law. AI systems operating in employment, finance, housing, and insurance do not exist outside the law. Longstanding anti-discrimination laws, from Title VII to the Fair Housing Act, apply whether humans or machines make decisions.

Organizations considering deploying AI in these sensitive contexts must:

- Realistically evaluate the economic or competitive advantage that an AI system will bring
- Understand and comply with applicable laws
- Anticipate new regulations
- Audit models for disparate impact and proxy bias

- Provide transparency and explanations to affected individuals
- Document governance procedures to demonstrate proactive management and good faith

Fairness in AI is not just a moral or ethical imperative; it's a legal obligation, one that regulators and plaintiffs are enforcing.

Red Flags That Trigger Legal Scrutiny

Regulators, courts, and advocacy groups may focus on AI systems that are observed to:

- Rely on opaque models that lack explainability
- Use credit history, postal codes, or education without reasonable justification
- Be unable to produce a specific reason for adverse decisions
- Be deployed without testing for disparate impact
- Deny requests for reasonable accommodations

Avoiding these pitfalls begins with embedding legal compliance into AI design, development, and deployment from the start. Business purpose and design reviews prior to development help organizations avoid these and other issues.

How Consumer Protection Laws Apply to AI

As AI becomes more embedded in consumer products and services, it increasingly influences how we shop, learn, get healthcare, and manage our finances. When these systems mislead, deceive, manipulate, or harm users, they may violate legal prohibitions against *unfair or deceptive acts or practices (UDAP)*, otherwise known as false or misleading advertising. This section explores how U.S. consumer protection laws apply to AI systems and provides practical guidance for governance, transparency, and risk mitigation.

Legal Foundations: Unfair and Deceptive Acts or Practices

Consumer protection laws in the United States derive from both federal and state statutes prohibiting unfair, deceptive, and fraudulent conduct in commerce. These laws are intentionally broad and flexible, making them a powerful tool for regulating AI behavior.

The FTC Act and Section 5 Authority

The U.S. Federal Trade Commission (FTC) is the primary federal enforcer of consumer protection laws. Under Section 5 of the *FTC Act*, the agency may take legal action against:

- **Unfair acts or practices:** Those that cause or are likely to cause substantial injury to consumers, are not reasonably avoidable, and are not outweighed by benefits
- **Deceptive acts or practices:** Those that mislead or are likely to mislead a reasonable consumer, based on material representations or omissions

These standards apply whether harm arises from a human sales tactic, a website, or an AI model.

UDAP vs. Fraud

Unfair and deceptive acts and practices (UDAP) differ from fraud as follows:

Unfair or Deceptive Act	Fraud
Civil liability	Often criminal
No intent required	Requires proof of intent
Broad application	Narrow, specific conduct
Enforced by the FTC, states	Prosecuted by DOJ, AGs, or plaintiffs

State-level Consumer Protection Laws

Every U.S. state enforces its own version of UDAP statutes, sometimes referred to as "mini-FTC Acts." These laws often provide additional protections, broader definitions of harm, and private rights of action. Examples include:

- California's Unfair Competition Law (UCL) borrows heavily from the FTC Act and prohibits "any unlawful, unfair or fraudulent business act or practice."
- Massachusetts Consumer Protection Law 93A allows consumers to recover triple damages for certain deceptive conduct.
- New York's General Business Law (GBL) §349 focuses on deceptive conduct.

These laws create a patchwork of risk for companies deploying AI tools across jurisdictions. As there is no U.S. federal prohibition on state-level AI laws, this patchwork promises to become more complex over the next few years. If history is any guide (thinking of federal and state privacy laws), it's unlikely that a sweeping federal AI law will supersede U.S. state laws.

Emerging Focus Areas for AI Enforcement

Consumer protection authorities are aware of the trend of organizations that are grafting AI capabilities into their products and services. Regulators are increasingly concerned with:

- **Undisclosed AI use:** Consumers interacting with AI systems may be unaware that they are not dealing with a human (thus far, regulations are not universally requiring this disclosure, yet it may be ethical to do so).
- **Manipulative design:** AI that nudges, tricks, or coerces consumers into behavior they wouldn't otherwise choose.
- **Inaccurate or harmful content:** Chatbots or recommendation engines that produce false or unsafe information.
- **Opaque decision-making:** AI systems that affect eligibility or pricing but offer no explanation or recourse.
- **Data privacy and security:** AI systems collecting, inferring, or sharing more personal data than is necessary or in ways that consumers did not consent to.

These concerns are leading to new enforcement actions, guidance documents, and policy proposals—all aiming to hold AI developers and deployers accountable.

Deception or Superior Service Practices

Virtually everyone is familiar with their social media "feed," as well as viewing and purchasing opportunities from streaming sites (for example, Netflix, Prime Video) and e-commerce sites (for example, Amazon, Shopify, Squarespace). Most, if not all, of these employ some sort of ML or AI algorithm to choose what we see. Would such a practice be considered manipulative (because we might be "tricked" into purchasing more) or simply a superior service practice by showing us what we're more likely to want?

FTC Enforcement and Policy in the AI Era

As it has for consumer privacy, the FTC has positioned itself as a key AI watchdog, leveraging its existing consumer protection authority to police harms caused by AI. While it has not yet published formal AI regulations, the agency has issued guidance and brought enforcement actions under Section 5.

FTC Guidance on AI Fairness and Transparency

In 2021 and 2022, the FTC released guidance via blog posts and policy statements, warning companies that:

- AI systems that produce biased, unfair, or deceptive outcomes may violate the law.
- Claims about AI accuracy and AI capabilities must be demonstrable and substantiated.

- Companies are responsible for AI behavior, even when using third-party AI tools (this harkens back to established third-party risk management practices).
- Data hygiene and provenance are critical for lawful model development.

These warnings reinforce the FTC's view that AI is not exempt from accountability, simply because it is complex or challenging to explain.

FTC's Five AI Governance Questions

According to the FTC, companies should determine whether:

- Their AI datasets are biased or incomplete
- Their AI-powered decisions and results discriminate against protected groups
- Consumers can opt out of automated decisions
- They are transparent about when and how AI is used
- They can explain and justify their AI-powered decisions and outcomes

Failing to address these issues can bring unwelcome attention.

Key FTC Enforcement Actions Involving AI

Although many legal cases are settled without admission of liability, they offer valuable insight into regulatory expectations.

Everalbum (2021)

Everalbum offered a photo storage app that employed facial recognition without clearly informing its users. The FTC alleged that the company misled consumers by enabling facial recognition by default without notice and retaining biometric data even after users deactivated their accounts.

Everalbum was required to delete its facial recognition models and data and affirmatively disclose its biometric practices going forward.

Weight Loss App "CuraLife" (2022)

Although not strictly AI-based, this case targeted false claims about automated health analysis. CuraLife allegedly marketed AI-based diabetes insights without adequate scientific backing.

Companies must disclose uses of AI and substantiate any claims about AI accuracy, health benefits, or personalization, just as they would for traditional products.

Amazon Alexa and Children's Data (2019)

The FTC and DOJ investigated and fined Amazon's Alexa product for violating the Children's Online Privacy Protection Act (COPPA) by retaining children's voice recordings,

failing to delete requested data, and not adequately disclosing how Alexa used the information.

While not strictly an AI case, this example illustrates how voice assistants, speech recognition, and personalized targeting can intersect with consumer protection and privacy law.

AI-powered Deception and Manipulation

Not all AI applications raise consumer protection concerns, but some use cases may be susceptible to manipulation, misinformation, or deception. These include generative models, chatbots, recommendation engines, and decision-assistance tools. This section explores how consumer protection law responds to AI systems that shape user behavior or mislead the public.

Dark Patterns and Manipulative Interfaces

As first discussed in Chapter 4, *dark patterns* refer to user interface designs that trick, coerce, or confuse consumers into making choices against their best interests, such as:

- Subscribing to services unintentionally
- Sharing more data than they intended
- Clicking on misleading or pre-checked options
- Struggling to cancel or opt out

AI exacerbates these risks by dynamically adapting content based on real-time behavioral predictions, optimizing for engagement or monetization rather than consumer welfare, and personalizing manipulation at scale. Why manipulate customers one at a time when AI can do so more efficiently?!

For example, imagine an AI-based service learns that late-night users are more impulsive and adjusts its interface to highlight premium upsells after 10 p.m. If these nudges cause consumers to subscribe unintentionally and make cancellation difficult, regulators may view this as unfair or deceptive. Many of us have experienced services that can be purchased with just a few clicks. However, canceling requires a lengthy navigation through a confusing auto-attendant's menu, together with long hold times. Many would simply give up, which is often the goal.

Dark Patterns and AI: Key Warning Signs

- Interfaces that hide or obscure opt-out and unsubscribe buttons
- "Confirm-shaming" (for example, "No thanks, I like paying full price")
- Deceptively labeled toggles or options
- Real-time A/B testing that increases conversion but reduces clarity

Several states, including California and Colorado, have adopted dark pattern bans in their consumer and AI-related laws.

Chatbots, Impersonation, and Misleading AI Interactions

As conversational AI becomes indistinguishable from real human conversation, regulators are focused on ensuring that users are disclosed when they are interacting with a machine, particularly in commercial or advisory contexts.

Suppose a chatbot offers guidance in categories such as legal, health, spiritual, or financial, but consumers reasonably believe they are interacting with a human expert. In that case, a lack of disclosure may constitute a material omission under UDAP standards.

AI-generated voices and deepfakes are increasingly used in scams or misleading endorsements. With or without malicious intent, impersonating someone for commercial or personal gain may violate consumer deception laws, right of publicity statutes, or guidance from the FTC regarding endorsements.

Fake Reviews, AI-generated Endorsements, and Influencer Fraud

Consumer trust in reviews and testimonials is a foundational aspect of online commerce. Increasingly, however, many online commerce sites have reviews of questionable repute, whether generated by AI or reviews to artificially inflate product quality. AI now enables:

- Fake reviews generated en masse
- Positive feedback auto-generated for one's own products
- Deepfake influencers promoting goods or services without disclosure

In 2023, the FTC updated its Endorsement Guides and proposed rules requiring disclosure of AI-generated testimonials, prohibition of fake or unverified reviews, and accountability for businesses that benefit from deceptive reviews, even if generated by affiliates or third-party vendors. AI-generated reviews that appear authentic but are fabricated violate the "reasonable consumer" standard and are likely to be deemed deceptive, resulting in legal action and sanctions.

Emerging Risks: Deepfakes, Misinformation, and Emotional Manipulation

AI's ability to generate realistic images, videos, and audio introduces new frontiers of consumer and business deception, especially as tools become more accessible and powerful. As this book is being written, the author notes that the tools available for creating deepfakes are widely available and produce remarkably realistic content.

Deepfakes can be used to fabricate celebrity endorsements, create false evidence of events or statements, forge false governmental IDs, imitate voices or faces for scams, and disseminate false political or financial information. In May 2025, the United States enacted the Take It Down Act, which criminalizes the distribution of non-consensual intimate images, including AI deepfakes. Also, the United Kingdom's Online Safety Act considers the sharing of non-consensual explicit deepfake content as a criminal offense. Deepfakes may also be prosecuted under:

- Unfair or deceptive practice laws, if used for marketing or profit
- Right of publicity laws, if likeness is misappropriated
- Wire fraud statutes, if used for impersonation scams

Several states (for example, California, Texas, and New York) have enacted laws targeting malicious deepfake use in political or pornographic content, and more states are likely to follow.

Deepfake Red Flags

- Unrealistically high-quality endorsements, confessions, and directives
- Videos shared without metadata or original source
- Synthetic voice used in customer service or support calls without notice
- Out-of-character statements—a person says things that contradict their usual views or style

AI Misinformation in Health, Finance, and Safety

Consumers may suffer harm when generative AI models provide inaccurate health advice, false financial guidance, and erroneous safety instructions. Even without malicious or misleading intent, companies may be held liable when AI is marketed as trustworthy or authoritative, consumers rely on the information in their research and decision-making, and if the harm was foreseeable and preventable.

Here's a real-world example of this: while ChatGPT includes disclaimers that it is not a medical provider, many rely on its outputs for diagnosing symptoms or medication interactions. Inappropriate reliance could lead to legal claims.

Emotional Manipulation and Vulnerable Populations

AI tools that engage with emotionally vulnerable consumers, such as children, seniors, or those experiencing grief or mental health issues, may be subject to heightened scrutiny. This concern is particularly relevant in therapeutic chatbots, AI grief companion apps, and AI-generated content for children.

If these tools misrepresent their capabilities or exploit emotional states for engagement, profit, or fraud, they may violate unfairness standards under Section 5 of the FTC Act.

Romantic Relationships with AI

AI tools are increasingly being used for romantic companionship. A broad survey in the United States by the Wheatley Institute found that about 19% of young adults aged 18–30 have interacted with romantic AI chatbots. Further, 31% of young men and 23% of

young women have reported such interactions. A YouGov-backed study revealed that 7% of unmarried adults claimed to be open to, or currently involved in, a romantic relationship with an AI partner.

The Institute for Family Studies conducted research that revealed that one in four young adults believe that AI partners could replace real-life romance.

FTC regulations likely apply to AI romance services if those services misrepresent their capabilities or fail to protect personal information.

Risk Mitigation and Consumer Protection Compliance Strategies

Avoiding unfair or deceptive AI outcomes is not just a legal imperative; it's central to building trust and minimizing reputational and regulatory risks. This section outlines how organizations can operationalize consumer protection principles through documentation, design, and accountability.

Design for Transparency and Disclosure

Consumers must explicitly understand when AI is involved in decision-making or interactions. Organizations must ensure that this transparency is a part of the design of every AI system. This transparency can be achieved through:

- Clear disclosures that users are interacting with AI (for example, "I'm an AI assistant, not a human representative")
- Accessible explanations (not buried in terms of service) of how decisions are made (for example, credit denials, pricing)
- Contextual disclaimers that AI responses are not professional advice in healthcare, finance, or law
- Model limitations statements, indicating where AI may be incomplete, uncertain, or prone to error
- Easily found help resources or escalation paths to a human representative

These "you are interacting with an AI" disclosures must be prominent (not buried in terms of service). These disclosures must be expressed in plain language and provided when the AI is used, not just at service or session onboarding.

When Google introduced its Google Duplex AI voice assistant capable of placing phone calls, it explicitly programmed the tool to identify itself as an AI during the call, a transparency step praised by some regulators and advocates. Incidentally, Google shut down Duplex in 2022 and now offers outbound calling through its Search product.

Substantiate AI Claims

Marketing claims by organizations about their AI systems must be accurate and demonstrable. Common violations include:

- Overstating capabilities (for example, "guaranteed accuracy" or "human-level judgment")
- Misleading endorsements or testimonials
- Implying that AI is safer or more objective without providing evidence

Quantitative claims, such as accuracy rates or cost savings, must be supported by robust data and reproducible testing.

Often, organizations will bolt AI onto an existing service and then boast about their superior "AI-powered" capabilities. Because so many organizations lack effective AI governance, these claims are sometimes little more than wishful thinking.

Checklist for AI Marketing Compliance

As they create updated marketing content touting AI capabilities, marketers should check that:

- All AI performance claims are supported by quantitative or qualitative evidence.
- The scope of AI's role is clearly stated (human-assisted vs. fully automated).
- Disclaimers are visible where needed.
- Third-party tools' limitations are disclosed.
- Testimonials are authentic and appropriately labeled.

Monitor for Algorithmic Drift and Harm

AI systems may degrade over time, adapt to new data in unexpected ways, or interact with user behavior in emergent, risky patterns. Moreover, even if an AI model doesn't degrade over time, its relevance still can suffer: last year's AI model may not serve the business as well this year. These and other forms of degradation make ongoing monitoring essential. Best practices include:

- Routine performance and fairness audits
- Red teaming to test for abusive or manipulative behavior
- Logging and reviewing user complaints
- Monitoring for false positives, hallucinations, or unsafe responses
- Security monitoring for prompt injection, data poisoning, or model exfiltration attempts

AI systems should have *kill switches*, safeguards, event and incident monitoring, and escalation paths to allow human intervention.

Implement AI Governance Policies

As stated many times in this book, policy is "the law" in organizations. When building AI governance processes, they should be backed up with policy. To ensure consumer protection is embedded in AI design, development, and operations, organizations should establish:

- AI product review committees that include legal, UX, cybersecurity, privacy, and ethics input
- A cross-functional AI risk register
- Documented model cards or datasheets that describe purpose, limitations, and known risks
- Incident response plans for AI-related consumer complaints

Governance should cover the entire AI lifecycle, from data sourcing and model training to deployment and sunsetting.

Vendor and Third-party Risk Controls

Many consumer-facing AI services rely on third-party tools, APIs, and datasets. Organizations remain legally responsible for how these components function, including operational events outside of their control.

To manage this risk, organizations must vet vendors for compliance history and auditing practices, include consumer protection obligations in contracts, require right to audit provisions, and establish incident notification terms in case of model failure or user harm. These requirements are a part of core expectations in the practice of third-party risk management.

A Proactive Approach to AI and Consumer Protection

AI is transforming how consumers access services, make decisions, and engage with the world, but it also introduces new avenues for manipulation, deception, and harm. Consumer protection laws provide a flexible, technology-neutral framework for holding companies accountable when AI causes unfair or misleading outcomes.

When identifying applicable laws and regulations in AI and privacy, organizations must document their approach to compliance with those regulations. However, because many aspects of AI are yet to be regulated, organizations should tread carefully and assume that those regulatory voids will eventually be filled. Hence, organizations should proactively:

- Design transparency and fairness into AI products and capabilities from the outset
- Audit for emerging harms and unintended behaviors
- Avoid overpromising AI capabilities
- Disclose when users are interacting with machines
- Monitor how real-world consumers experience AI tools

Trust in AI will depend more on an organization's transparency, legal integrity, and ethical responsibility regarding AI systems and less on technical performance. When organizations follow these principles, they are less likely to have to retrofit their AI systems when regulations are enacted.

AI Red Flags for Product Teams

Red Flag	Why It Is Risky
"100% accurate" AI claim	May be deceptive unless provable
AI chatbot provides medical or legal advice	Could mislead consumers into unsafe decisions
No disclosure when users interact with AI	Violates material disclosure expectations
Optimizing UI for engagement over clarity	May trigger dark pattern enforcement
AI-generated reviews or endorsements	Likely deceptive unless clearly disclosed
Claims based on outdated or biased training data	May cause unfair outcomes without recourse

These red flags should trigger internal review and legal vetting before deployment.

How Product Liability Laws Apply to AI

AI systems are increasingly embedded in products that interact with the physical world: cars, medical devices, drones, appliances, and industrial and domestic robots. When these systems malfunction or behave in unexpected ways, the consequences can be catastrophic and even life-threatening. Product liability law provides a legal framework for determining who is responsible when products cause harm. However, the traditional categories of design, manufacturing, and warning defects were developed in an era before autonomous and adaptive systems. This section explores how product liability laws apply to AI and outlines emerging legal challenges in assigning responsibility for AI-enabled harms.



Even when you do not see a direct linkage between product liability laws and AI systems, assume the linkage exists.

Product Liability and Its Application to AI

Product liability is a branch of *tort law* that holds manufacturers, designers, and sellers liable for producing and selling defective products. Product liability is well established for tangible goods; however, applying it to software-driven, autonomous, or learning systems is not straightforward.

Theories of Product Liability

Product liability claims in the United States are typically based on one or more of the following:

- **Design defect:** A product is manufactured as intended, but the design itself is unreasonably harmful or dangerous.
- **Manufacturing defect:** The product deviates from its intended design due to an error in manufacturing or assembly.
- **Failure to warn (marketing defect):** The product lacks adequate instructions or warnings about risks.

These are applied through:

- **Strict liability:** Plaintiff does not need to prove negligence, only that the product was defective and caused harm.
- **Negligence:** Plaintiff must show that the defendant failed to act with reasonable care in design, manufacturing, testing, or oversight.
- **Breach of warranty:** Based on contractual promises about product safety or performance.

Strict Liability vs. Negligence

Feature	Strict Liability	Negligence
Standard	Defective product causes harm	Failure to act with reasonable care
Intent	Not required	Required (failure to prevent foreseeable harm)
Common Use Case	Design and manufacturing defects	Inadequate testing, supervision, or warning
AI Example	Autonomous vacuum explodes due to a flaw	AI decision causes harm due to poor oversight

What Constitutes a "Defect" in AI-driven Products?

In traditional product liability law, a defect refers to a flaw that makes a product unreasonably dangerous when used as directed. With AI systems, the concept of a defect becomes murkier for several reasons. First, an AI system's behavior is dynamic and context-sensitive. Additionally, AI models may *drift* (change after initial deployment), and outcomes are probabilistic, not deterministic.

Legal scholars and courts are still contending with questions such as:

- Is an AI defect a coding error, a data flaw, or an unexpected decision?
- Who is liable when a product performs differently based on user behavior or data input?

Regarding product liability, manufacturers are often liable for foreseeable misuse, but with AI, "foreseeability" is harder to define because of emergent behaviors. Foreseeability is becoming a key legal battleground.

In 2018, an autonomous Uber vehicle struck and killed a pedestrian in Arizona during a testing phase. Investigations found the software failed to classify the person as a hazard due to poor object detection design. Was the defect in the algorithm, sensor configuration, or testing protocol? The ambiguity illustrates how difficult it is to pinpoint and prove "defects" in systems that make independent decisions.

In 2025, Hertz, Thrifty, Avis, Sixt, and other car rental companies were using AI scanning systems to check returned rental cars for damage. Then the complaints began rolling in—customers were being charged for damage that, in most cases, was either a minor scuff or not apparent at all. However, the worst part is that there was no human oversight or appeal process available. At the time of writing this book, this is a developing story.

Expanding the Definition of "Product" to Include Software and AI

Under traditional legal frameworks, courts were reluctant to treat pure software as a product because software was viewed as a service (not a tangible good), licensing agreements (not sales) governed transactions, and software behavior could not be fully predicted.

As more court cases on high-tech product liability are tried, and as trial lawyers and judges learn more about AI, legal positions are shifting. Courts are increasingly willing to treat embedded software and AI as part of a product system, particularly when the software directly controls physical hardware, the software's function is essential to safety, and AI performs tasks that replace human judgment.

Restatement (Third) of Torts: Products Liability

The influential Restatement (Third) defines a product as "tangible personal property distributed commercially for use or consumption." While software isn't always included, embedded software that operates a physical product (for example, autopilot, insulin pumps, drones) may qualify.

Design Defects in AI Systems

Design defect claims generally allege that a product's design is inherently flawed and creates unnecessary risk for its users. With AI, these claims focus on the model architecture, training data selection, risk assumptions, or failure to account for foreseeable misuse.

Risk-utility Balancing and the "Reasonable Alternative Design"

In the United States, courts often apply a rule called the risk-utility test that asks, "Was the risk of harm outweighed by the product's utility, and could a reasonable alternative design have reduced the risk without compromising functionality?"

In the AI context, plaintiffs in a product liability lawsuit may allege that the AI system used insufficient or improper training data, that safer model architectures were available, the design failed to include human oversight, or a safer default behavior was feasible.

In a hypothetical example, a robotic surgery assistant makes an improper incision due to a misclassified image from its AI vision system, and the patient suffers harm. If safer classification algorithms had been available and were ignored, or human verification could have prevented the error, a design defect claim may succeed.

Manufacturing Defects in AI-integrated Products

Manufacturing defects occur when the final product differs from its intended design due to a manufacturing error. These claims are generally easier to prove than design defects because they involve clear deviations from their design.

Manufacturing defects in the context of AI include:

- Faulty sensors or cameras installed in autonomous vehicles
- Corrupted firmware or software deployment
- Hardware failures that impair AI perception or control
- Integration errors between AI components and mechanical systems
- Compatibility issues with components manufactured by third parties

AI as an Assembly of Interdependent Parts

Because AI-enabled products rely on hardware, software, connectivity, and data inputs, even small variances in one component can lead to unpredictable behavior. One could say that AI manufacturers are not manufacturers but integrators, who assemble components from several sources, resulting in the final AI system.

An industrial robot guided by AI is designed to stop when encountering a human. During production, an improperly calibrated proximity sensor is installed, and the robot fails to halt, injuring a worker. Though the AI software performed correctly, the hardware deviation constitutes a manufacturing defect. Were this to result in a legal proceeding, courts or arbitrators are likely to find strict liability where the defect is traceable to the manufacturing

process, the end product was used as intended, and harm was foreseeable and directly caused by the defect.

Challenges in Detecting AI Manufacturing Defects

Because AI systems are non-deterministic, distinguishing normal variation from a true defect can be a challenge, complicating the plaintiff's burden of proof in traditional liability claims. Key issues in legal challenges include a lack of transparency in AI system decision-making, model drift, and supply chain complexity.

These difficulties have prompted discussion of shifting burdens, requiring manufacturers to prove a system was *not* defective after harm occurs. Such shifts remain controversial and jurisdiction-dependent.

Imagine, if you will, an incident between an autonomous vehicle that strikes a kangaroo carrying a ladder. The kangaroo's owner might sue based on a manufacturing defect. How would the vehicle's manufacturer be required to prove its vehicle was free of defects, including the most unlikely use cases in safety testing?

Failure to Warn: Disclosing AI Risks

Product liability law also requires product manufacturers to warn consumers about non-obvious dangers associated with product use. This duty is ongoing: it applies at the time of sale and may extend as risks evolve post-deployment.

The makers of consumer AI systems may want to consider notifying customers when:

- User instructions omit known limitations of AI accuracy
- Product disclaimers are buried or not written in plain language
- Users are unaware that AI is controlling or influencing outcomes
- Users are not informed about edge cases or risk conditions

Given the proclivity for high-tech startups to focus first on time-to-market and later (if at all) on compliance with applicable laws, we may see a season of product defect enforcement actions against AI system vendors.

Human-machine Interface and User Misunderstanding

Users may become overreliant on AI systems, assuming they are accurate and more capable than they really are. This risk increases when AI systems are marketed as "autonomous," "self-learning," or "safe," or when the UI gives no indication of uncertainty or error, or when user guides are insufficient for real-world use.

Although Tesla's "Autopilot" and "Full Self-Driving" features require constant driver supervision, some consumers have misunderstood the capabilities. This confusion has led to litigation and regulatory scrutiny, with critics arguing that the branding itself constitutes a failure to warn. In mid-2025, a federal jury awarded over $240 million in damages to the family of a Tesla owner who died in a 2019 Tesla Autopilot-related crash.

Updating Warnings Post-deployment

Because AI systems evolve and new risks may surface after launch, AI system vendors may have a continuing duty to warn, including software update notices, safety alerts, recall announcements, and public disclosures of known issues.

Courts have yet to definitively establish how far this duty extends, especially for AI systems that learn or adapt in the field.

Preemption and Regulatory Tension

In regulated industries such as healthcare, automotive, and aviation, companies may argue that federal regulations preempt state-level product liability claims. For instance, when the FDA (Food and Drug Administration) approves an AI-powered medical diagnostic tool for commercial use, U.S. states should not be permitted to enforce state product liability laws. And similarly, automobile manufacturers whose autonomous driving systems have been certified by the NHTSA (National Highway Traffic Safety Administration) would likewise argue that U.S. states should not enforce state product liability laws.

As new case law emerges, courts will need to balance the role of federal oversight with the consumer's right to recover for injury.

Preemption and AI

Claim	Potential Preemption
Design defect in an AI medical device	May be preempted by FDA approval
Failure to warn about AI in drones	Unlikely to be preempted
Post-sale updates causing harm	Less likely to be protected

While preemption can limit liability, courts tend to favor warnings and consumer rights when there's ambiguity.

Continuing Legal Uncertainty

Because AI-driven product injuries are relatively new, case law is limited but growing. Courts have yet to establish firm precedents on evaluating causation in systems that rely on probability, who is liable when an autonomous system makes harmful decisions, and whether AI can be considered a product.

Multi-party Liability and the AI Supply Chain

The supply chains for AI products are quite complex and differ significantly from those of traditional consumer products. Product liability traditionally allows lawsuits against product designers, manufacturers, distributors, and sellers. However, AI systems' supply chains are quite different and consist of hardware vendors, cloud service providers, model developers, data brokers, and system integrators. In a liability-related legal action, an AI system's supply chain actors are likely to deny responsibility and blame others.

Here is a hypothetical example: a firmware update in an AI-powered smart-home system malfunctions and disables smoke detectors, contributing to a housefire. Liability may be disputed between the smart device manufacturer, the cloud AI provider, and the software vendor that created and pushed the update. This example reminds the author of a situation in which the owner of an automobile experiences a drivability problem, whereby the wheel vendor blames the tires, and the tire vendor blames the wheels, leaving the owner without a remedy.

Legal ambiguity about joint and several liability in AI contexts makes legal proceedings and litigation more complex.

The European Union's Approach to AI and Product Liability

The European Union is working to modernize its product liability regime to include AI-related risks. In 2022, the European Commission proposed a revised Product Liability Directive to cover software and digital updates, together with a presumption of defect in high-risk AI cases (for example, autonomous vehicles), as well as expanded liability to developers and deployers, not just manufacturers.

While not directly applicable in the United States, these proposals may influence U.S. regulations, litigation, product design, and overall compliance strategy.

Proactive Risk Mitigation Strategies for AI Product Liability

Because product liability litigation is reactive (triggered only after harm occurs), organizations deploying AI in consumer or industrial products must adopt preventive risk strategies. These include engineering controls, documentation, cross-functional governance, and external validation. In other words, AI products must be safe, secure, transparent (where explainability is a factor), and compliant *by design*.

This section includes several proactive measures that organizations in the AI system supply chain can consider.

Conduct AI-specific Safety Testing

Expand AI product testing to account for:

- Edge-case behavior in non-deterministic environments
- Bias and fairness audits for outputs affecting physical outcomes

- Simulation testing for rare but catastrophic events
- Human-AI interaction risks, including overreliance and confusion
- Vulnerability risks, including adversarial patch, prompt injection, and data poisoning

Essential practices in AI system safety testing include:

- Using *digital twins* or sandboxed environments to simulate real-world use
- Stress testing AI models using unpredictable inputs (akin to *fuzzing* in cybersecurity)
- Validating model decisions against regulatory benchmarks and guidelines

Before deployment, an autonomous, AI-driven drone system is tested against GPS spoofing, wind shear, pedestrian crossings, and sensor failure. These simulations reveal unanticipated altitude drop patterns—allowing correction before real-world harm occurs.

Maintain Traceability Across the AI Lifecycle

Product liability claims often hinge on proving (or disproving) the origin of a product defect. AI governance for any organization in the AI supply chain must ensure traceability of training data provenance and archival, model versions and update logs, test results and validation artifacts, and source code commits and deployment records. In other words, every artifact in the component's development lifecycle should be captured and archived.

AI supply chain organizations should maintain an *AI system bill of materials (AI-SBOMs)* that documents dependencies and subcomponents of an AI system. This recommendation is akin to a *software bill of materials (SBOM)*, the inventory of components used to build software, which is increasingly common in the software industry.

What to Track in an AI-SBOM

Vendors in the AI system supply chain can begin building their AI-SBOM to include:

Element	Purpose
Model version and ID	Identify which version caused harm
Dataset sources	Reveal bias or licensing issues
Preprocessing steps	Explain unexpected model behavior
Firmware/software logs	Validate configuration at the time of error
Update timestamps	Determine if recent changes were causal

Organizations evaluating AI systems should consider whether AI system vendors maintain an AI-SBOM. This level of documentation can reduce liability by demonstrating due diligence and control.

Build Cross-functional Product Safety Governance

AI liability risk management must not sit solely with any one team, such as engineering or legal. Instead, organizations should establish a multidisciplinary product review board as a part of their overall AI governance structure. Such a review board would involve engineering, cybersecurity, privacy, risk management, legal/compliance, UX and human factors, and risk and safety specialists. An AI product review board would oversee product safety, hazard analysis and mitigation, regulatory disclosures, and *post-market surveillance*.

Before launching an AI-driven insulin pump, a medical device company conducts a Failure Modes and Effects Analysis (FMEA), convenes a product safety board, and retains third-party auditors. These steps mitigate design and marketing liability while aligning with FDA expectations.

Integrate AI Update, Governance, and Field Monitoring

Because the behavior of some AI models can change after deployment, AI governance must include ongoing operations, including real-world performance monitoring, event and incident escalation procedures, and contingency plans that include *model rollback* and *human override*.

Organizations should consider implementing automated alerting for performance changes, unusual patterns, claims of harm, and model behavior that is inconsistent with validation results.

Regulators are expected to increase their expectations for post-market surveillance of high-risk systems, particularly in transportation, healthcare, and infrastructure.

Modernizing the Liability Mindset

Traditional product liability law is based on physical products, static designs, and linear causality. AI challenges all these assumptions, introducing autonomy, adaptivity, and inference, often without direct human control. These characteristics don't eliminate liability; instead, they complicate it, for the most part, because regulations and case law have not caught up with AI's societal impact.

Hence, organizations building or deploying AI-enabled products should assume that harms will occur and implement defensible controls beforehand. These controls may include:

- Design safety into model and system requirements, architecture, design, testing, and operations.
- Test beyond normal conditions to include edge cases, fuzzing, and drift.

- Threat-model the entire end-to-end AI system.
- Communicate limitations and risks clearly to users.
- Document every step, from requirements to design to update.

Legal systems are adapting but more slowly than the pace of AI innovation. In the meantime, product teams must bridge the gap between 20th-century legal frameworks and 21st-century intelligent systems by thoughtfully anticipating developments in practices, standards, regulations, and case law.

Summary

As AI systems become more integrated into critical sectors of society, they intersect with a broad range of existing legal frameworks across multiple jurisdictions and industries. This chapter examined four major categories of sectoral and civil laws (intellectual property, anti-discrimination, consumer protection, and product liability) to understand how each applies to the design, deployment, and use of AI.

Intellectual property law plays a foundational role in regulating how AI systems are trained and how their outputs may be used or protected. Copyright law presents new challenges, as training data often includes copyrighted content scraped from the web without owners' consent. While developers argue that training constitutes fair use, courts are increasingly scrutinizing whether the use of copyright data to train AI systems infringes on existing rights. Patent law raises separate concerns about whether AI-generated inventions may be patented, and who (or what) qualifies as an inventor. Meanwhile, some organizations utilize trade secret protections to shield their proprietary model architectures and datasets. Globally, differences in database rights and copyright exceptions complicate cross-border deployments. Organizations must track data provenance, obtain proper licenses, and manage IP risk through contractual and technical controls.

Anti-discrimination laws govern the use of AI in domains such as employment, credit, housing, and insurance. Laws like Title VII of the Civil Rights Act, the Equal Credit Opportunity Act, and the Fair Housing Act prohibit not only intentional discrimination but also unintentional outcomes with a disparate impact on protected groups. AI systems that rely on biased training data or use proxy variables, such as postal code or education level, may unintentionally produce discriminatory outcomes, triggering legal exposure. Regulators like the EEOC and CFPB have clarified that liability remains with organizations, even when AI tools are developed by third parties (which is consistent with established third-party risk management doctrine). State and local laws reflect a growing trend toward proactive bias detection and governance.

Consumer protection laws, especially Section 5 of the Federal Trade Commission (FTC) Act, as well as state UDAP ("unfair, deceptive, and abusive practices") laws, prohibit unfair and deceptive practices in commerce. These laws apply to AI systems that mislead consumers, exploit behavioral vulnerabilities, or obscure material facts. Key risks include

concealed AI use, fake reviews, dark patterns, overhyped claims, and deception. The FTC has issued guidance and brought enforcement actions against companies that misused AI in facial recognition, health claims, and data retention. Best practices include ensuring transparency, substantiating AI-related marketing claims, and implementing safeguards for emotionally or economically vulnerable users.

Product liability laws compound complexity for AI systems that interact with the physical world. When an autonomous vehicle, surgical robot, or smart appliance causes harm, courts may apply existing theories of design defect, manufacturing defect, or failure to warn. However, these doctrines struggle with adaptive behavior, probabilistic (vs. deterministic) decision-making, and complex, high-tech supply chains. Liability may attach to any entity involved in designing, integrating, or deploying the AI, depending on the nature of the defect and the foreseeability of harm. Companies must proactively test AI systems under real-world conditions, maintain detailed documentation, and update risk disclosures as systems evolve.

Together, these legal regimes demand a lifecycle approach to AI governance that incorporates risk assessments, cross-functional oversight, and a commitment to fairness, safety, and transparency in every stage of the AI development and operations pipeline.

Exam Essentials

Appreciate that regulations and standards have not caught up to AI's new paradigms. With history as a guide, AI capabilities are creating challenges with existing regulations and case law. Organizations should err on the side of caution and conduct rigorous diligence when implementing new AI systems.

Understand how intellectual property laws affect AI training and outputs. AI systems trained on copyrighted content may infringe on intellectual property rights, particularly if they reproduce or closely mimic protected works. Organizations must assess whether copyrighted training data use qualifies as fair use, secure appropriate licenses, and document data provenance to mitigate IP risk.

Recognize how non-discrimination laws apply to AI in regulated sectors. Laws like Title VII, ECOA, and FHA prohibit intentional and unintentional discrimination and practices that disproportionately impact protected groups. AI systems used in hiring, credit, housing, and insurance must be audited for disparate impact and must not rely on proxy variables tied to race, gender, or other protected traits.

Identify consumer protection risks posed by AI deception or opacity. Unfair or deceptive acts—such as fake reviews, undisclosed AI interactions, misleading claims, or manipulative interfaces—can violate the FTC Act and state UDAP laws. Organizations must be transparent about how AI is used, substantiate marketing claims, and provide users with clear, accessible information.

Apply existing product liability principles to AI-driven physical systems. When an AI system causes harm in the real world, courts may apply traditional product liability doctrines, including design defect, manufacturing defect, and failure to warn. AI's autonomous and adaptive behavior complicates causation and responsibility, requiring rigorous testing, version tracking, and post-market monitoring.

Integrate legal and compliance risk management into the AI lifecycle. Compliance with sectoral and civil laws cannot be an afterthought. Organizations must embed legal review, bias mitigation, documentation, and human oversight throughout the AI development pipeline, especially when systems affect rights, safety, or access to critical goods and services.

Review Questions

The following questions are designed to test your understanding of this chapter's material. For more information on how to obtain additional questions, see this book's Introduction. You can find the answers in Appendix A.

1. Which of the following best describes a potential copyright issue with AI training data?
   - **A.** Using data from government websites
   - **B.** Training on public domain materials
   - **C.** Using copyrighted images without permission
   - **D.** Using statistical summaries of datasets
2. Under U.S. law, which of the following is a protected characteristic in employment discrimination cases involving AI tools?
   - **A.** Internet usage history
   - **B.** Political affiliation
   - **C.** Education level
   - **D.** Race
3. What is a key principle behind the FTC's approach to AI under Section 5 of the FTC Act?
   - **A.** AI is not subject to consumer protection laws.
   - **B.** Only AI used for profit is regulated.
   - **C.** Undisclosed use of AI may be considered deceptive.
   - **D.** AI disclosures must be approved by a federal judge.
4. In product liability law, what type of defect occurs when a product's original design is inherently unsafe?
   - **A.** Marketing defect
   - **B.** Manufacturing defect
   - **C.** Design defect
   - **D.** Distribution defect
5. Which legal framework prohibits discrimination in loan decisions made by AI?
   - **A.** Computer Fraud and Abuse Act (CFAA)
   - **B.** Equal Credit Opportunity Act (ECOA)
   - **C.** Digital Millennium Copyright Act (DMCA)
   - **D.** Sherman Antitrust Act

6. A chatbot giving unlicensed medical advice without disclosure most likely violates which type of law?
   - **A.** Trade secret law
   - **B.** Employment law
   - **C.** Consumer protection law
   - **D.** Patent law

7. Which of the following is not typically a valid defense against a product liability claim involving AI?
   - **A.** The AI was too complex to predict.
   - **B.** The product was used as intended.
   - **C.** A safer design was available.
   - **D.** The product caused foreseeable harm.

8. Which regulatory body issued guidance clarifying employer obligations when using AI tools in hiring?
   - **A.** Federal Reserve
   - **B.** National Labor Relations Board
   - **C.** Equal Employment Opportunity Commission (EEOC)
   - **D.** U.S. Copyright Office

9. What is a key risk of using proxy variables in AI decision-making?
   - **A.** They may have a disparate impact on protected groups.
   - **B.** They can be more efficient than direct attributes.
   - **C.** They reduce model complexity.
   - **D.** They are required by most civil rights laws.

10. Which of the following best describes the FTC's stance on AI performance claims?
    - **A.** They are allowed as long as results are plausible.
    - **B.** They require substantiation like any other product claim.
    - **C.** They are exempt from scrutiny if the AI is open source.
    - **D.** They must be approved by a third-party certifier.



Chapter

6

The EU AI Act

THIS CHAPTER COVERS AIGP DOMAIN II, "UNDERSTANDING HOW LAWS, STANDARDS AND FRAMEWORKS APPLY TO AI," SUBDOMAIN II.C, "UNDERSTAND THE MAIN ELEMENTS OF THE EU AI ACT." THE FOLLOWING TOPICS ARE COVERED:

- ✔ Risk classification levels
- ✔ Requirements for each risk classification level
- ✔ Requirements for general-purpose AI models
- ✔ Enforcement and penalties
- ✔ Requirements for providers, deployers, importers, and distributors

THE AI GOVERNANCE PROFESSIONAL (AIGP) PERFORMANCE INDICATORS IN THIS SUBDOMAIN INCLUDE:

- ✔ Understand the risk classification framework for AI (that is, prohibited AI, high-risk, limited-risk and minimal-risk) and what systems fall into each category.
- ✔ Understand the key requirements for high-risk, limited-risk and minimal-risk AI including risk management, data governance, technical documentation, conformity assessment, record keeping, human oversight, transparency and notification, quality management (as applicable).
- ✔ Understand the distinct requirements for general-purpose AI models.
- ✔ Understand the enforcement framework and penalties for non-compliance.
- ✔ Understand the differences in requirements based on organizational context (for example, providers, deployers, importers, and distributors).



The EU AI Act represents a landmark legislative framework that governs the design, deployment, and oversight of AI within the European Union. This chapter explores the Act's risk-based classification system, specific compliance requirements based on risk level, and the responsibilities that vary by organizational role. Special attention is given to general-purpose AI systems, which present unique challenges due to their broad applicability and potential systemic impact. Readers will also gain an understanding of the enforcement mechanisms and penalties for non-compliance. By the end of this chapter, you will be equipped with a clear grasp of how the EU AI Act applies across different AI systems, contexts, and stakeholder roles.

Risk Classification

The EU AI Act introduces a structured framework for classifying AI systems based on the risk they introduce to fundamental rights, safety, and societal values. This risk-based approach emphasizes the regulation's rationale and dictates the obligations AI developers, providers, and deployers must follow. This section explores the EU AI Act's four-tier classification: *prohibited AI, high-risk AI, limited-risk AI,* and *minimal-risk AI.* Each classification level has corresponding requirements and constraints. Understanding these categories is essential for compliant and ethical AI deployment within the EU and potentially beyond.



Organizations outside the EU are adopting the EU AI Act classification scheme, which is rational and straightforward. Organizations that need to develop a classification scheme are advised to start with the EU AI Act's scheme and tailor it as required.

Prohibited AI Systems

The EU AI Act's risk classification scheme includes one category of AI systems that are not allowed.

Definition and Legal Basis

Prohibited AI refers to use cases that pose an unacceptable risk to safety, human dignity, or fundamental rights and are therefore *banned outright* under the EU AI Act. These systems are *not allowed* to be placed on the market, put into service, or used anywhere within the European Union.

Examples of Prohibited AI

These are examples of AI systems not permitted in the EU:

- **Subliminal manipulation:** AI systems that manipulate behavior in ways that bypass consciousness, such as neuro-influencing technologies.
- **Social scoring by governments:** Systems that rank individuals based on behavior or personality traits, leading to discriminatory outcomes.
- **Exploiting vulnerable groups:** AI systems that target children, the elderly, or disabled individuals to distort their or others' behavior.
- **Real-time remote biometric identification in public spaces:** With limited exceptions, such as for serious crimes and under judicial oversight.
- **Individual criminal offense risk assessment or prediction.**
- **Emotional recognition in workplaces and educational institutions:** Explicitly banned due to risks of manipulation, discrimination, and pseudoscience.

Exceptions to the Biometric Identification Ban

Some real-time biometric identification may be permitted under strict conditions, such as national security threats or ongoing criminal investigations, but must be individually authorized by a legal authority.

Implications for Developers

AI system developers (referring to vendor organizations, not engineers) must ensure their system design, training data, and use cases do not cross into prohibited territory. The EU's approach reflects a precautionary stance toward AI deployments that could irreversibly harm individuals or society.

High-risk AI Systems

The EU AI Act's classification scheme includes three levels of AI systems that are permitted. High-risk AI is the highest risk of the three.

Definition and Criteria

High-risk AI systems are those that have the potential to significantly affect people's rights, lives, or safety, and thus are subject to stringent regulatory controls. These include systems used in regulated sectors or those that influence critical decisions.

Categories of High-risk AI (Annex III of the Act):

- Biometric identification and categorization
- Critical infrastructure management (for example, energy grids, wastewater management, and traffic control)
- Education and vocational training (for example, AI grading tools, lesson planning)
- Employment, worker management, and access to self-employment (for example, AI interviews, AI-driven performance evaluations)
- Access to and enjoyment of essential private and public services (for example, credit scoring, welfare benefits)
- Law enforcement (for example, crime prediction tools)
- Migration, asylum, and border control
- Administration of justice and democratic processes

Compliance Requirements:

- ***Conformity assessments*:** Organizations must demonstrate compliance before placement on the market.
- **Risk management and data governance:** Developers must maintain data quality, minimize bias, and manage operational risks.
- **Human oversight:** Systems must allow for meaningful human intervention and escalation.
- **Technical documentation and logging:** Lifecycle documentation and automatic logging must be maintained.

CE Marking for AI

Like other product safety marks, high-risk AI systems must bear the CE mark to indicate conformity with EU standards. All requirements will apply.

For example, an employment recruiting platform that uses AI to filter candidates based on prior employment or behavioral profiles would be considered high-risk under the employment category.

Limited-risk AI Systems

The EU AI Act's classification scheme includes three levels of AI systems that are permitted. Limited-risk AI is the mid-level risk of the three.

Definition

Limited-risk AI refers to AI systems that interact with users but do not significantly impact safety or rights.

Examples of limited-risk AI include:

- Chatbots and virtual assistants
- AI-generated media (deepfakes, images, or synthetic voices)
- Educational tools (AI-based learning assistants)
- Social media filters/avatars

These systems impose minimal obligations focused mainly on transparency.

Transparency Obligations

- **Disclosure of AI use:** Users must be informed when they are interacting with an AI system.
- **Synthetic media disclosure:** Deepfakes and other generated content must be labeled as such.
- **User autonomy:** Users should have the opportunity to opt out or refuse certain AI interactions where reasonable and feasible.

For instance, a chatbot providing customer support must disclose that it is not a human. Similarly, AI-generated images used in advertising must be clearly labeled as synthetic.

Dark Patterns in Limited-risk Systems

Even though these systems carry limited obligations, the use of dark patterns or manipulative UI in AI interfaces may trigger scrutiny under consumer protection laws, which are separate from the EU AI Act and the GDPR. Dark patterns are discussed in Chapter 4.

Minimal-risk AI Systems

The EU AI Act's classification scheme includes three levels of AI systems that are permitted. Minimal-risk AI is the lowest-level risk of the three levels of permitted AI systems.

Definition

Minimal-risk AI systems are those AI systems that pose negligible or no risk to users or society. They represent the majority of AI systems and are largely unregulated under the EU AI Act.

Examples of minimal-risk AI systems include:

- Spam filters
- Recommendation engines (for example, video streaming found in streaming services like Netflix, YouTube, and Amazon Prime)
- Productivity tools (for example, AI-enabled grammar checkers)

Encouraged Practices

While not mandated, developers of minimal-risk systems are encouraged to follow voluntary codes of conduct, ethical guidelines, and transparency practices, particularly when their systems are scaled or widely deployed. AI regulations could change in the future and include additional scrutiny for AI systems currently considered minimal risk.

Risk Reclassification and Monitoring

The EU AI Act anticipates that some AI systems may evolve over time or be repurposed. Systems initially categorized as minimal- or limited-risk may become high-risk depending on their use context. Continuous monitoring and reassessment are advised.

Table 6.1 provides a side-by-side comparison of the EU AI Act risk classification scheme.

This classification system is the heart of the EU AI Act's risk-based approach. AI governance personnel must understand the thresholds and boundaries between each tier, which is essential for ensuring legal compliance, managing reputational risk, and building trustworthy AI.

TABLE 6.1 EU AI Act Risk Classifications

Risk Category	Definition	Examples	Regulatory Requirements
Prohibited AI	Poses unacceptable risk to rights/safety	Social scoring, subliminal manipulation	Completely banned
High-risk AI	Impacts fundamental rights or safety	Credit scoring, AI hiring tools	Conformity assessments, logging, oversight
Limited-risk AI	Requires transparency but limited impact	Chatbots, synthetic media	Disclosure, opt-out mechanisms
Minimal-risk AI	Poses negligible risk	Spell-checkers, streaming recommendations	No requirements (voluntary best practices)

Classification Requirements

The obligations stipulated for each AI risk classification under the EU AI Act are not merely labels; they represent specific operational and compliance expectations tailored to the system's risk profile. This section explores the requirements for high-risk, limited-risk, and minimal-risk AI systems, with particular emphasis on operational controls such as risk management, data governance, documentation, conformity assessments, transparency measures, and oversight. These requirements define how AI systems must be designed, validated, deployed, and monitored.

You may recall that in the previous section, four risk categories were described. However, there are no requirements for the *prohibited AI* classification, as these systems are not allowed at all.

High-risk AI: Full Compliance Framework

High-risk AI is the highest-risk classification for AI systems and, hence, has the most stringent requirements.

Readers may believe that the whole body of these requirements is onerous. However, these represent similar requirements found in mature systems' development lifecycle frameworks. Put another way, if you are shocked at the amount of paperwork required for high-risk AI systems, then perhaps you are not well-versed in mature systems' development frameworks.

Risk Management System

Developers of high-risk AI systems must implement a documented and ongoing risk management system. This system must:

- Proactively identify known and foreseeable risks related to the AI system's intended use and reasonably foreseeable misuse
- Evaluate and monitor risks throughout the system lifecycle
- Implement risk control measures to eliminate or reduce identified risks

Several good risk management frameworks are available for organizations needing to implement or improve their risk management practice, including:

- NIST SP 800-37 ("Risk Management Framework for Information Systems and Organizations")
- ISO/IEC 27001 ("Information Security, Cybersecurity, and Privacy Protection—Information Security Management Systems—Requirements")
- ISO/IEC 31000 ("Risk Management—Guidelines")
- ISO/IEC 23894:2023 ("Information Technology—Artificial Intelligence—Guidance on Risk Management")

Note that the legal guidance in these NIST and ISO standards is helpful but not a substitute for the EU AI Act's explicit requirements.

Data and Data Governance

High-risk AI systems must be trained, validated, and tested on datasets that are high-quality, relevant, representative, and as free from errors and bias as possible.

Requirements include:

- Bias mitigation strategies
- Data relevance to the intended purpose
- Data lineage documentation and version control

Organizations training AI systems with personal information must also comply with all applicable privacy laws, including the EU GDPR. All privacy-related requirements must be implemented in AI systems.

Organizations lacking a sound data governance function must first implement data governance prior to implementing high-risk AI systems. Organizations needing guidance may refer to these resources:

- *Data Management Body of Knowledge (DMBOK)*, which defines data governance principles, roles and responsibilities, policies and procedures, and a data governance maturity model.
- *COBIT (formerly, "Control Objectives for Information and Related Technology," now just COBIT)*, which addresses business alignment, risk management, performance measurement, compliance, and legal requirements.
- *Data Governance Institute (DGI) Framework*, which contains four components: (1) value statement, goals, and metrics; (2) data rules, decision rights, and accountability; (3) controls and standards; and (4) people and processes.

Consulting firms, including SAS, BCG, PWC, and Eckerson, have also published data governance frameworks.

Data governance is discussed in detail in Chapter 9.

Technical Documentation

The AI system provider must maintain detailed and formally managed documentation to demonstrate conformity to all applicable portions of the EU AI Act, including:

- System architecture and specifications
- Algorithms and training methodologies
- Risk management and human oversight protocols

To avoid overlap and duplicity, organizations should develop their technical documentation to meet all applicable regulatory requirements.

Record keeping and Logging

Organizations must create and maintain logs and other records throughout the lifecycle of the AI system. These logs support:

- Post-market monitoring
- Event and incident investigation
- Audit readiness and accountability
- Change management at the business process and system levels

Conformity Assessment

Before being placed on the market, organizations must undergo:

- Internal control (self-assessment) in some cases
- Third-party assessment by a *Notified Body* in other cases
- CE marking once conformity is confirmed

Documentation and records for these activities must be captured and maintained.

Human Oversight

AI systems must be designed to allow for:

- **Detection of anomalies or signs of malfunction:** Call this disaster recovery planning if you must: if an AI system (or any other information system) begins to malfunction, it must be taken offline and fixed.
- **Effective override or shut-off mechanisms:** When situations warrant, organizations must be able to pause the use of an AI system experiencing problems, without adversely impacting critical business processes.
- **Human-in-the-loop decision review where necessary:** Examples include AI-driven loan approval, employment screening, medical diagnosis, and more.

Transparency and User Instructions

Users of an AI system must receive clear instructions for its use, including:

- The system's intended purpose and limitations
- Human oversight responsibilities
- Performance metrics and conditions of use (accuracy levels, expected error rates, environmental or data quality requirements)
- Contact points for reporting incidents or questions

Quality Management System

Organizations implementing high-risk AI systems must implement a *quality management system (QMS)*, which must include:

- Policies and procedures for design, development, testing, and control
- Corrective and preventive action procedures
- Documentation control
- Processes for monitoring, reporting, and logging during operation and post-market surveillance
- Resource and training management to ensure competence

ISO 9000 ("Quality Management Systems—Requirements") is the definitive industry standard for building a quality management system.

Lifecycle Documentation Responsibilities

Developers of high-risk AI systems must maintain documentation not only pre-market but throughout the system's use. Required documentation includes updates to the system, retraining events, and downstream adaptations by deployers.

Limited-risk AI: Targeted Transparency Obligations

Limited-risk AI is the mid-tier-risk classification for AI systems and, hence, has less stringent requirements than those of high-risk AI.

Disclosure Requirements

AI providers must ensure that the users of limited-risk AI systems are informed:

- When interacting with AI systems (for example, chatbots and virtual assistants)
- When viewing synthetic or manipulated content (for example, deepfakes, images, and text)

Notification and User Autonomy

Limited-risk AI systems must support:

- Conspicuous notifications that a system is AI-driven
- Easily found means for users to opt out or switch to human interaction where appropriate
- Transparent instructions on system use and limitations so users understand capabilities and boundaries

Documentation and Accountability

Though not subject to the full documentation burden of high-risk systems, limited-risk AI developers should:

- Maintain records of how transparency mechanisms are implemented
- Monitor for misleading outputs that could impair user autonomy

For instance, a retail site using an AI-driven stylist chatbot must notify users they are interacting with AI and offer a human option where feasible.

Regulatory Intersection

While limited-risk systems are lightly regulated under the EU AI Act, they may still fall under broader EU privacy and/or consumer protection laws, especially when personal information or misleading/manipulative design choices are involved.

Minimal-risk AI: Voluntary Best Practices

No mandatory requirements apply to minimal-risk AI systems. However, developers of minimal-risk AI systems are encouraged to provide simple notices where AI functionality may affect user decisions, avoid misleading or manipulative design patterns, and maintain internal logs and quality checks.

AI systems posing negligible risk should still consider accessibility accommodations for diverse users and avoid reinforcing harmful stereotypes or disinformation.

For example, an AI-enhanced grammar editing tool may voluntarily include a disclosure that suggestions are AI-generated, promoting trust without regulatory pressure.

Minimal Risk Today, but What About Tomorrow?

Minimal-risk systems deployed at scale or adapted to new contexts (for example, education or public policy) could face reclassification in the future. Voluntary practices now (especially, documenting the details of the design, development, and training of an AI system) may reduce compliance costs later.

Understanding classification-specific requirements enables organizations to implement proportionate governance, ensuring that regulatory burdens scale with potential harm.

Hence, it minimizes overregulation for low-impact tools while applying full safeguards to high-risk systems.

Avoid Checkbox Compliance Through Business Alignment

Numerous organizations in every industry express a minimalist attitude toward compliance, known as "checkbox compliance," in which leadership strives to achieve only the minimum stated requirements in the least costly ways. Often, this results in the mere *appearance* of compliance that lacks true risk reduction.

A healthier approach involves pursuing the *spirit* and *intention* of stated requirements, while also considering what additional measures the organization should implement to align with its mission, objectives, and culture.

Requirements for General-purpose AI

As general-purpose AI (GPAI) systems grow in scale and impact, the EU AI Act imposes a tailored set of requirements to address their systemic influence. Unlike narrowly scoped AI tools, GPAI systems can serve multiple downstream applications, including high-risk use cases. Their complexity, versatility, and general availability introduce unique governance challenges. This section outlines the specific legal, technical, and operational requirements that apply to GPAI developers and deployers, with distinctions based on whether the model is open-source, commercial, or designated as having systemic risk.

Defining General-purpose AI Under the EU AI Act

Legal Definition

The EU AI Act defines *general-purpose AI (GPAI)* systems as models that can serve multiple intended purposes and be integrated into a variety of downstream systems or applications. These models are not tied to a single task or deployment context. Examples of GPAI systems include:

- Large language models (LLMs), such as the GPT-family, Llama, and Claude
- Foundation models for vision tasks, such as CLIP and SAM
- Multimodal models capable of interpreting text, images, video, and audio

Foundation Model vs. General-purpose AI

While often used interchangeably, "foundation model" refers to the base technical artifact, while "general-purpose AI" refers to its usage and availability across diverse applications.

General Obligations for GPAI Providers

GPAI providers are required to develop documentation, evaluate and test their AI models, and practice data governance, as detailed in this section.

Technical Documentation and Transparency

All GPAI providers must prepare comprehensive technical documentation, including:

- Model architecture, training data sources, and capabilities
- Intended and foreseeable uses
- Known limitations, biases, and failure modes

These providers must also disclose whether their GPAI model was fine-tuned on sensitive data, as well as its mechanisms for risk identification and mitigation, monitoring, and deactivation.

Model Evaluation and Testing

GPAI providers must perform and document risk assessments that evaluate:

- AI model accuracy and robustness across various tasks
- Bias and fairness across demographics
- Environmental impact (energy consumption)

Where GPAI is expected to be integrated into high-risk applications, additional testing and safeguards are required, as stipulated in the EU AI Act's risk classification scheme.

Data Governance

Though GPAI providers are not required to fully disclose raw training datasets, providers must document training dataset composition and provenance, and mitigate copyright risks, personal data inclusion, and representational bias. Data governance is discussed in detail in Chapter 9.

Transparency Summary Sheet

The EU AI Act encourages the creation of a standardized GPAI model card or factsheet that summarizes key characteristics and risks for downstream users. As stated elsewhere in this chapter, organizations should consider fully documenting the development and training of their AI systems, whether required by regulations or not.

Additional Obligations for GPAI with Systemic Risk

According to the EU AI Act, GPAI models with systemic risk have additional compliance obligations.

What Qualifies as Systemic Risk?

A GPAI system is considered to pose *systemic risk* if it:

- Has exceptionally high compute requirements (for example, $>10^{25}$ FLOPs; not a hard limit but an illustration of scale)
- Is integrated into critical infrastructure or widely deployed across industry sectors

or

- Has demonstrated ability to autonomously influence behavior at scale

Enhanced Requirements

GPAI providers whose systems meet systemic risk criteria must:

- Submit to external audits
- Provide real-time access to model behavior—under certain conditions
- Publish adversarial testing results
- Notify EU regulators of significant changes or incidents
- Publish transparency summaries about capabilities, limitations, and risks, so downstream deployers and the public understand potential impacts

Governance and Accountability

GPAI providers with systemic risk AI systems must:

- Establish an EU-based legal liaison
- Maintain incident response and escalation mechanisms
- Provide APIs and documentation for third-party safety researchers

Comparison: Standard GPAI vs. Systemic Risk GPAI

Requirement	Standard GPAI	GPAI with Systemic Risk
Technical documentation	Yes	Yes
External audits	No	Mandatory
Adversarial testing publication	Encouraged	Required
Regulator notification	No	Required
Legal representative in the EU	No	Mandatory

Obligations for Distributors and Deployers of GPAI

Organizations that integrate GPAI into downstream systems, especially high-risk applications, are required to:

- Verify the model's conformity with published documentation
- Assess their own use case for potential elevation in risk classification
- Ensure transparency toward end users

For example, an EU-based bank integrating a GPAI model into its loan underwriting platform must verify:

- The accuracy and completeness of its model documentation and risk profile
- That human oversight is maintained
- That it complies with all high-risk system obligations if triggered

Open-source GPAI Considerations

Open-source GPAI developers may be exempt from some obligations if they do not derive commercial benefit or maintain operational control of the systems they develop. However, deployers remain responsible for downstream compliance.

Encouraged Practices Across All GPAI Models

Even when obligations are not mandatory, the EU AI Act encourages the adoption of voluntary codes of conduct that address accessibility, inclusiveness, environmental sustainability, and misuse prevention (for example, election manipulation, malware generation, or fraud). Further, the Act encourages GPAI developers to collaborate with academia and civil society and support open benchmarks and transparency initiatives.

The versatility of GPAI necessitates a distinct governance approach compared to fixed-function AI tools, as a GPAI system can be adapted to various environments and support diverse use cases, including some that may carry high risks. By regulating GPAI through both general and risk-triggered provisions, the EU AI Act aims to ensure these powerful systems are developed and deployed responsibly, even when their applications are not yet fully known.

Enforcement and Penalties

Regulations lacking consequences of non-compliance have little value. The EU AI Act establishes robust enforcement to ensure compliance with its provisions across all risk tiers, contexts, and organizational roles. Enforcement is designed to be proportionate to the level of risk posed by a system and the severity of the non-compliance. This section explains how the regulation will be monitored and enforced by EU and national authorities, and describes the penalties organizations may face if they fail to comply.

Enforcement under the EU AI Act reflects a balance between innovation and accountability. By establishing strong supervisory powers and deterrent penalties, the Act reinforces its risk-based approach, by holding developers and providers more accountable where AI use presents greater danger to rights, safety, or democracy.

Oversight and Supervisory Authorities

The EU AI Act provides for two levels of enforcement authorities—the European AI Office and the member states.

The EU AI Act establishes a centralized body known as the European AI Office, housed within the European Commission. It is required to coordinate the supervision of GPAI systems and models with systemic risk, maintain a public database of high-risk AI systems, and issue guidance and best practices across member states.

Each member state is required to designate one or more national competent authorities to conduct market surveillance, handle complaints from individuals and civil society, investigate breaches, and impose penalties within the member state's jurisdiction.

Coordination Between the European Union and Member States

While enforcement is nationally driven, the European AI Office plays a central harmonization and supervisory role, particularly for GPAI and cross-border AI deployments.

Inspection Powers and Remedies

As a part of market surveillance, authorities may require access to technical documentation, perform unannounced inspections (including digital audits), mandate and verify corrective actions, and suspend or prohibit marketing of AI systems.

Depending on the severity of audit results or a breach, authorities can require that non-compliant systems be withdrawn or recalled. They may also direct AI developers or deployers to perform updates, retrain their systems, or label system changes. They may require providers to notify affected users.

Individuals affected by non-compliant AI systems may be able to seek remedies under national civil liability frameworks, or file complaints through data protection or consumer protection channels.

Whistleblower Protections

The EU AI Act encourages reporting of systemic non-compliance and may include provisions to protect whistleblowers within AI organizations, similar to provisions in GDPR and other EU laws.

Penalties for Non-compliance

The EU AI Act introduces a tiered penalty framework of administrative fines as follows:

- Up to €35 million or 7% of global annual turnover (whichever is higher) for violations of banned practices (for example, prohibited AI)
- Up to €15 million or 3% of turnover (whichever is higher) for non-compliance with obligations for high-risk systems
- Up to €7.5 million or 1% of turnover (whichever is higher) for supplying incorrect, incomplete, or misleading information to authorities

Penalties Compared to GDPR

The EU AI Act's penalties resemble those under GDPR but may exceed them for the most egregious violations involving prohibited AI or systemic risk GPAI.

For example, if a provider of an AI-based hiring tool fails to implement required bias mitigation or human oversight, and this leads to discriminatory outcomes, they could face fines of up to €15 million, plus a mandatory recall of the system.

Escalation and Recurrence

The EU AI Act includes provisions for compliance issues that are not easily solved.

In assessing penalties, authorities will consider:

- Nature and significance of the breach
- Whether there was a matter of non-compliance, and, if so, whether intentional or negligent
- Any previous history of violations
- The degree of cooperation with authorities

After an enforcement action, organizations subject to penalties may also be required to submit to third-party audits, implement compliance remediation plans, and report on remediation progress periodically.

Voluntary Disclosure Benefits

Organizations that voluntarily disclose incidents or deficiencies and cooperate with authorities may receive reduced penalties or avoid punitive sanctions.

Cross-border Enforcement and Cooperation

The EU AI Act includes stipulations related to enforcement matters that involve two or more member states, as well as developers and providers outside the European Union.

Mutual Assistance Framework

The EU AI Act directs national authorities to assist each other when enforcement matters cross state boundaries. National authorities will cooperate through:

- Sharing information on investigations
- Supporting cross-border enforcement actions
- Coordinating these and other activities through the European AI Office

Jurisdictional Clarity for GPAI Providers

GPAI developers outside the EU are required to appoint an EU-based legal representative to represent the organization, wherever it is headquartered. GPAI developers are also required to ensure all systems placed on the EU market meet all applicable compliance standards.

Extraterritoriality

The EU AI Act applies not only to organizations based in the EU but also to non-EU entities if they place AI systems on the EU market or make AI systems available to users located in the EU. This includes AI systems and general-purpose AI systems marketed or used in the European Union, and deployers outside the European Union if the system affects EU persons. The regulation also affects GPAI developers outside the European Union when their models are integrated into EU-facing applications (even if those applications are not physically in the European Union).

Non-EU providers are required to appoint an EU-based legal representative, ensure all conformity assessments are in place, and meet all transparency and documentation requirements based on the system's classification.

Organizational Context

Compliance with the EU AI Act depends not only on the type of AI system in question but also on the role an organization plays in the AI supply chain. Whether an entity is a provider, deployer, importer, or distributor significantly shapes its legal obligations. This section explains the distinct requirements for each role, highlighting how organizational context drives accountability under the regulation.

AI Providers: Core Compliance Burdens

A *provider* is any natural or legal person or entity that develops an AI system or GPAI model intending to place it on the EU market, under its own name or brand.

The primary obligations of providers include:

- Conduct and document risk management procedures
- Complete and document conformity assessments
- Maintain technical documentation and logs
- Apply CE marking (for high-risk AI systems)
- Register the system in the EU AI database
- Ensure post-market monitoring and incident reporting

Providers of GPAI vs. Narrow AI

GPAI providers face additional, distinct obligations, including model transparency, systemic risk analysis, and documentation of capabilities and limitations, beyond what traditional narrow AI providers must do.

An example company that trains and sells a facial recognition model for commercial use is a provider and bears full responsibility for documentation, risk controls, and transparency.

Deployers: Operational Accountability in Practice

A *deployer* is a person or organization that uses AI systems in the context of their professional activities. This includes public and private organizations applying or installing AI systems developed by others. Think of a deployer as a person, team, or organization that deploys an AI system *for its professional activities* (not for private use).

Obligations of deployers vary by the type of AI system.

High-risk AI:

- Ensure human oversight is applied during system use
- Monitor system performance and report serious incidents
- Ensure input data quality and appropriateness for the local context
- Maintain system logs and usage records
- Use the system according to provider instructions

GPAI Integration:

- Validate documentation and test for compatibility
- Consider the role in risk reclassification if adapting or repurposing the AI

When Deployers Become Providers

If a deployer substantially modifies an AI system or develops its own downstream application using GPAI, it may legally be reclassified as a provider and inherit those obligations.

An example of a deployer is a hospital deploying an AI-based diagnostic tool, which must ensure that oversight protocols are in place and that the system performs safely in the clinical setting.

Importers: Bridging the Legal Jurisdictions

An *importer* is an organization that places AI systems from non-EU countries onto the EU market. They act as the legal intermediary between a foreign provider and EU-based regulators.

Obligations by importers include:

- Verify the system's CE marking and conformity assessment
- Ensure technical documentation is available, correct, and complete
- Register the AI system in the EU database (unless done by the provider)
- Cooperate with authorities on compliance checks

Importer Risk Exposure

Importers are legally liable if they place a non-compliant AI system on the market. Contracts between importers and providers should clarify responsibility boundaries.

An example German-based tech reseller imports a U.S.-developed education AI tool into the EU market. The reseller, in the role of an importer, must ensure that all documentation, registrations, and regulatory markings are in place.

Distributors: Passive but Not Exempt

Distributors make AI systems available in the EU market without modifying them or acting as importers. Examples include app stores, platform resellers, and marketplaces.

The obligations of distributors per the EU AI Act include:

- Verify CE marking is present
- Confirm provider/importer contact information is visible
- Suspend distribution if non-compliance is suspected
- Cooperate with authorities during investigations

An example e-commerce platform listing AI-powered home devices is a distributor according to the EU AI Act. The organization must take down listings if it receives credible evidence of non-conformity.

Distributors, Notifiers, and Whistleblowing

Distributors are encouraged (but not required) to establish mechanisms for consumers or competitors to report potential non-compliance or AI risks.

Shared and Overlapping Responsibilities

Let's take a brief look at situations of shared and overlapping responsibilities.

Horizontal Duties

Across all roles, organizations are required to:

- Cooperate with national competent authorities and regulators
- Respond to incident reports
- Maintain records (logs, updates, notices)

AI-as-a-Service and Role Ambiguity

Cloud-based AI platforms can blur roles:

- Whether the cloud provider is the deployer, importer, distributor, or user
- Which party performs conformity checks on modular AI services

Clarity may require contractual designation of responsibilities linked to the EU AI Act–defined roles.

Mapping Your Role in the AI Lifecycle

Organizations should conduct an "AI role assessment" to determine whether they are acting as a provider, deployer, importer, or distributor for each system—and what compliance obligations result.

Understanding the regulatory burden associated with each organizational role ensures not only compliance but also clearer accountability, governance readiness, and supply chain coordination. The EU AI Act demands that all parties, from developers to resellers, accept and document their respective responsibilities for trustworthy AI.

EU AI Act Implementation

The EU AI Act follows a phased implementation timeline, with staggered obligations and grace periods, depending upon the type of AI system and stakeholder role. Table 6.2 contains a concise breakdown.

Readers are advised to periodically review news regarding the enforcement of the EU AI Act, as these dates could change.

TABLE 6.2 Key Milestones in the EU AI Act

Date	Event/Obligation
July 12, 2024	AI Act published in the Official Journal of the EU
August 1, 2024	Entry into force—formal start of the regulation but no immediate obligations
February 2, 2025	Prohibited AI practices and AI literacy requirements begin to apply
May 2, 2025	Codes of practice for GPAI models are expected to be finalized
August 2, 2025	Significant obligations begin: GPAI rules, governance structures, penalties, notified bodies
August 2, 2025	Providers of GPAI models already on the market must comply by August 2, 2027
August 2, 2026	High-risk AI system requirements take effect (for example, conformity assessments, CE marking)
By 2027	Full enforcement expected across all applicable provisions

Here are some of the grace periods and transitional provisions:

- **GPAI providers:** If a GPAI model was placed on the market before August 2, 2025, the provider has until August 2, 2027, to achieve full compliance.
- **Codes of practice:** While voluntary, signing the GPAI Code of Practice may offer legal assurances. Non-signatories won't benefit from presumptions of conformity.
- **Member states:** They must designate enforcement authorities and establish penalty frameworks by August 2, 2025.

Meta and the GPAI Code of Practice

In July 2025, Meta announced that the company would not sign the GPAI Code of Practice, citing legal uncertainties, regulatory overreach, and concerns over AI innovation. Other AI providers, including OpenAI and Mistral AI, have agreed to sign the Code.

As of the writing of this book, this matter continues to develop.

Summary

The EU AI Act introduces a tiered risk-based framework to govern the development and use of AI systems, classifying them into four categories: prohibited, high-risk, limited-risk, and minimal-risk. Each category corresponds to the level of harm the system could pose to fundamental rights, safety, or democratic values. *Prohibited AI* systems include practices such as social scoring by governments or subliminal manipulation, and are banned outright. *High-risk AI* systems (those affecting critical areas like employment, education, or law enforcement) are subject to the most stringent controls, including conformity assessments, documentation, oversight, and CE marking. *Limited-risk AI* systems, such as chatbots and deepfakes, must meet transparency requirements. *Minimal-risk AI* systems like spam filters and grammar checkers are largely unregulated, although they are encouraged to document their development, management, and operational processes.

Each risk tier comes with its own regulatory requirements. *High-risk AI* must implement robust risk management protocols, ensure high-quality training data, maintain comprehensive technical documentation, and support meaningful human oversight. These systems also need a quality management system and post-market monitoring. *Limited-risk AI* systems are primarily bound by transparency and notification duties to inform users that they are interacting with AI. For *minimal-risk AI* systems, the Act encourages (but does not mandate) voluntary good practices around transparency and ethics. Across all tiers, the regulation encourages proportionality and lifecycle thinking.

General-purpose AI (GPAI) systems, such as large language models, receive special, additional regulatory treatment under the EU AI Act. GPAI providers must disclose model architecture, training data practices, known limitations, and intended uses. They are also expected to conduct testing for robustness, fairness, and environmental impact. Models deemed to pose systemic risk (based on scale, impact, or societal influence) must meet elevated standards, including external audits, adversarial testing publication, and regulator notification. The Act also imposes obligations on organizations that deploy or integrate GPAI into downstream systems, especially when such use results in a high-risk or limited-risk designation.

The AI Act distinguishes obligations based on an entity's organizational role, whether it is a provider, deployer, importer, or distributor. Providers bear the greatest responsibility, including conducting risk assessments, completing conformity checks, and ensuring post-market compliance. Deployers must apply oversight, manage system performance, and respond to adverse outcomes. Importers ensure foreign systems meet EU compliance before entering the market, and distributors must act if they become aware of non-compliance. These distinctions clarify accountability across the AI lifecycle and help ensure supply chain integrity. An organization can have more than one role; for instance, an organization can be both an importer and a distributor.

The EU AI Act includes a comprehensive enforcement framework. The European AI Office leads centralized supervision, especially for GPAI, while national authorities handle inspections, market surveillance, and penalties. Fines for non-compliance are tiered based

on the nature and severity of violations, ranging up to €35 million or 7% of global revenue (whichever is greater) for prohibited practices. Authorities may suspend or recall systems, require changes, or mandate user notification. Entities found in breach may also face ongoing monitoring or third-party audits. The Act encourages voluntary disclosure and cooperation, offering potential penalty mitigation as an incentive.

Organizations should assume that the EU AI Act, as well as other applicable AI regulations, will change over time, imposing new or different requirements on organizations in their various AI supply chain roles.

Exam Essentials

AI systems are classified into four risk tiers: prohibited, high-risk, limited-risk, and minimal-risk. This classification scheme determines the level of regulatory control imposed. Prohibited systems are banned outright due to their potential to cause unacceptable harm. High-risk systems are subject to the most stringent obligations, while limited- and minimal-risk systems face more targeted or voluntary requirements.

High-risk AI systems must meet detailed compliance obligations before market placement. Requirements include conformity assessments, technical documentation, risk management procedures, human oversight, post-market monitoring, and business records. The goal is to ensure safety, reliability, and accountability throughout the AI lifecycle.

General-purpose AI (GPAI) systems face unique requirements based on scale and systemic impact. GPAI providers must document system architecture, disclose intended uses and risks, and conduct model evaluations. If the system poses systemic risk, enhanced obligations such as external audits and regulator notification apply. Requirements for GPAI systems are in addition to the requirements corresponding to the system's risk classification (high-risk, limited-risk, and minimal-risk).

The role an organization plays (provider, deployer, importer, or distributor) determines its legal responsibilities. Providers carry the heaviest burden, but deployers must ensure safe implementation and oversight. Importers verify compliance for non-EU systems, and distributors must act upon credible non-compliance alerts. Clear role definitions are essential to fulfilling obligations.

Penalties under the EU AI Act can reach up to €35 million or 7% of global turnover. The Act establishes a tiered enforcement system, with penalties calibrated to the type and severity of the violation. Prohibited practices incur the highest fines, but even documentation failures or non-cooperation can result in substantial financial penalties and mandatory corrective action.

Review Questions

The following questions are designed to test your understanding of this chapter's material. For more information on how to obtain additional questions, see this book's Introduction. You can find the answers in Appendix A.

1. Under the EU AI Act, which of the following is considered a prohibited AI practice?
   - **A.** AI used in autonomous vehicles
   - **B.** AI used in spam filtering
   - **C.** Real-time biometric identification in public spaces (without exceptions)
   - **D.** AI used in personalized shopping recommendations
2. What is the key documentation requirement for high-risk AI systems under the EU AI Act?
   - **A.** Patent registration
   - **B.** Social impact assessment
   - **C.** Technical documentation covering design, risk, and oversight
   - **D.** Environmental footprint report
3. Which type of organization is responsible for placing an AI system on the EU market under its own name?
   - **A.** Importer
   - **B.** Deployer
   - **C.** Distributor
   - **D.** Provider
4. What is a key obligation for deployers of high-risk AI systems?
   - **A.** CE marking
   - **B.** Human oversight during use
   - **C.** Notifying customs authorities
   - **D.** Conducting environmental testing
5. Which of the following AI system categories is mostly exempt from mandatory regulation?
   - **A.** GPAI with systemic risk
   - **B.** Minimal-risk AI
   - **C.** High-risk AI
   - **D.** Limited-risk AI

6. What triggers enhanced requirements for a GPAI system?
   - **A.** Use of open-source code
   - **B.** Use in a single domain
   - **C.** Systemic risk designation
   - **D.** Training with European data
7. Importers of AI systems must ensure that the system:
   - **A.** Includes a software license
   - **B.** Is CE marked and has passed conformity assessment
   - **C.** Is developed in an EU member state
   - **D.** Contains only anonymized data
8. What is the maximum financial penalty for prohibited AI practices?
   - **A.** €1 million
   - **B.** €7.5 million
   - **C.** €15 million
   - **D.** €35 million or 7% of global turnover
9. A distributor becomes legally liable under the EU AI Act if they:
   - **A.** Use AI internally
   - **B.** Manufacture hardware for AI use
   - **C.** Ignore known signs of non-compliance
   - **D.** Provide AI training datasets
10. What is the purpose of a "model card" in the context of GPAI?
    - **A.** It describes the hardware compatibility of the system.
    - **B.** It registers the system with the European Patent Office.
    - **C.** It summarizes technical specifications and known limitations.
    - **D.** It licenses the model for public use.



Chapter

7

AI Standards and Frameworks

THIS CHAPTER COVERS AIGP DOMAIN II, "UNDERSTANDING HOW LAWS, STANDARDS, AND FRAMEWORKS APPLY TO AI," SUBDOMAIN II.D, "UNDERSTAND THE MAIN INDUSTRY STANDARDS AND TOOLS THAT APPLY TO AI." THE FOLLOWING TOPICS ARE COVERED:

- ✔ OECD Principles
- ✔ NIST AI RMF and NIST ARIA
- ✔ ISO standards
- ✔ Other AI standards and frameworks

THE AI GOVERNANCE PROFESSIONAL (AIGP) PERFORMANCE INDICATORS IN THIS SUBDOMAIN INCLUDE:

- ✔ Understand the OECD Principles, framework, policies, and recommended practices for trustworthy AI.
- ✔ Understand the NIST AI Risk Management Framework and playbook (for example, the core functions, categories, and subcategories).
- ✔ Understand the NIST ARIA program for methodologies, tools, metrics, and measurements on AI safety.
- ✔ Understand the core ISO AI standards (i.e., 22989 and 42001).



The rapid advancement and deployment of artificial intelligence systems across sectors has triggered global efforts to establish trustworthy, accountable, and risk-aware governance. This chapter explores the leading AI standards and frameworks that shape these efforts. It begins with the OECD AI Principles, which offer a high-level ethical foundation adopted by dozens of countries. It then examines the NIST AI Risk Management Framework and the ARIA program, both of which provide U.S.-based technical and operational guidance for managing AI risks. The chapter continues with an in-depth review of ISO/IEC standards, including 22989 for AI terminology and 42001 for AI management systems. It concludes with an overview of additional frameworks from entities such as the G7, IEEE, UNESCO, and the Singapore government. Together, these resources provide a global framework for responsible AI development, helping professionals align their organizational practices with both emerging regulations and evolving expectations for trustworthy AI.

OECD Principles

The *OECD AI Principles*, adopted in 2019 by member countries of the OECD (Organisation for Economic Co-operation and Development), is an early set of guidelines for trustworthy AI. These principles guide the design and deployment of responsible AI systems. While not legally binding, they have influenced many national and international AI strategies and regulatory approaches, including the EU AI Act and the G7's Hiroshima Process.

You can find out more about the OECD AI Principles at `https://oecd.ai/en/ai-principles`.

Key OECD Principles for Responsible Stewardship of Trustworthy AI

These are the five principles of trustworthy AI:

- **Inclusive growth, sustainable development, and well-being:** Quoting OECD, "Stakeholders should proactively engage in responsible stewardship of trustworthy AI in pursuit of beneficial outcomes for people and the planet, such as augmenting

human capabilities and enhancing creativity, advancing inclusion of underrepresented populations, reducing economic, social, gender and other inequalities, and protecting natural environments, thus invigorating inclusive growth, well-being, sustainable development and environmental sustainability."

- **Human rights and democratic values, including fairness and privacy:** "AI actors should respect the rule of law, human rights, democratic and human-centred values throughout the AI system lifecycle. These include non-discrimination and equality, freedom, dignity, autonomy of individuals, privacy and data protection, diversity, fairness, social justice, and internationally recognised labour rights. This also includes addressing misinformation and disinformation amplified by AI, while respecting freedom of expression and other rights and freedoms protected by applicable international law."
- **Transparency and explainability:** "AI actors should commit to transparency and responsible disclosure regarding AI systems. To this end, they should provide meaningful information, appropriate to the context, and consistent with the state of art:
  - To foster a general understanding of AI systems, including their capabilities and limitations.
  - To make stakeholders aware of their interactions with AI systems, including in the workplace.
  - Where feasible and useful, to provide plain and easy-to-understand information on the sources of data/input, factors, processes and/or logic that led to the prediction, content, recommendation or decision, to enable those affected by an AI system to understand the output.
  - To provide information that enable those adversely affected by an AI system to challenge its output."
- **Robustness, security, and safety:**
  - "AI systems should be robust, secure, and safe throughout their entire lifecycle so that, in conditions of normal use, foreseeable use, or misuse, or other adverse conditions, they function appropriately and do not pose unreasonable safety and/or security risks.
  - Mechanisms should be in place, as appropriate, to ensure that if AI systems risk causing undue harm or exhibit undesired behaviour, they can be overridden, repaired, and/or decommissioned safely as needed.
  - Mechanisms should also, where technically feasible, be in place to bolster information integrity while ensuring respect for freedom of expression."
- **Accountability:** "AI actors should be accountable for the proper functioning of AI systems and for the respect of the above principles, based on their roles, the context, and consistent with the state of the art. To this end, AI actors should ensure traceability, including in relation to datasets, processes, and decisions made during the AI system lifecycle, to enable analysis of the AI system's outputs and responses to inquiry, appropriate to the context and consistent with the state of the art. AI actors should, based on their roles, the context,

and their ability to act, apply a systematic risk management approach to each phase of the AI system lifecycle on an ongoing basis and adopt responsible business conduct to address risks related to AI systems, including, as appropriate, via cooperation between different AI actors, suppliers of AI knowledge and AI resources, AI system users, and other stakeholders. Risks include those related to harmful bias, human rights including safety, security, and privacy, as well as labour and intellectual property rights."

Implementation Guidance for Government Policymakers

In addition to these five principles, the OECD offers implementation recommendations for governments, which include:

- Fostering a digital ecosystem for AI through investment and infrastructure
- Promoting a skilled workforce and inclusive education
- Ensuring public sector leadership in trustworthy AI adoption
- Facilitating international cooperation on AI research and governance

Relevance to AI Governance Professionals

AI governance professionals should consider the OECD Principles as a foundational lens for policy development. They serve as baseline expectations in cross-border collaboration and multilateral agreements, providing a helpful starting point when creating internal AI ethics policies or benchmarking compliance efforts against anticipated or emerging regulations.

NIST AI Risk Management Framework (AI RMF)

With the emergence of industry practices and regulations, ensuring responsible and trustworthy AI use has become a priority. The National Institute of Standards and Technology (NIST) released the *AI Risk Management Framework (AI RMF)* to guide organizations to manage risks associated with AI systems throughout their lifecycle. The AI RMF, along with its accompanying playbook, is a structured framework built around four core functions: Govern, Map, Measure, and Manage. Categories and subcategories support these functions to help organizations address AI-specific risks, including safety, fairness, privacy, explainability, and resilience. The accompanying AI RMF playbook offers practical implementation guidance, and organizations can tailor "profiles" to their sector, risk appetite, and use cases. This framework serves as a valuable complementary tool for risk-based AI governance.

Organizations building an AI governance function can utilize the AI RMF as their cornerstone. Organizations operating in, or offering AI systems to, the EU must still comply with the EU AI Act, as discussed in Chapter 6.

Overview of the NIST AI RMF

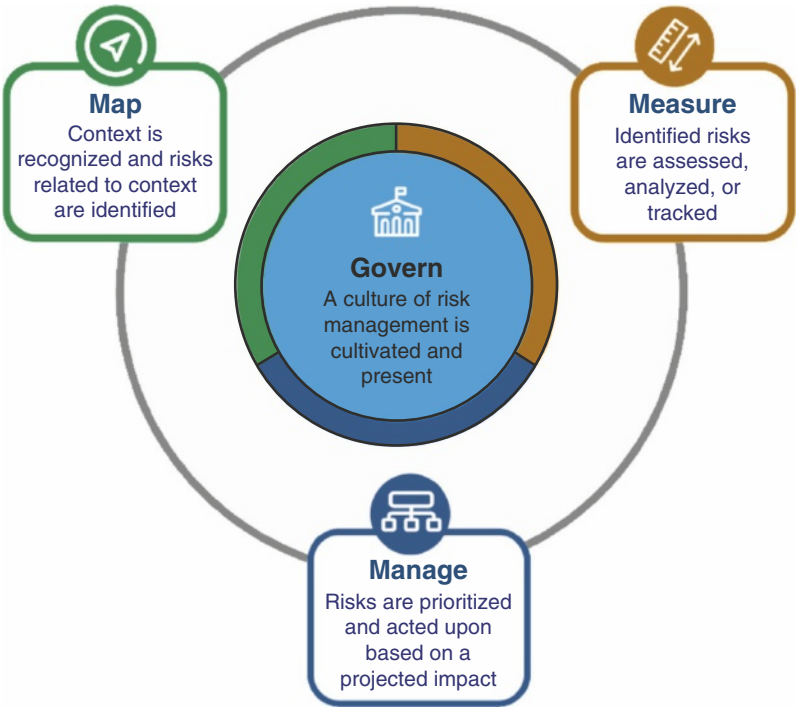
Published in January 2023, the NIST AI Risk Management Framework (AI RMF) addresses concerns about the unintended consequences of AI deployment. Unlike traditional cybersecurity or privacy frameworks, the AI RMF centers on the unique characteristics of AI risk, which can be context-dependent and difficult to quantify.

The key goals of the AI RMF are as follows:

- **Enable trustworthy and responsible AI** through structured risk management.
- **Provide a flexible, voluntary framework** suitable for a wide variety of sectors and maturity levels.
- **Foster collaboration** among developers, deployers, evaluators, and users of AI.
- **Support innovation** while proactively identifying and mitigating harm.

The AI RMF aims to establish a unified language and methodology for managing AI-related risks. It draws from NIST's extensive experience in cybersecurity and privacy, adapting those foundational concepts to the complex socio-technical landscape of AI. The AI RMF functions are depicted in Figure 7.1.

FIGURE 7.1 NIST AI RMF functions. Image from the U.S. National Institute of Standards and Technology.



What Makes AI Risk Unique?

Risk arises not only from technology but from context.

- Many AI risks come from biased data, proxies, or deployment contexts that reinforce inequality.
- Behavior may emerge that was not intended or foreseen at design time.
- Outcomes may be unpredictable, due to opaque or adaptive algorithms.
- Harms may affect individuals, groups, or society, even when performance appears optimal.
- Compliance with privacy laws presents unique challenges.
- Governance must consider data quality, design intent, system behavior, and human oversight.
- AI risks scale rapidly. Once deployed, outputs can affect millions instantly (for example, content ranking, automated decisions).

The Govern Function: Building the Foundation for AI Risk Management

The Govern function is the cornerstone of the AI RMF, and it's most relevant for those responsible for implementing and managing AI governance. It establishes the organizational context, culture, and structures necessary to manage AI risks effectively. Without strong governance, efforts to map, measure, or manage risks may be fragmented, misaligned, or might not occur at all. The Govern function supports institutional accountability, risk awareness, and cross-functional collaboration throughout the AI lifecycle.

Govern is vital in organizations that develop, deploy, or use AI systems. It lays the groundwork for trustworthy AI by embedding risk management into organizational policy and practice.

The details of the Govern function are explained in the following sections.

Category 1: Policies, Processes, and Procedures

Organizations should define, implement, and maintain formal policies, procedures, and controls to guide the development, acquisition, and use of AI systems. Key subcategories include:

- Risk management policies and procedures are documented.
- Procedures are defined that include AI-specific considerations (for example, fairness, explainability, robustness, privacy autonomy).
- Policies that reflect organizational values and legal/regulatory requirements are established.

For example, a healthcare organization that deploys diagnostic AI tools adopts policies requiring every algorithm to undergo pre-deployment fairness audits and post-deployment drift monitoring.

Category 2: Roles and Responsibilities

Organizations should establish clear roles and accountabilities across the AI lifecycle. Key subcategories include:

- Roles and responsibilities for risk management are documented and communicated.
- Oversight bodies (for example, AI ethics committees, internal audit) are designated and empowered to act.
- Escalation mechanisms are established for unresolved risks.

The RACI Model in AI Governance

Many organizations use a RACI model (Responsible, Accountable, Consulted, Informed) to assign responsibilities. For example, data scientists may be responsible for model development, while legal/compliance teams are accountable for ensuring regulatory alignment. RACI models are discussed in Chapter 2.

Category 3: Risk Tolerance and Risk Appetite

Organizations must define their risk tolerance thresholds for AI systems and align them with enterprise-wide objectives. Ideally, organizations can extend their existing risk appetite and risk tolerance to incorporate AI-related considerations. Key subcategories include:

- Risk appetite and risk tolerance statements are developed and documented.
- Decisions about acceptable risk levels are informed by business impact and societal context, and they harmonize with risks in other contexts.
- Stakeholder input is gathered from technical, business, and compliance teams when setting thresholds.
- Periodic reviews of risk tolerance are performed as technologies, regulations, and use cases evolve.

For example, a financial institution defines a low risk tolerance for bias in loan approvals but a higher tolerance for experimentation in internal productivity tools.

Category 4: Organizational Culture and Internal Communication

Culture plays a significant role in the success of AI governance and risk management. Governance should promote a culture of ethical awareness, open communication, and risk mindfulness. Key subcategories include:

- Leaders model and promote risk-informed decision-making.
- Employees are encouraged to raise AI risk concerns without fear of retaliation.
- Training and communication programs reinforce ethical AI use.

AI governance must be business-aligned and, to the greatest extent possible, integrated into existing frameworks and structures, thereby increasing the likelihood of success and yielding better outcomes with AI systems.

AI Risk Culture vs. Security Culture

While both emphasize vigilance and proactive risk awareness, AI risk culture also requires continuous questioning of assumptions, attention to context, and inclusion of ethical and societal considerations—not just technical safeguards.

The Map Function: Understanding the Context of an AI System

The AI RMF Map function helps organizations identify and understand the characteristics of an AI system and its operating context. The Map function includes mapping technical components, as well as their intended and foreseeable business purposes, limitations, societal impacts, and stakeholder expectations. Mapping is an essential governance activity because AI risks are highly contextual; the same AI system may be high-risk in one use case and low-risk in another. Mapping should therefore be performed iteratively across the AI lifecycle to reflect evolving technologies, regulations, and use contexts.

The Map function helps organizations avoid treating AI as a "black box." Instead, Map promotes clarity, visibility, and understanding, which informs the subsequent steps of measuring and managing risk.

Category 1: Context and Intended Purpose

This involves understanding why and how each AI system instance is foundational. Organizations should clearly define the system's intended purpose and the environment in which it operates. Key subcategories include:

- Intended uses, goals, and limitations are documented.
- Foreseeable misuses and unintended applications are analyzed and documented.
- The socio-technical context (for example, legal, cultural, and demographic factors) is analyzed.

- AI system alignment with mission and values is assessed.
- Documentation is made traceable and auditable for governance and accountability purposes.

For example, a public transit agency mapping an AI-powered passenger route planner considers local geography, commuter demographics, and service equity goals before deployment, to ensure safe, fair, and ethical operation.

Category 2: System Design and Functionality

As mentioned throughout this book, Category 2 asserts that organizations should document how each AI system works, including data inputs, model architecture, and decision logic. Key subcategories include:

- Data sources, preprocessing steps, and model features are identified.
- Human-AI interaction design is described.
- Model performance metrics and boundaries are defined.

Avoiding Overreliance on Performance Metrics

Performance and trustworthiness are separate and loosely related (if at all). A model with high accuracy may still reinforce harmful stereotypes or erode human autonomy. Mapping must include qualitative context, not just quantitative performance.

Category 3: Stakeholders and Affected Populations

Stakeholder analysis ensures the understanding of direct and indirect impacts, particularly for vulnerable or underrepresented groups. AI systems must treat affected persons fairly. Key subcategories include:

- Primary users and decision-makers are identified.
- Impacted individuals, communities, and third parties are considered.
- Stakeholder engagement processes are documented.

For example, when designing an academic admissions algorithm, a school district collects feedback from several stakeholders—including students, parents, civil rights groups, and school staff—to assess potential impacts.

Category 4: Risk Identification and Prioritization

Mapping results in the identification of potential risks, including those related to safety, discrimination, privacy, manipulation, and environmental impact. Subcategories include:

- Known and foreseeable risks are documented.
- Risks are prioritized based on likelihood and severity.
- Sources of uncertainty and gaps in understanding are identified.

Epistemic vs. Aleatoric Uncertainty

- Epistemic uncertainty refers to unknowns from limited knowledge (for example, insufficient training data).
- *Aleatoric uncertainty* refers to inherent randomness (for example, noise in sensor data).

Mapping should identify both types of risks.

The Measure Function: Assessing AI Risk and Trustworthiness

As the term implies, the AI RMF Measure function focuses on evaluating how well an AI system meets expectations related to performance, reliability, trustworthiness, and risk mitigation. Measure emphasizes *ongoing* evaluation using quantitative and qualitative methods. Measuring is not a one-time task but a continuous activity aligned with the dynamic and evolving nature of AI systems.

The Measure function helps organizations determine whether each AI system instance operates as intended, and whether it remains within acceptable risk thresholds throughout its lifecycle.

Category 1: Risk Metrics and Indicators

Organizations should define metrics and indicators to track AI-related activities and risks over time, enabling them to monitor progress and identify potential issues. These metrics may include technical, operational, ethical, and societal dimensions. Key subcategories:

- Risk measurement criteria are established for each identified risk.
- Metrics are tailored to the system's context and stakeholders.
- Metrics are updated as the understanding of system behavior and risk evolves.

For example, an employment applicant tracking and screening algorithm is monitored using fairness metrics (for example, disparate impact ratio), accuracy by demographic group, and stakeholder satisfaction surveys.

Category 2: Evaluation Methods

The evaluation of AI systems should encompass a range of methods, including testing, simulation, auditing, and human review, to gain a comprehensive understanding of how the system behaves. Subcategories include:

- Testing methods (for example, unit testing, red teaming, blue teaming, and stress testing) are applied.
- Qualitative methods (for example, expert panels, user studies) complement quantitative data.
- Third-party audits are considered for high-risk use cases.

The Importance of Multimodal Evaluation

The evaluation of fairness, explainability, and harm requires more than just numbers. Qualitative insight, such as stakeholder interviews, focus groups, and case studies, adds depth to understanding how systems may cause harm or produce unexpected outcomes.

Category 3: System Performance Monitoring

Organizations should monitor system behavior in live environments to detect performance degradation, drift, anomalies, and unintended consequences. Subcategories include:

- Continuous performance monitoring and logging.
- Event monitoring systems detect errors and anomalies.
- Feedback loops refine and adapt the system.

For example, a content moderation system on a social media platform uses real-time monitoring to detect increases in errors and false positives after a system update.

Category 4: Risk Documentation and Reporting

Transparent and actionable risk documentation supports accountability and helps avoid unacceptable risks. Subcategories include:

- Risk measurement results are documented and archived.
- Reports are available to stakeholders with appropriate detail and context.
- Results inform future system updates and policy decisions.

Model Cards and System Cards

Tools like model cards and *system cards* (used in large language models) can help document AI system performance, limitations, and risks in a structured, accessible format.

The Manage Function: Taking Action to Address AI Risk

The AI RMF Manage function focuses on proactively responding to identified AI risks and events through design adjustments, operational controls, policy interventions, and stakeholder engagement. While the Govern, Map, and Measure functions build understanding and structure, Manage operationalizes that insight into concrete risk mitigation and system improvement.

Management activities should be iterative, context-sensitive, and responsive to change. Risks may evolve as data shifts, user behavior changes, or social expectations develop, and the Manage function helps organizations adapt accordingly.

The Manage function can be thought of as AI system operations.

Category 1: Risk Treatment and Mitigation

Organizations must implement and maintain *controls* (policies, processes, or procedures) to reduce risk to acceptable levels based on AI model context and risk tolerance. Key subcategories include:

- Mitigation actions are implemented based on *risk assessment* results.
- Controls include technical (for example, algorithmic constraints) and organizational (for example, policy enforcement) measures.
- Risk treatments are tailored to the system's use case and potential harms.

According to standard risk management practices, all *risk treatment* decisions (whether *risk acceptance*, *risk transfer*, *risk avoidance*, or *risk mitigation*) should be approved by management and documented.

For example, a voice recognition system used for telephone-based banking deploys additional speaker verification for users with various accents to reduce bias without degrading security.

Category 2: Incident Response and Recovery

Even with the best controls, mitigation efforts cannot eliminate all risk, meaning that incidents and breaches may still occur. Thus, organizations must plan for AI-specific incidents and establish clear protocols for detecting, escalating, and resolving them. Key subcategories include:

- AI-specific incident response playbooks are documented and rehearsed.
- AI incident response should align with the organization's overall cybersecurity and incident response framework.
- Escalation procedures are tied to defined risk thresholds.
- Recovery and after-action reviews include learning loops to prevent incident recurrence.

What Is an "AI Incident"?

An AI incident might include:

- A self-driving car veers off path due to a model failure or unforeseen circumstance.
- An AI assistant giving dangerous advice.

- A bias audit reveals disparate treatment in high-stakes decisions.

Organizations should treat AI incidents as learning opportunities, not just compliance or operational failures.



The practice of incident management is fully described in other books published by this author, specifically for the CISM (Certified Information Security Manager), CISA (Certified Information Systems Auditor), and CRISC (Certified in Risk and Information Systems Control) professional certifications.

Category 3: Risk Management Integration

Risk management should be an integral part of routine organizational processes, rather than being treated as an isolated, compliance-only function. Key subcategories include:

- AI risk is integrated with broader cybersecurity, privacy, and *enterprise risk management (ERM)* practices.
- Cross-functional teams coordinate risk decisions across departments.
- Procurement, vendor management, and third-party tools follow AI risk protocols.

For example, an enterprise aligns its AI model lifecycle governance with existing ISO 27001–based risk management frameworks, enabling unified risk visibility across cybersecurity, privacy, and AI ethics domains.

Category 4: Risk Evolution and System Retirement

The risk landscape for AI systems will continue to evolve due to changes in corporate objectives, business models, data, regulations, and societal values. Organizations should evaluate whether systems continue to provide value, or if they should be updated, repurposed, or decommissioned. Key subcategories include:

- Systems are periodically and routinely reviewed for ongoing fitness and relevance.
- End-of-life planning is in place for AI systems with known risks or obsolescence.
- Decommissioning includes proper documentation and post-use impact assessment.

Don't Let Models Outlive Their Governance

AI systems can persist long after their original purpose has been fulfilled and their original teams have disbanded. Without sunset planning, outdated or poorly monitored models may continue to operate in shadow, causing unmonitored harm and material consequences.

Putting It All Together: Integrating the AI RMF into Governance Practice

The four core functions of the NIST AI Risk Management Framework (Govern, Map, Measure, and Manage) are intended to be iterative, interdependent, and adaptive. But fulfilling these objectives is up to the organization. Unlike linear or checklist-based compliance tools, the AI RMF emphasizes continuous alignment between organizational values, stakeholder expectations, and system behavior.

When applied effectively, the AI RMF helps organizations:

- Build and enable organizational readiness for responsible AI through governance structures and culture.
- Achieve situational awareness by mapping systems and their context, and through continuous monitoring.
- Conduct evidence-based assessments of risk, performance, and trustworthiness.
- Implement targeted and intentional responses to events, incidents, threats, failures, and ethical concerns.

Many organizations utilize the AI RMF to develop AI-specific overlays on top of existing governance and risk management processes, aligning with governance models such as ISO/IEC 27001, enterprise risk management (ERM), or the NIST CSF.

The AI RMF is designed to scale with system complexity, organizational size, and organizational maturity. A small startup building a chatbot can apply it in a lightweight form; a multinational deploying AI in financial services or healthcare may implement all categories and subcategories rigorously, with multiple layers of oversight.

AI RMF vs. NIST Cybersecurity Framework (CSF)

Although developed separately, the NIST AI RMF and *NIST Cybersecurity Framework (CSF)* share a common ancestry. Both are voluntary, outcome-focused frameworks designed to promote trust, resilience, and responsible risk management. However, their domains, assumptions, and mechanisms differ significantly. Their differences are depicted in Table 7.1. Table 7.2 illustrates various use cases where either CSF or AI RMF is an appropriate framework.

Fundamental Philosophical Shift

The CSF treats risk as something to be defended against using well-known controls.

The AI RMF treats risk as emergent, systemic, and value-laden, requiring social, organizational, and technical strategies.

TABLE 7.1 Differences Between NIST AI RMF and NIST CSF

Element	AI RMF	CSF
Focus domain	AI system lifecycle and socio-technical risks	Cybersecurity controls and resilience
Types of risk addressed	Bias, opacity, misuse, autonomy, drift, societal harm	Confidentiality, integrity, availability (CIA), operational resilience
Treatment of uncertainty	Explicit handling of epistemic/aleatoric uncertainty	Generally assumes known or measurable threats
Emphasis on explainability and fairness	Core theme	Not addressed
Reliance on system context	Central to risk mapping	Less context-sensitive, control-based
Ethical dimensions	Integral to the framework	Not in scope

TABLE 7.2 When to Use the AI RMF vs. the CSF

Use Case	Appropriate Framework
Preventing malware in IT systems	CSF
Evaluating bias in an HR screening algorithm	AI RMF
Securing cloud infrastructure for AI workloads	CSF (with AI-specific overlays)
Governing the ethics and impacts of autonomous vehicles	AI RMF (with crosswalks to CSF for security)

In organizations with mature cybersecurity programs, the AI RMF can be layered onto the CSF to provide a holistic view of AI system risk, blending technical security with socio-technical risk governance.

NIST ARIA: Layered AI Model Evaluation

AI systems are becoming increasingly capable and are being deployed to support more consequential and complex use cases, creating a need for reliable, repeatable, and transparent assurance mechanisms. In May 2024, NIST launched the *Assurance of AI Systems (ARIA)* program, a forward-looking initiative designed to develop methods, tools, and measurement techniques for evaluating the safety, trustworthiness, and effectiveness of AI systems.

ARIA is still evolving and not yet a formal standard. At present, ARIA focuses on risks and impacts of large language model (LLM) AI systems. Future work may include other types of generative AI. You can find information about ARIA at `https://ai-challenges.nist.gov/aria`.

Purpose and Positioning of ARIA

While the AI RMF is intended to be voluntary and flexible, ARIA takes a more targeted approach. Its primary audience is the U.S. federal government, particularly agencies that build or use AI systems in support of their missions. However, private-sector organizations, research institutions, vendors, and contractors may also benefit from ARIA as a practical blueprint for managing AI risk in regulated or complex environments.

Rather than replacing the AI RMF, ARIA complements it through a layered approach to AI system testing. In doing so, it acts as a guide for operationalizing the assessment portion of AI risk management.

The Three Levels of Testing

ARIA evaluations of AI systems include three distinct testing levels, each with a specific focus.

Model Testing

The purpose of model testing in ARIA is to confirm an AI system's claimed capabilities. According to the ARIA pilot evaluation plan published in 2024, model testing questions include:

- Does the system demonstrate required capabilities?
- Does the system demonstrate required guardrails?

Red Teaming

The purpose of red teaming in ARIA is to determine whether testing can cause the AI system to malfunction or behave in unexpected or unplanned ways. Similar to red teaming in other contexts (applications, information systems, networks), red team testing attempts to trick an AI system into performing actions it is not designed to perform. Per the ARIA pilot evaluation plan, testing questions include:

- Can the system be induced to produce violative outcomes?
- Under what conditions do violative outcomes occur?

Field Testing

The purpose of field testing in ARIA is to explore the potential positive and negative impacts of the AI system in ordinary use by end users. The ARIA pilot plan includes questions such as:

- Are users exposed to positively or negatively impactful information during regular use?
- Do users perceive the positively or negatively impactful information presented to them?
- Based on impactful exposure, what are users' subsequent actions?

Practical Application of ARIA

ARIA is not a checklist but a guide that enables organizations to prioritize actions based on their mission, resources, and risk profile. While ARIA itself exists as a blueprint for testing AI systems, the results of such testing may inform broader aspects of AI governance. For instance, if AI model testing produces numerous exceptions, business leaders may infer the need for more thorough or effective governance functions, such as improved requirements development.

ARIA and Executive Branch Compliance

In the U.S. federal government, ARIA supports compliance with a growing number of federal requirements surrounding AI development and use. Notably, it helps agencies meet the expectations of:

- **OMB Memorandum M-24-10,** which mandates that agencies conduct risk assessments, maintain AI inventories, and submit governance plans.
- **Executive Order 14110,** signed on 30 October 2023, which directed federal agencies to promote the safe, secure, and trustworthy development of AI technologies. This order was rescinded in January 2025, representing a significant policy shift.
- **Other executive orders,** including 14319 and 14320.

By aligning internal AI evaluation practices (such as those promoted in ARIA) with these federal mandates, agencies can streamline compliance reporting and support robust risk management.

ARIA Beyond the Federal Government

While initially designed with the U.S. public sector in mind, ARIA's structure and approach have broader relevance. Private-sector companies, NGOs, and academic institutions can adapt ARIA principles to:

- Strengthen their responsible AI programs.
- Build institutional capacity around AI risk.
- Align internal controls with external regulatory expectations.
- Complement governance frameworks such as the NIST AI RMF or ISO/IEC 42001.

ARIA's modular nature makes it broadly adaptable, even outside of federal contexts.

ISO Standards

International standards play a crucial role in enabling global alignment, interoperability, and trust in existing and emerging technologies; AI is no exception. Within the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC), specialized committees of experienced professionals have been actively building and shaping the landscape of AI standards to support consistent terminology, methods, lifecycle management, and governance practices. The two most relevant and important standards for AI governance professionals are ISO/IEC 22989 and ISO/IEC 42001.

This section explains the role and structure of these two ISO AI standards. It explores how 22989 provides foundational definitions and concepts, while 42001 describes a management system framework (similar to ISO 27001 for information security) for organizations that develop, deploy, or use AI systems responsibly. A short list of related ISO standards follows this.

ISO standards are licensed to users and organizations and you can obtain them from `https://www.iso.org/`.

ISO/IEC 22989: Concepts and Terminology for AI

Standardization begins with a standard lexicon. In a typical professional conversation, it's crucial that all parties share a common understanding of terms like bias, drift, or transparency. *ISO/IEC 22989* defines the core concepts and terminology related to AI, offering a foundational vocabulary for policy, development, and governance. This standard does not prescribe how AI should be built or managed. Instead, it ensures that stakeholders, including developers, auditors, regulators, and the public, use consistent and interoperable terms when discussing AI systems.

Why Terminology Matters in Governance

The absence of shared definitions creates miscommunication and misalignment, particularly when AI systems cross organizational, jurisdictional, or cultural boundaries. ISO/IEC 22989 supports:

- Interdisciplinary collaboration
- Contractual clarity for AI procurement and legal agreements

- Cross-border regulatory alignment
- Audit readiness and documentation standards

Not All "AI" Is Actually AI

ISO 22989 emphasizes precision: an "AI system" must involve elements of perception, reasoning, or learning. This distinguishes AI from simple automation or business rules, which is an essential distinction for risk classification and regulatory scope.

Key Definitions and Concepts in ISO 22989

Table 7.3 includes selected core terms defined in ISO 22989 with particular relevance for AI governance.

Organizations that use ISO 22989 terms in RFPs and contracts are more likely to avoid ambiguity.

TABLE 7.3 Selected AI Governance Definitions in ISO/IEC 22989

Term	Definition (Summarized)*
Artificial Intelligence (AI)	The capability of a system to perform tasks that typically require human intelligence.
AI System	An engineered system that generates outputs such as predictions, recommendations, or decisions, for a given set of objectives.
Machine Learning (ML)	A set of methods allowing systems to improve performance on a task based on data.
Autonomy	The ability of a system to operate without ongoing human intervention.
Transparency	The degree to which the internal workings of a system and its decision-making processes can be understood.
Explainability	The degree to which a human can understand and interpret the output or behavior of an AI system.
Trustworthiness	The ability of an AI system to operate as intended while minimizing risk and promoting confidence among stakeholders.

*Definitions are summarized here; the full definitions are protected by copyright.

AI Capabilities and System Behaviors

ISO/IEC 22989 provides conceptual groupings of AI capabilities, helping frame discussions about what a system's capabilities and interactions with the world are. These include:

- Perception (for example, image recognition)
- Reasoning (for example, optimization, rule-based logic)
- Learning (for example, neural networks, supervised learning)
- Interaction (for example, natural language, gesture interfaces)
- Decision-making (for example, autonomous planning, risk scoring)

Lifecycle Relevance

Understanding an AI system's capabilities helps determine which risks are relevant at each lifecycle stage, from data sourcing to user interface deployment.

AI System Lifecycle Stages

ISO/IEC 22989 defines a lifecycle model for AI systems in high-level conceptual terms. Remember, though, that ISO/IEC 22989 is a vocabulary standard that provides no information on implementing the lifecycle. More detailed lifecycle frameworks, such as ISO/IEC 23053 (machine learning lifecycle) and ISO/IEC 5338 (AI lifecycle management), describe stages like the following:

- Business requirements and design
- Data acquisition and processing
- Model development and training
- System integration and testing
- Deployment and operation
- Monitoring, maintenance, and retirement

An example retail company aligns its AI development process with the ISO/IEC 22989 lifecycle stages, enabling auditors to know where human oversight is required during design, training, deployment, and operations.

ISO/IEC 42001: AI Management Systems

ISO/IEC 42001, published in 2023, is the world's first certifiable management system specifically designed for organizations that develop, provide, or use AI systems. Modeled after established standards such as *ISO/IEC 27001* (information security) and *ISO 9001*

(quality), ISO/IEC 42001 helps organizations operationalize AI governance by embedding trust, transparency, and risk controls into their daily practices.

Where ISO/IEC 22989 provides common terminology, ISO/IEC 42001 defines a structured governance system, complete with roles, policies, documentation, internal audits, and continual improvement cycles, enabling organizations to demonstrate responsible AI use through formal certification. Like other ISO standards, ISO/IEC 42001 is scalable and flexible, equally fit for startups and multinational manufacturing organizations. ISO/IEC 42001 and 22989 are hand-in-glove.

The Purpose and Scope of ISO/IEC 42001

ISO/IEC 42001 can be used by any organization involved in the AI value chain, including AI developers (model designers, data scientists), AI deployers (platform providers, integrators), AI users (enterprises using third-party AI tools), and even organizations indirectly affected by AI (those using embedded AI features).

ISO/IEC 42001 does not prescribe how to build AI models. It defines how organizations manage AI systems, through leadership, policies, controls, roles, and governance processes—for the entire AI system lifecycle.

Core Requirements of ISO/IEC 42001

ISO/IEC 42001 follows the same high-level structure as other ISO management standards, making it compatible with ISO 27001, 31000, and others. Key components include:

- **Context of the organization:** Understanding internal and external factors affecting AI use, including applicable laws and standards
- **Leadership and commitment:** Top management setting the tone, allocating resources, and owning responsibility
- **Planning:** Risk assessment, objectives, and actions for trustworthiness goals
- **Support:** Resources, competence, awareness, and documented information
- **Operation:** Design, acquisition, development, and deployment processes
- **Performance evaluation**: Monitoring, audit, and management review
- **Improvement:** Corrective actions and continuous improvement mechanisms

Risk Management in ISO/IEC 42001

Risk-based thinking is central to 42001. According to the standard, organizations must identify and assess risks related to AI systems, define criteria for acceptable levels of risk (for example, risk appetite and risk tolerance), and implement controls, mitigations, and monitoring procedures to ensure that risks remain within accepted bounds.

The types of risk sources that ISO/IEC 42001 cites include:

- Algorithmic bias or unfair outcomes
- Unintended behavior from machine learning models
- Overreliance on AI outputs without human oversight
- Regulatory non-compliance or reputational harm

Alignment with AI RMF

ISO/IEC 42001's risk requirements align well with the NIST AI RMF functions (Govern, Map, Measure, Manage), providing a bridge between strategy and formal certification. Both standards are excellent sources for AI governance structure and related processes.

Trustworthiness Dimensions in 42001

Per the ISO/IEC 42001 standard, organizations are required to consider several trustworthiness factors in their AI management practices, including:

- Transparency and explainability
- Fairness and non-discrimination
- Robustness and resilience
- Accountability and human oversight
- Privacy and data governance

You can think of these as a measure of risk tolerance. These considerations influence risk treatment plans, system design principles, and stakeholder engagement strategies.

For example, a financial services firm includes fairness and transparency as documented trust objectives in its 42001 AI management system, triggering mitigation actions whenever key measures fall below defined thresholds.

Implementing ISO/IEC 42001 in Practice

Implementing ISO/IEC 42001 is not just about gaining certification or checking a box; it is about achieving a higher level of quality and performance. Moreover, it's about embedding governance discipline into an organization's processes for conceiving, building, and managing AI systems. This section outlines an approach to certification, including gap analysis, alignment with other ISO standards, readiness for internal audits, and scaling governance across diverse teams and systems.

Conducting a Readiness Assessment

Before a full implementation of ISO/IEC 42001, organizations should perform a *gap analysis* to compare current practices with the requirements of ISO/IEC 42001. This assessment includes:

- Inventory of AI systems and use cases
- Inventory of current policies and procedures governing AI development and use
- Inventory of existing roles and accountabilities with AI oversight
- Inventory of internal documentation, training, and testing protocols

For example, a retail chain conducting an ISO 42001 gap assessment found that it has robust model development processes but no formal oversight of vendor-provided AI solutions. This gap highlights the need for extending its third-party risk management process to third-party AI vendors.

Integrating with Existing Management Systems

Because ISO/IEC 42001 follows the same high-level structure as other ISO governance standards, it can be integrated with existing ISO-inspired processes, including:

- ISO/IEC 27001 (Information Security Management)
- ISO 9001 (Quality Management)
- ISO/IEC 27701 (Privacy Information Management)

Shared Controls Across Systems

A single risk register, incident response plan, or audit function can often support multiple standards. ISO/IEC 42001 encourages the reuse of resources where appropriate. Why have multiple, separate, stove-piped processes when combining them into a single process would be more effective and efficient?

Developing AI-specific Policies and Procedures

As organizations develop AI governance processes, key documents supporting implementation include AI system lifecycle management policies, ethical AI principles aligned with organizational values, procedures for bias detection and mitigation, logging and monitoring requirements for AI system outputs, and stakeholder engagement protocols.

These documents should reflect the organizational context, including sector-specific risks, regulatory obligations, and stakeholder expectations.

An example of a healthcare provider implementing 42001 is creating a cross-functional AI oversight committee and developing new policies that require algorithmic validation before use in clinical settings.

Training, Awareness, and Organizational Culture

ISO/IEC 42001 places strong emphasis on competence and awareness. Successful implementation requires that technical staff understand governance obligations, business users and managers know how to recognize and escalate AI-related risks, and leadership models a culture of ethical and responsible AI use.

Beyond Compliance

Certification is not the ultimate goal but rather an indicator of maturity. Organizations that embed ISO/IEC 42001 as a living governance discipline will build AI systems that are more adaptable, resilient, and trusted over time.

Other Relevant ISO Standards for AI Governance

While ISO/IEC 22989 and ISO/IEC 42001 serve as the foundational standards for AI governance, several additional ISO and IEC standards support specific disciplines of responsible AI design, deployment, and oversight. These additional standards help organizations enhance their governance programs by addressing specific domains of trustworthiness, operational practices, and risk categories.

ISO/IEC 23894: AI Risk Management

This standard complements ISO/IEC 42001 by providing a dedicated framework for risk management in AI systems. It aligns closely with ISO 31000 (Enterprise Risk Management) and offers tailored guidance on identifying and analyzing AI-specific risks, determining appropriate controls, and reviewing risk postures throughout the AI lifecycle.

Note that *ISO/IEC 23894* is not about risk management *using* AI but rather about the risk management *of* AI. The standard helps organizations operationalize risk assessment within AI system design and deployment, especially when integrated with 42001.

ISO/IEC 38507: Governance of IT (the Governance Implications of AI)

Part of the ISO 38500 series on IT governance, this standard provides strategic-level guidance for board members and executives on overseeing AI use. The standard discusses roles and responsibilities for AI oversight, risk appetite, and ethical alignment, plus accountability and transparency at the leadership level.

This standard bridges the gap between executive oversight and operational controls, making it valuable for board briefings and governance committee planning.

ISO/IEC TR 24028: Trustworthiness in AI

This standard is a technical report (not a certifiable standard) that introduces the concept of AI trustworthiness and describes mechanisms to achieve it. Topics include transparency, reliability, resilience, privacy, and fairness.

ISO/IEC TR 24028 provides a conceptual foundation for trustworthiness claims and can, in conjunction with risk management, inform control design in ISO/IEC 42001 implementations.

ISO/IEC 42005: Artificial Intelligence (AI) (the AI System Impact Assessment)

This standard defines methodologies for assessing the impacts of AI systems—and their foreseeable applications—on individuals, groups, and society as a whole. Impact assessments are increasingly required under laws like the EU AI Act. *ISO/IEC 42005* could become a standard method for performing and documenting these evaluations.

ISO AI Standards in the Global Governance Landscape

ISO standards are a vital component of global AI governance. As national and regional regulations emerge—from the EU AI Act to the U.S. Executive Orders and regulations on AI—ISO offers a non-regulatory, globally harmonized baseline that supports:

- Standard vocabulary and methodologies
- Cross-border consistency
- Sector-specific adaptation
- Vendor-neutral implementation
- Certification and audit readiness

Organizations aligning with ISO/IEC 22989 and 42001 position themselves to meet emerging legal, ethical, operational, and societal expectations while signaling leadership in trustworthy AI.

ISO vs. Legal Compliance

ISO compliance is voluntary but often serves as a proof point for due diligence, risk maturity, and regulatory alignment. However, ISO compliance and certification are frequently cited in legal agreements, allowing organizations to hold others accountable to these international standards.

ISO standards do not stifle innovation; instead, they create pragmatic guardrails that enable safe innovation, fair deployment, and transparent accountability. As regulatory bodies and certification ecosystems mature, ISO standards are likely to become:

- Benchmarks for AI conformity assessments
- Reference models in enforcement guidance
- Building blocks for sector-specific codes of practice

Other AI-related Standards

While the OECD, NIST, and ISO initiatives form the core of today's AI governance standards, other influential frameworks are emerging. These include region-specific regulations, industry-led guidelines, and multi-stakeholder frameworks that support responsible AI development and deployment. This section explores some of these, emphasizing their influence, unique contributions, and relevance to AI governance professionals.

The EU High-level Expert Group (HLEG) Ethics Guidelines

Prior to the EU AI Act, the European Commission convened an expert group to define ethical principles for trustworthy AI. In 2019, the *High-level Expert Group (HLEG)* on AI released its ethics guidelines for trustworthy AI. The key components include:

- Seven core requirements for trustworthy AI:
  1. Human agency and oversight
  2. Technical robustness and safety
  3. Privacy and data governance
  4. Transparency
  5. Diversity, non-discrimination, and fairness
  6. Societal and environmental well-being
  7. Accountability
- Four ethical principles for trustworthy AI:
  1. Respect for human autonomy
  2. Prevention of harm
  3. Fairness
  4. Explicability
- A self-assessment checklist for organizations to evaluate adherence

The HLEG Guidelines directly influenced the EU AI Act, particularly in risk categorization and obligations for high-risk systems.

The G7 Hiroshima Process and Code of Conduct

The G7 Hiroshima Process was launched in 2023 as a response to the rapid evolution of generative AI. In October 2023, the G7 published a Guiding Principles and a Code of Conduct for Advanced AI Systems.

The Guiding Principles and Code of Conduct for organizations developing advanced AI systems focuses on these points:

1. "Identify, evaluate, and mitigate risks across the AI lifecycle.
2. Rectify patterns of misuse, after deployment, including placement on the market.
3. Publicly report advanced AI systems' capabilities, limitations and domains of appropriate and inappropriate use.
4. Work towards responsible information sharing and reporting of incidents among organizations developing advanced AI systems, including with industry, governments, civil society, and academia.
5. Develop, implement, and disclose AI governance and risk management policies.
6. Invest in and implement robust security controls.
7. Develop and deploy reliable content authentication and provenance mechanisms.
8. Prioritize research to mitigate societal, safety, and security risks and prioritize investment in effective mitigation measures.
9. Prioritize the development of advanced AI systems to address the world's greatest challenges.
10. Advance the development of and, where appropriate, adoption of international technical standards.
11. Implement appropriate data input measures and protections for personal data and intellectual property."

These and other artifacts are available from `https://g7g20-documents.org/`; search for AI Code of Conduct and AI Guiding Principles.

The IEEE Ethically Aligned Design Framework

The Institute of Electrical and Electronics Engineers (IEEE) had a significant role in shaping the ethical foundations of AI system design through its Ethically Aligned Design (EAD) initiative and the IEEE 7000 series of standards. The relevant IEEE standards are:

- **IEEE 7000:** Model process for addressing ethical concerns in system design
- **IEEE 7001:** Transparency in autonomous systems

- **IEEE 7002:** Data privacy in AI systems
- **IEEE 7003:** Algorithmic bias considerations
- **IEEE 7010:** Well-being metrics for autonomous systems

The IEEE 7000 series considers ethics as an integral design consideration, incorporated into requirements analysis, system modeling, and validation.

You can find these resources at `https://standards.ieee.org/`.

World Economic Forum (WEF) AI Governance Toolkits

The World Economic Forum (WEF) has published toolkits and playbooks that focus on AI governance for policymakers and organizational leaders. Key contributions include:

- **AI Governance Alliance Toolkit (2023):** Cross-sector governance guidance for responsible AI deployment
- **Presidio Recommendations on Responsible AI (2020):** A multi-stakeholder model for risk-based governance
- **ChatGPT and Generative AI Toolkit for Policymakers (2023):** Risk mapping and regulatory design guidance

You can obtain these resources from `https://initiatives.weforum.org/ai-governance-alliance/home`.

OECD Framework for Classification of AI Systems

Separate from its ethical principles discussed earlier in this chapter, the OECD released a framework for the classification of AI systems in 2022. This framework categorizes AI systems based on context, autonomy, human interaction, impact scale, and affected parties. While similar to the risk classifications in the EU AI Act, the OECD classification framework is not a binding standard.

United Nations UNESCO Recommendation on the Ethics of AI

In 2021, UNESCO adopted the first global standard on the ethics of AI, applicable to all 194 member states. Currently, this standard is not a binding regulation but rather a proposed practice. The UNESCO recommendation on the ethics of AI promotes human rights–centered design and includes:

- Ethical impact assessments
- Ethical governance and stewardship
- Data policy

- Environmental sustainability
- Gender equality
- Protection of cultural diversity
- Algorithmic transparency
- Health and social well-being

You can find the full content of the UNESCO document at `https://unesdoc.unesco.org/ark:/48223/pf0000381137`.

Singapore Model AI Governance Framework

Singapore has emerged as a leader in model-based, sector-neutral AI governance. Its Model AI Governance Framework, released by the Singapore Infocomm Media Development Authority (IMDA) in May 2024, emphasizes:

- Accountability
- Data quality
- Trusted development and deployment
- Incident reporting
- Testing and assurance
- Security
- Content provenance
- Safety and alignment of R&D
- AI for public good

You can find this framework at `https://aiverifyfoundation.sg/resources/mgf-gen-ai/`.

ISO/IEC JTC 1/SC 42 Roadmap and Future Standards

The ISO subcommittee responsible for AI, known as ISO/IEC JTC 1/SC 42, has released a roadmap for future AI standards. Upcoming areas of standardization include:

- AI system impact assessments
- AI quality and performance evaluation
- Bias mitigation methodologies
- AI lifecycle assurance

These are all areas in which practices exist today but will warrant improvement through industry experience and new AI innovations.

AI governance professionals should monitor the SC 42 roadmap, as upcoming standards are likely to influence future certification schemes, procurement contracts, and cross-border compliance.

Summary

AI systems are no longer experimental tools operating in isolation; they are embedded across industries and societal functions, carrying both great promise and significant risk. In this context, a growing body of AI standards and frameworks has emerged to guide organizations in building, deploying, and managing AI responsibly. This chapter introduced key international and national efforts to provide structured, interoperable, and trustworthy approaches to AI governance.

The OECD AI Principles laid the ethical foundation for much of the world's AI policy activity. By articulating five core tenets (including human-centered values, transparency, and accountability), the OECD created a flexible, non-binding framework that has influenced formal regulatory systems, such as the EU AI Act. These principles serve as a moral and strategic compass, providing a shared vision of what responsible AI should aim to achieve, even as technologies and contexts evolve.

In the United States, the National Institute of Standards and Technology (NIST) has developed the AI Risk Management Framework (AI RMF), a voluntary but comprehensive tool to help organizations manage AI-specific risks. Structured around four core functions (Govern, Map, Measure, and Manage), the AI RMF encourages organizations to take a mature, lifecycle view of AI systems (similar in structure to SDLC), acknowledging both technical and socio-technical risk vectors. The framework is flexible enough to be applied across industries, yet rigorous enough to support assurance and audit functions.

Building on the AI RMF, NIST's ARIA (Assessing Risks and Impacts of AI) program tackles a different question: how can we verify that AI systems actually meet their intended safety, fairness, and trustworthiness goals? ARIA provides action-oriented methodologies, tools, and metrics that support testing, monitoring, and validation of AI behavior under real-world conditions. ARIA emphasizes not only performance metrics but also explainability, robustness, and the dynamics of human-AI interaction, enabling a deeper operationalization of trustworthy AI principles.

Internationally, ISO has developed several AI standards that contribute structure, terminology, and certification pathways. ISO/IEC 22989 offers foundational definitions of AI concepts, systems, and lifecycle stages, promoting consistency across technical and regulatory domains. ISO/IEC 42001, by contrast, introduces a certifiable management system for AI, helping organizations embed governance practices, risk controls, and continuous improvement mechanisms into their everyday operations. Additional ISO standards further support risk management, impact assessment, and trustworthiness.

Several other global frameworks play a vital supporting role. The EU High-level Expert Group's Ethics Guidelines, the G7 Hiroshima Process, IEEE 7000 series, UNESCO's ethics recommendations, Singapore's Model AI Governance Framework, and tools from the World Economic Forum each contribute to a broader mosaic of safe, ethical, and responsible AI. While diverse in scope, they share common themes of accountability, fairness, ethics, transparency, and human-centered design. Together, these standards and frameworks enable AI governance professionals to build systems that are not only compliant but also ethically aligned, socially responsive, and resilient in the face of future challenges.

Exam Essentials

The NIST AI RMF is built around four core functions: Govern, Map, Measure, and Manage. These four functions form the backbone of the AI RMF and guide organizations through the full lifecycle of AI risk management, which is an essential component of AI governance. The RMF Govern function establishes accountability and oversight structures. Map involves identifying the system, its context, and relevant risks. Measure focuses on evaluating system performance and risk posture. Manage includes mitigation strategies and incident response. Understanding each function and its supporting categories is essential for applying the framework effectively.

ISO/IEC 22989 provides a standard AI governance lexicon. This standard provides standard definitions of key terms used in AI governance and management. The standard defines terms but lacks operational guidance, which is found instead in ISO/IEC 42001.

ISO/IEC 42001 is the first certifiable AI management system standard. This standard provides a formal structure for operationalizing AI governance within organizations. Modeled after ISO 27001, it includes clauses for leadership, planning, risk management, performance evaluation, and continual improvement. Certification under ISO/IEC 42001 allows organizations to demonstrate due diligence and alignment with emerging regulatory expectations.

The OECD AI Principles establish an ethical foundation for international AI governance. Adopted by over 40 countries, these voluntary principles promote inclusive growth, human-centered values, transparency, robustness, and accountability. They serve as a global reference point and have directly influenced regulations such as the EU AI Act and policies from G7 member states.

NIST ARIA provides technical methods to assure AI system safety and trustworthiness. ARIA complements the AI RMF by offering specific tools and metrics for testing AI systems, including stress testing, conformance evaluation, explainability assessments, and simulation. It supports qualitative and quantitative methods for verifying AI behavior under real-world conditions.

IEEE, UNESCO, the G7, and the EU HLEG have each contributed significant AI frameworks. These global initiatives address various ethical, societal, and technical aspects of AI governance. For example, IEEE 7000 emphasizes ethically aligned design, UNESCO promotes human rights–centered AI, the G7 Hiroshima Process outlines AI governance, and the EU HLEG provided early guidance on trustworthy AI. While non-binding, these frameworks shape regulatory and organizational priorities worldwide.

Review Questions

The following questions are designed to test your understanding of this chapter's material. For more information on how to obtain additional questions, see this book's Introduction. You can find the answers in Appendix A.

1. What is the primary purpose of the OECD AI Principles?
   - **A.** To create mandatory AI regulations for OECD countries
   - **B.** To define technical testing protocols for AI safety
   - **C.** To establish ethical guidelines for trustworthy AI
   - **D.** To categorize AI systems by risk level
2. Which NIST AI RMF function is responsible for identifying the context and intended purpose of an AI system?
   - **A.** Manage
   - **B.** Measure
   - **C.** Map
   - **D.** Govern
3. What distinguishes NIST ARIA from the NIST AI RMF?
   - **A.** ARIA addresses regulatory compliance, while RMF does not
   - **B.** ARIA focuses on assurance methods and testing techniques
   - **C.** ARIA is a certification framework
   - **D.** ARIA replaces the RMF in high-risk sectors
4. What is a defining feature of ISO/IEC 42001?
   - **A.** It is a privacy-focused AI framework
   - **B.** It standardizes AI ethics assessments
   - **C.** It offers a certifiable AI management system
   - **D.** It provides a taxonomy of AI capabilities
5. What is the purpose of ISO/IEC 22989?
   - **A.** To regulate foundation models
   - **B.** To define global AI enforcement mechanisms
   - **C.** To standardize AI terminology and concepts
   - **D.** To guide AI procurement strategies

6. Which G7 initiative provides guidance on AI governance?
   - **A.** Hiroshima Process
   - **B.** Kyoto Protocol
   - **C.** Tokyo AI Alignment Act
   - **D.** OECD Code of Conduct

7. The IEEE 7000 series focuses on which aspect of AI systems?
   - **A.** Model compression and optimization
   - **B.** Ethical alignment during system design
   - **C.** High-frequency trading algorithms
   - **D.** National security applications of AI

8. Which of the following is a key trustworthiness characteristic emphasized by NIST ARIA?
   - **A.** Profitability
   - **B.** Interpretability
   - **C.** Processing speed
   - **D.** Licensing model

9. What distinguishes ISO/IEC 23894 from ISO/IEC 42001?
   - **A.** 23894 is focused on risk management, while 42001 is a management system
   - **B.** 23894 defines training standards for AI teams
   - **C.** 23894 only applies to AI in manufacturing
   - **D.** 23894 provides legal enforcement guidance

10. Why is UNESCO's AI recommendation particularly influential in the global south?
    - **A.** It provides hardware specifications for AI systems
    - **B.** It offers a prescriptive legal framework
    - **C.** It aligns with principles of cultural inclusion and human rights
    - **D.** It enforces AI bans on specific use cases

Governing the AI Lifecycle

PART III





Chapter

8

AI System Design

THIS CHAPTER COVERS PORTIONS OF AIGP DOMAIN III, "UNDERSTANDING HOW TO GOVERN AI DEVELOPMENT," SUBDOMAIN III.A, "GOVERN THE DESIGNING AND BUILDING OF THE AI MODEL," AS WELL AS PORTIONS OF SUBDOMAIN III.B, "GOVERN THE COLLECTION AND USE OF DATA IN TRAINING AND TESTING THE AI MODEL," SUBDOMAIN IV.A, "EVALUATE KEY FACTORS AND RISKS RELEVANT TO THE DECISION TO DEPLOY THE AI MODEL," AND SUBDOMAIN IV.B, "PERFORM KEY ACTIVITIES TO ASSESS THE AI MODEL." THE FOLLOWING TOPICS ARE COVERED:

- ✔ Lifecycle documentation
- ✔ Developing business context and use cases
- ✔ Performing impact assessments
- ✔ AI system design
- ✔ AI development risk management
- ✔ Considerations when developing proprietary models

THE AI GOVERNANCE PROFESSIONAL (AIGP) PERFORMANCE INDICATORS IN THIS SUBDOMAIN INCLUDE:

- ✔ Document the designing and building process (for example, to establish compliance and manage risks).
- ✔ Document the training and testing process (for example, to validate results, establish compliance and manage risks).



- ✔ **Define the business context and use case of the AI model.**
- ✔ **Understand the context of the AI use case (for example, business objectives, performance requirements, data availability, ethical considerations and workforce readiness).**
- ✔ **Perform or review an impact assessment on the AI model.**
- ✔ **Perform or review an impact assessment on the selected AI model.**
- ✔ **Apply the policies, procedures, best practices and ethical considerations to designing and building the AI model (for example, purpose of AI, requirements gathering, architecture and model selection, human oversight, data analysis, metric and threshold evaluation, stakeholder engagement and feedback and operational controls).**
- ✔ **Identify and manage the internal and external risks and contributing factors related to designing and building the AI model (for example, using probability/ severity harms matrix, using a risk mitigation hierarchy, stakeholder mapping, use case evaluation, benchmarking, pre-deployment pilots and testing).**
- ✔ **Identify and understand issues that are unique to a company deploying its own proprietary AI model (for example, increased obligations and higher potential liability).**



Designing AI systems involves aligning the system's purpose with organizational goals, documenting the full development lifecycle, anticipating risks, and embedding governance at every stage. This chapter explores the critical considerations involved in designing AI systems that are ethical, effective, accountable, and compliant. Topics include business context and use case, design and testing phases, impact assessments, design best practices, design-related risks, and the unique obligations associated with proprietary models. Whether for a simple internal tool or a high-risk decision-making system, design lays the foundation for compliance, transparency, and trust. You will read practical guidance for integrating policy, oversight, and technical rigor into the earliest phases of AI system development.

About "Applicable Laws" in This Chapter

The official AIGP outline for this chapter includes an item titled "Identify laws that apply to the AI model." While this is unquestionably an essential consideration in AI system design, the topic has been addressed in depth in earlier chapters—specifically in Chapters 4 and 5. To avoid redundancy, this chapter does not revisit those legal frameworks in detail. Readers are encouraged to refer to those chapters for comprehensive coverage of applicable laws, regulatory obligations, and jurisdiction-specific requirements. This chapter instead focuses on the design process itself, assuming that legal compliance is being addressed in parallel through proper governance channels.

Refer to the AIGP domain chart in this book's Introduction to better understand how the AIGP domains are covered by the chapters in this book.

Documenting the Lifecycle

Documenting the AI system's lifecycle is essential for ensuring compliance, managing risks, promoting accountability, maintaining transparency, and ensuring operational continuity. Each stage of development introduces potential risks, governance obligations, and

opportunities to embed trustworthy AI practices. This section explores how organizations can systematically document each phase of AI systems development, with an emphasis on traceability, alignment with internal policies and external standards, and readiness for audit or investigation.

Documenting Requirements

Before the design of an AI system can begin, it's essential to formally identify and capture the *requirements* of the proposed system. By documenting requirements, management will be able to determine whether the system operates correctly by confirming whether it successfully fulfills each requirement.

Organizations with a mature system development lifecycle (SDLC) process will have established practices for documenting, reviewing, and baselining a requirements specification.

Essential practices in requirements management include:

- Each requirement must be verifiable, traceable, unambiguous, and measurable.
- Categorize requirements by type, including security, privacy, user experience (UX), regulatory, operational, and established standards (OS, protocols, tools, and so forth). Some organizations maintain a catalog of standard requirements to apply to every system development project.
  - *Functional requirements* describe the AI system's functions.
  - *Technical requirements* (also known as non-functional requirements) describe elements of the tech stack and include performance, scalability, security, privacy, usability, and resilience.
  - *Regulatory requirements* describe AI system functions and characteristics required by specific regulations.
  - *Security requirements* describe AI system functions and characteristics that protect the AI system and associated data.
- Record the source, owner, priority, and rationale for each requirement.
- Set data governance requirements (consent, lawful basis, PII handling, lineage, retention, deletion).
- Define evaluation metrics with thresholds (for example, accuracy, calibration, latency, cost per inference).
- Include the criticality or weighting of each requirement.
- The final set of requirements should be approved by management.

Designers and developers should be provided with a set of approved requirements to serve as a guide during the design and development phases of the project.

Don't Let the Loudest Voice Win

Stakeholders with organizational influence may dominate the requirements-setting process. Counterbalance this by explicitly incorporating feedback from marginalized or impacted groups. Remember to document who provided each requirement, along with its criticality.

Documenting the Designing and Building Process

Documentation for an AI system has two dimensions: first, a detailed description of the design and build process, and second, a detailed description of the design of the AI system. The design and build process needs to be documented only once (and periodically revised thereafter), and each AI system's design must also be documented. This section focuses primarily on the documentation for an AI system.

Documentation in the Design Phase

Documenting the design phase of an AI system serves several purposes: it creates a business record of intent, key decisions, and helps align system features with business and regulatory goals. In SDLC/AI governance circles, this is known as a "design baseline" or "design artifact." This documentation should begin early in the system's conception and capture design decisions as they emerge.

Key elements to capture in design documentation include:

- **System purpose and scope:** Define the AI system's primary functions, decision-making roles, and context of use.
- **Stakeholders and roles:** Identify who will be affected by or interact with the AI system, including users, customers, operators, third parties, as well as those responsible for building, operating, and maintaining it.
- **Ethical and regulatory considerations:** Document how the system aligns with principles such as fairness, transparency, accountability, privacy, robustness, and compliance with applicable laws.
- **System boundaries and interfaces:** Clarify what the system will and will not do, how it will interface with other systems, and where human oversight is expected.
- **Data and model design**: Capture data sources, quality criteria, labeling strategies, model selection rationale, target performance, and explainability constraints.
- **Risk and lifecycle management**: Identify foreseeable risks, misuse cases, safety constraints, fallback mechanisms, and plans for monitoring, retraining, and eventual retirement.

For example, an AI system designed to screen loan applications should document its data sources, how fairness will be ensured in the decision-making process, and what recourse applicants have to challenge decisions.

Minimum Viable Documentation for AI Design

Capturing the following elements will provide significant value to the development project:

- Problem statement (as defined and expressed by the business, not by developers)
- Governance and compliance context
- Intended users and use cases
- Assumptions and constraints
- System inputs and outputs
- Training data to be used
- Initial risk analysis
- High-level model or architectural design

Documentation in the Development Phase

The development phase transitions requirements and design into code and hardware. This is where decisions are made regarding system architecture, data ingestion pipelines, feature engineering, and initial model selection. In this phase, organizations should develop:

- **Model cards and system cards:** Discussed in earlier chapters, these are essential AI-specific artifacts that detail an AI model's characteristics, intended use cases, limitations, ethical considerations, and performance.
- **Architecture diagrams:** Include high-level and detailed visuals of system components and their interactions, with version control.
- **Code and model repositories:** Include proper security, source control, including branches, change history, and contributor records.
- **Tooling and dependencies:** Track software libraries, cloud services, and frameworks used, all of which are critical for debugging, patching, and future maintenance and development efforts.
- **Security considerations:** Ensure code and data are protected during development, including any use of sandboxing or secure enclaves and vulnerability scanning. All these artifacts should reside in official business record repositories and be subject to formal review and approval.

An example team builds a natural language processing (NLP) chatbot. It documents the preprocessing steps applied to user queries, the tokenization method used, and any proprietary language models being fine-tuned.

Documenting the Training and Testing Process

Once the AI system's architecture has been finalized and approved, the next phase involves model training, validation, and testing, which are critical stages for refining model performance, detecting bias, and validating outputs. Complete documentation at this stage helps ensure that testing results are repeatable, policy and regulatory obligations are met, and unexpected behaviors can be identified and resolved.

Training Documentation

Training documentation should include details on how the model was trained, as well as background information, including the rationale behind specific training decisions and how performance tradeoffs were addressed. Training documentation should include:

- **Training data sources and characteristics:** Include all data sources, exclusions, embellishments, filtering, and the population(s) represented. All training data should be archived and immutable to ensure future retraining or troubleshooting is possible.
- **Labeling process and quality assurance:** If supervised learning is involved, record how labels were assigned and verified.
- **Model training configuration:** Include hyperparameters, batch size, epoch count, optimization strategy, and evaluation metrics and results.
- **Model performance benchmarks:** Baseline performance, ongoing metrics, and information in support of selected thresholds.
- **Bias detection and mitigation:** Capture fairness assessments, de-biasing techniques, and tradeoffs.
- **Testing outcomes**: Record robustness, generalization, and privacy/security test results to ensure the model performs reliably under expected and unexpected conditions.

For example, a team training a computer vision model to detect defects in manufactured parts creates documentation describing how 20,000 images were sourced from different production lines, labeled by trained technicians, and augmented using rotation and contrast filters.

Regulatory Expectations for Training Logs

In high-risk AI systems under the EU AI Act and other frameworks, training documentation should support:

- Comprehensive documentation sufficient to reproduce the design and development effort
- Explanation of design rationale
- Long-term auditability
- Evidence of risk mitigation

Testing and Validation Documentation

Testing reveals a model's reliability and safety. Like training, it must be well-documented, to verify performance and to provide evidence of risk management. Testing documentation should include the following:

- **Testing methodology:** Include the types of tests that were run (unit, integration, performance, adversarial, and so forth). NIST ARIA, discussed in Chapter 7, includes guidance on model testing.
- **Test data description:** State whether test data was withheld from the training set and whether it represents real-world conditions.
- **Test results:** Include accuracy, false positives/negatives, precision/recall, and other metrics.
- **Stress testing and edge case analysis:** Note boundary conditions, adversarial examples, and inputs and outputs outside expected norms.
- **Failure modes and responses:** Include known weaknesses and the details of the system's behavior in the event of failure.
- **User testing and feedback:** Include details on how usability and human-AI interaction were evaluated.

For example, an organization developing an AI-powered hiring tool tests the system before deployment using a synthetic dataset to explore how it handles edge cases, such as incomplete resumes, inconsistent formatting, or unusual job histories.

Testing for Trust

Trustworthy AI requires more than just high accuracy. Testing should answer the following:

- Does the system fail gracefully?
- Can outputs be explained or justified?
- Are error patterns acceptable under ethical and legal standards?

Maintaining Traceability and Audit Readiness

Thorough documentation of the development of any system is essential, and arguably critical for AI systems, given the unique array of risks associated with AI. Comprehensive documentation enables organizations to trace AI behaviors back to their origin, a crucial capability when responding to audits, public inquiries, or litigation. Traceability builds trust, supports compliance, and facilitates more effective long-term risk management.

Version Control and Change Management

All changes in requirements, design, data, and model configurations should be tracked meticulously. Essential practices:

- Maintain linked *version control* across code, data, and documentation
- Require peer review and formal approval for key changes
- Ensure requirements are linked to design elements, test plans, and test results
- Capture not just decisions but the rationales and the outcomes
- Tag releases with associated risk assessments or approvals

Explainability and Documentation Artifacts

As a model feature, explainability has documentation implications. Organizations need to create and preserve artifacts that help them interpret how AI models made decisions. Essential artifacts include:

- Annotated decision flows or diagrams
- *Feature importance* summaries
- Case-level justifications or reversals
- Human-in-the-loop (HITL) escalation paths and results
- Model and system cards

Integration with Broader Governance Frameworks

As mentioned throughout this book, AI governance should be integrated with other supporting governance, including IT, cybersecurity, and privacy. AI system documentation should connect to broader organizational governance systems, including:

- Risk management
- Model inventories
- Incident response
- Ethical review board proceedings

Aligning documentation with internal controls helps make governance more repeatable, auditable, and defensible. When documentation practices are integrated with enterprise risk and compliance systems, they not only support regulatory preparedness but also create a scalable foundation for future AI development.

Business Context and Use Cases

An AI system's effectiveness depends on both its technical performance and whether it effectively serves its intended business purpose. Defining the business context and the specific use case for an AI model is a foundational step in responsible development, and it must be formally completed before finalizing requirements and design, as well as before development. Context and use case(s) anchor the system's organizational objectives, clarify performance expectations, and highlight operational, regulatory, and ethical constraints. Without a documented business

context and use cases, how can an organization be expected to develop a suitable AI system? This section examines the business context and outlines the use case, as well as how to evaluate the broader readiness and feasibility of deploying AI in real-world environments.

Defining the Business Context

Business context is a foundational element of AI system development, as it defines the organizational, strategic, and operational environment in which an AI system will be used. Similarly, a *business case* is a statement of the expected benefits to the organization that are anticipated to be realized as a result of developing an AI system.

Understanding Organizational Objectives

Every AI project should begin with clearly articulated business objectives—the reasons for the AI system and its expected contributions to the organization. This helps ensure that technical teams are not creating a system in search of a problem but are instead building a solution aligned with strategic needs.

Key considerations in support of a business context include:

- The problem that the AI system is intended to solve
- Expected success metrics from a business perspective
- How the system will support broader organizational goals

An example insurance company strives to improve fraud detection and intends to deploy a machine learning system to flag suspicious claims. The business goals are to identify outliers and reduce fraud losses by a measurable percentage, while minimizing false positives.

Organizational Readiness and Governance

Before starting development, organizations must assess their ability to support and operate the AI system, including infrastructure, policies, leadership alignment, and business processes. Considerations include:

- Leadership support and ongoing budget commitment
- Data governance maturity
- Cybersecurity and privacy maturity
- Risk tolerance and ethical culture
- Availability of cross-functional expertise (legal, IT, data science, operations, cybersecurity, and privacy)

AI Project Risk Scoring

Some organizations assign risk tiers to AI projects during planning. A high-impact, customer-facing model may require stronger documentation, pre-deployment review,

and executive sponsorship than a low-risk internal tool. Risk tiers defined in the EU AI Act (covered in Chapter 6) are a good start, even for organizations not subject to it (future regulations may use the EU AI Act as a model).

Stakeholder Identification and Engagement

Understanding who will interact with, benefit from, or be affected by an AI system is essential for shaping its scope, requirements, constraints, and evaluation methods. Stakeholders generally include:

- Business sponsors (usually, senior executives)
- System developers and data scientists
- Legal and compliance teams
- Cybersecurity and privacy teams
- Data governance
- Regulators and external auditors
- End users and customers
- Individuals indirectly impacted by decisions (for example, rejected applicants)

For example, for a university admissions AI model, the stakeholders identified include admissions officers (users), applicants (subjects), regulators (oversight), and institutional leadership (sponsors).

Articulating the AI Use Case

A well-defined *use case* translates the business context into a high-level specification for an AI application. The use case describes how the system will operate, what data it will use, and what outcomes it is expected to deliver. This allows stakeholders to assess feasibility, resource requirements, compliance requirements, and ethical implications before committing to design and development.

Defining the Scope and Boundaries

It's critical that the organization defines what the AI system will do, and what it won't do. Clear definitions help avoid scope creep, mismatched expectations, and unmanageable risk. The definition of scope should include:

- **Intended function** (for example, classification, prediction, assistance, recommendation)
- **Intended users and environments**
- **Assumptions and constraints** (for example, data sources, infrastructure, regulatory constraints)

- **Decision authority** (whether the system acts autonomously, or if there is a human in the loop)
- **Expected system lifetime, workload, and frequency of use**

An example retailer plans to use AI for dynamic pricing and should specify whether pricing decisions will be applied only online or in-store, whether prices will be updated hourly or daily, and whether human review is required for significant deviations.

Defining Success Metrics

The lack of clear metrics introduces the risk that an organization will optimize an AI system for the wrong outcomes. Metrics should be aligned with business goals, realistic given available data, and sensitive to unintended consequences. Example metrics may include:

- **For fraud detection:** Reduction in undetected fraud, false positive rate
- **For chatbot efficiency:** Resolution time, escalation rate, customer satisfaction
- **For hiring tools:** Quality of hire, time to fill, bias across protected classes

Metric Pitfalls

Relying too heavily on precision can lead to an increase in false negatives. Optimizing for efficiency might reduce transparency. Always weigh tradeoffs.

Assessing Data Availability and Suitability

The failure of many AI projects is not due to flawed models or designs but rather to poor data availability or misalignment between the data and the business objective. Evaluating data suitability early in the project ensures the use case is feasible and risks are manageable. Data suitability should be finalized before development efforts begin.

Inventorying Available Data

Organizations must assess what data they already have, where it resides, and how it was collected. Further, data that is a candidate must be evaluated for completeness, accuracy, consistency, and label availability if the data is to be used for supervised learning.

For example, a bank looking to predict loan defaults may discover that customer income data is inconsistently collected, requiring either additional data cleaning or revised model assumptions.

Data Governance Is a Key Prerequisite

Organizations lacking mature data governance policies, practices, and tools will be at a decided disadvantage, as they will have only partial awareness of their data inventory and management processes.

Legal and Ethical Constraints on Data Use

Even if suitable training data exists, there may be legal or ethical reasons why it cannot be used as is. Key concerns may include:

- Data collected for one purpose reused by the AI system without subject consent
- Inclusion of sensitive attributes (for example, race, health, or religious affiliation data)
- Inferences about individuals that may violate privacy expectations

Business Impact of GDPR Article 5(1)(b)

The "purpose limitation" in the GDPR can prevent organizations from using historical data for training, unless a new lawful basis for processing can be established.

Data Readiness for Deployment

Organizations need to assess whether the data ecosystem can support ongoing data needs, not only for training but also for monitoring, retraining, and compliance. Considerations include:

- Whether the data pipeline is stable and scalable
- Whether the model will be fed new, updated data over time
- Whether real-time inference can be supported, if required
- Whether monitoring is in place to detect data drift or degradation
- Whether continuous data ingestion complies with privacy, security, and access-control requirements

Ethical and Operational Considerations

It is distinctly possible that an AI use case that is feasible and valuable will still raise ethical or operational red flags. These essential considerations help an organization ensure that deployment is responsible, sustainable, and aligned with organizational values.

Risk of Harm and Ethical Tradeoffs

Even if legal, some use cases may lead to outcomes that harm individuals, exacerbate inequality, or reduce agency, any of which can enrage affected parties and draw the attention of regulators. Common areas of concern include:

- The use of AI in criminal justice, housing, hiring, education, or credit
- Reinforcement of historical bias
- Opacity in high-stakes decisions
- Algorithmic manipulation of behavior (for example, addiction loops in social media)

For example, an AI-powered mental health chatbot might reduce costs but fail to escalate users in crisis appropriately, creating serious risk. Escalation protocols may be critical in use cases like this.

Human Oversight and Decision Rights

Every AI system that makes business decisions should be designed as a part of a human-centric decision-making process. Consider the following:

- What decisions can AI make autonomously?
- When must humans be involved in decisions?
- How are overrides or appeals handled?
- Is the human oversight role well-defined and resourced?
- Are there sufficient business records to defend decisions and other actions?

RACI Matrix for AI Oversight

Here are suggested roles for human oversight of AI systems:

- Responsible: Data science lead
- Accountable: C-level product owner
- Consulted: Legal and ethics committee, privacy officers, security officers, and risk managers
- Informed: Frontline managers

Workforce Readiness and Culture

When AI systems are introduced into organizations, they don't just change workflows; they also alter roles, job requirements, and organizational culture. Points to consider regarding organizational readiness and culture include:

- Whether impacted teams have been consulted and their feedback heard
- Whether reskilling or new hiring is needed
- Whether job aids, support systems, and escalation paths are in place
- Whether regular employee engagement indicates changes in morale

Alignment with Organizational Values and Strategy

An AI use case may be technically excellent and legally sound. However, it may not be aligned with the organization's brand identity, reputation, or public commitments. Points to consider include:

- Whether the AI system or a specific use case contradicts public statements or corporate objectives
- Whether the AI system will erode trust with customers or partners
- Whether the AI system could introduce reputational risk disproportionate to its value
- Whether the staff will see the project as consistent with the company's values or could it trigger internal pushback

Impact Assessments

Impact assessments are a cornerstone of responsible AI governance. Whether required by regulation or driven by internal policy or risk management results, impact assessments help organizations anticipate, measure, and mitigate the potential consequences of deploying AI systems prior to their deployment. Performing or reviewing an impact assessment is not a one-time checkbox activity but a structured and iterative process that aligns the AI system with ethical principles, legal and regulatory requirements, and stakeholder expectations.

Understanding AI Impact Assessments

The primary goal of an *AI impact assessment (AIIA)* is to identify and manage potential harms or unintended consequences before deployment. These can include discrimination, lack of transparency, safety failures, excessive surveillance, or harm to fundamental rights.

When documented as a repeatable process, impact assessments provide a consistent method for risk identification and encourage transparency and accountability. When documented, impact assessments serve as documentation for internal governance and regulatory compliance. They can help improve stakeholder trust and public acceptance.

For instance, a public sector agency deploying AI-powered facial recognition technology uses an impact assessment to explore privacy risks, potential bias, and the need for public consultation. The results of the assessment lead to the implementation of new safeguards prior to deployment.

Legal and Regulatory Drivers

Many jurisdictions now mandate or recommend impact assessments for AI systems, particularly those deemed high-risk. Examples include:

- **EU AI Act:** Article 6 requires "conformity assessments" and documentation for high-risk AI systems.
- **Canada's Directive on Automated Decision-making (DADM):** Article 6.1 requires Algorithmic Impact Assessments (AIA) for AI decision-making systems.

- **California CPRA (and similar laws):** Grant regulators authority to require risk assessments for automated decision-making.
- **OECD and G7 principles:** Encourage impact assessments as a best practice.

Not All Risk Assessments Are Equal

While cybersecurity, data protection, and AI impact assessments may overlap, each focuses on different domains of risk. Aligning them can streamline governance and reduce redundancy.

When and How Often to Conduct Assessments

AI impact assessments should be performed during the AI system's design phase prior to deployment and whenever significant changes are made to the model or its data, as well as in response to incidents. High-risk and high-impact AI systems should be assessed regularly.

Preparing for an Impact Assessment

Organizations need to prepare for upcoming impact assessments. Key considerations include defining the scope, identifying stakeholders, and determining the necessary information. These should be documented prior to the start of the assessment.

Define the Assessment Scope

The scope of an assessment ensures it stays focused and aligns with the AI system's intended use. The definition of scope should include:

- System description and functionality
- Business context
- Use cases (for AI systems with more than one)
- Users and affected populations
- Geographic and legal jurisdiction(s)
- Level of autonomy and human oversight in supported business processes
- Assumptions, constraints, and exclusions

For example, an AI model used for employee performance review may need to be assessed differently in the EU (due to GDPR Article 22 on automated decisions made solely by AI) than in the United States, where labor laws differ.

Identify Stakeholders

Business and technical stakeholders should be included in the assessment team to ensure that all perspectives are represented. Assessment participants may include:

- Project sponsors and product owners
- Developers and data scientists
- Legal, security, compliance, and privacy officers
- Domain experts (for example, ethics, DEI, accessibility)
- Representatives of impacted groups, if feasible

Some organizations may want to include external stakeholders, such as advisors, consultants, vendors, and members of the public.

Inventory Required Information

Before an impact assessment can proceed, the team should assemble foundational information, including:

- Organizational documentation (policies, processes, code of ethics)
- System documentation (design, functionality, requirements, intended use)
- Applicable laws, regulations, and other legal obligations
- Stakeholders (identification of users, operators, and affected parties)
- Data flows, sources, provenance, and types
- Model performance metrics and known limitations
- User experience details
- Previous incidents or public concerns in similar domains

Conducting an Impact Assessment

The impact assessment process itself is a series of steps: identifying potential harms, evaluating risks, documenting mitigations, and capturing residual concerns. This section outlines common assessment steps and tools.

The process for conducting AIAs should be documented, including identifying responsible parties, and the location where results will be stored.

Identify Potential Impacts

The AIA begins by mapping where and how the AI system might create risk or harm. Areas of potential impacts include:

- Privacy and surveillance
- Theft of intellectual property or privacy data

- Discrimination and bias
- Freedom of expression or association
- Autonomy and human dignity
- Psychological harm or manipulation
- Access to rights or opportunities
- Environmental impact

For example, an AI-powered employment screening tool may penalize applicants with gaps in their employment history, indirectly punishing caregivers, people with disabilities, or veterans.

Evaluate Risk Likelihood and Severity

For each impact that's identified, it is essential to assess:

- **Likelihood:** Probability of the impact
- **Severity:** The level of harm to individuals or groups
- **Scope:** Number of people affected
- **Reversibility:** Whether harm can be undone

Scoring tools, such as risk matrices or heat maps, can help prioritize identified risks.

Analyze Discrimination and Fairness Risks

Due to the nature of AI models, bias and fairness risks require special attention. AI systems can discriminate, even when protected attributes, such as race or gender, are excluded from the training data. Assessment considerations include:

- Whether the system disproportionately affects protected groups
- Whether fairness metrics are analyzed
- Whether proxy variables or historical biases were embedded in training data

Even neutral input data may produce discriminatory outputs if AI systems are trained on biased historical patterns. It is essential to assess outcomes, not just features.

Assess Explainability and Transparency

In business, it's often critical for the organization to be able to understand and explain how an AI system did what it did. If stakeholders can't understand how or why an AI system makes specific decisions, the organization may suffer from trust and accountability concerns. Here are some important considerations:

- Whether end users can understand how the system works
- Whether decisions can be explained to affected individuals
- Whether documentation is available to regulators or auditors

For example, a credit union that wants to implement an AI-driven loan application evaluation system must include requirements concerning explainability, along with documentation and records to demonstrate this.

Review Human Oversight Mechanisms

AI systems should not be launched without ongoing human oversight, including human-in-the-loop capabilities for AI-driven actions and decisions. It's vital to be able to determine whether the system provides for effective human intervention. Some questions to ask during an assessment include:

- Are humans "in the loop" or "out of the loop"?
- Can decisions be reviewed, appealed, or reversed?
- Are personnel trained and empowered to override outputs?
- Are decision overrides guided by policy and documented?

Oversight Isn't Just a Checkbox

In a managed AI-driven business process, a human reviewer must have time to intervene, the authority to override, and visibility into how decisions were made. Oversight, whether it's human-in-the-loop or post-decision review, must be a standard business process, not a once-in-a-while or incident-driven activity.

Document Mitigation Strategies

When an AIA identifies risks, each identified risk should be documented, analyzed, and accompanied by a mitigation plan. Examples of risk mitigation include:

- Removal of sensitive features or proxy variables
- Applying fairness-aware algorithms, adversarial debiasing, or explainability techniques
- Addition of user consent or opt-out mechanisms
- Regular audits or bias testing
- Escalation of high-risk decisions to human reviewers

Like mainstream risk management, some risks may be accepted, and all instances of risk acceptance should be documented, including the party that approved the risk. Residual risks should also be documented, along with any accepted tradeoffs.

Reviewing and Operationalizing the Assessment

Like any risk assessment, an impact assessment is valuable only if its findings are reviewed, acted upon, and integrated into broader AI and risk governance. This final section describes how to ensure assessments are meaningful and actionable.

Internal Review and Signoff

Once completed, assessments should be reviewed by individuals or committees with governance authority (those authorities should be documented in AI governance documentation). Activities during reviews may include:

- Legal and regulatory compliance check
- Risk acceptance or escalation process
- Ethics review (if applicable)
- Executive or board-level endorsement for high-risk systems

In some organizations, the review and signoffs of AIAs for high-risk AI systems may be assigned to higher levels of organizational leadership than for lower-risk systems. AI system risk tiering should be extended to AIAs, considering how frequently and rigorously they are performed, as well as the level of management required to review and approve them.

Integration with Development and Deployment

Findings from the impact assessment must feed into development and deployment processes. This is where the AIA feedback loop leads to actual changes in an AI system. The activities that may result from an AIA include:

- Changes to model features or training data
- Adjustments to user interfaces for better explainability
- Revisions of deployment timelines or rollout strategies
- Additions of new monitoring or alerting mechanisms

For example, the deployment of an AI content moderation system is delayed after the impact assessment identifies a risk of over-censorship in minority languages, resulting in a retraining effort for the model.

External Transparency and Accountability

Depending on the regulatory context and contractual obligations, organizations may be required to publish the results or summaries of their impact assessments. Available methods include:

- Public summaries (as seen with Canada's Algorithmic Impact Assessments)
- Submission to regulatory bodies (for example, under the EU AI Act)
- Stakeholder briefings or community consultations

These techniques are often used for other forms of audits and assessments, including penetration tests, risk assessments, Sarbanes-Oxley, and PCI DSS audits.

What Not to Publish

Organizations refrain from publishing proprietary model architectures, protected data, or excessive details in the AIA that could be exploited by others if known.

Continuous Reassessment

Because AI systems change over time, impact assessments must be considered as standard periodic activities, and even as continuous monitoring. AIAs should be revisited when:

- AI models are retrained or redeployed
- New data sources are added
- Data filtering techniques are changed
- Use cases or environments change
- New risks emerge in the field
- New regulations are enacted

For example, an AI-powered traffic monitoring system initially trained on dry-weather data must be reassessed after it is extended to snowy regions, where camera visibility and road behavior differ significantly.

AI System Design Considerations

Designing an AI system is not only a technical challenge but a multidimensional process that integrates ethical reasoning, regulatory constraints, stakeholder engagement, and operational foresight. A thoughtful design process helps ensure that the AI system will be effective, fair, secure, compliant, and aligned with business goals and objectives.

This section addresses both technical and business process aspects of design. Certainly, any business process that is to be augmented by an AI model will need to be revised to address interactions with the model.

Defining the Purpose and Framing the Problem

Design begins with intentionality and clarity. A deep understanding of the AI system's purpose ensures that developers and stakeholders share a common understanding of its goals and limitations. In short, when success is defined in advance, designers will have a clearer idea of how to get there. Some questions to ask:

- What business or operational problem does the system aim to solve?
- Is AI the appropriate tool for this problem?
- What are the intended and unintended outcomes of the system?

AI systems often fail because the business problem was misframed, or embedded assumptions were not validated. The design phase is the right time to surface these assumptions. In fact, the design phase is the final opportunity to question everything before the system is built.

For example, a municipal agency wants to predict where potholes will emerge. Framing this as a predictive maintenance problem, rather than a real-time detection task, leads to a different model design and data strategy.

Architecture and Model Selection

Organizations must consider not just the design of the AI model but the full architecture: how components interact, how data flows, and how results are delivered. The choice of architecture and model must be driven by purpose, constraints, and risk tolerance.

Mature organizations often develop both a logical architecture and a physical architecture, allowing the system's functions to be defined at a conceptual level, followed by the design of the technical environment.

Designing for Modularity and Traceability

Modern information systems are often modular, which facilitates increased flexibility over the long run. Modular architectures support scalability, troubleshooting, model updates, and explainability. Essential practices include:

- Separate data ingestion, model inference, and output rendering
- Enable independent testing of each component
- Log internal system decisions and state transitions
- Documentation of all module- and system-level interfaces
- Track versions of data pipelines, models, and dependencies in registries for reproducibility and rollback
- Assign ownership and change-approval responsibilities for each module to ensure accountability

Choosing the Right Model Type

The AI model should be fit to effectively address the business problem's complexity, interpretability needs, and data environment. The speed at which AI models and types are evolving makes it difficult to enumerate the array of potential choices, but this is not the purview of AI governance; instead, qualified experts should explore the types of candidate AI systems and make a defendable, repeatable decision on the type to be implemented.

A few considerations include:

- High-stakes AI-driven decisions may demand simpler, explainable models.
- Resource-constrained environments may favor lightweight or cloud-based models.
- Proprietary vs. open-source choices may affect transparency.

Organizations are frequently attracted to "shiny objects," or new technologies and systems that staff want to get their hands on. Staff members are often clouded by bias rooted in the need to remain current and familiar with emerging technologies. Without proper precautions, organizations may find themselves choosing a solution based on its "cool factor" rather than its real suitability.

Avoiding "Model Envy"

The latest deep learning model may not be the right fit for an organization's specific business problem. Simpler models can outperform complex ones when data is structured, features are well-defined, and explainability is crucial.

Every model selection decision involves tradeoffs: there are no perfect solutions. Decisions must be documented and justified during the design phase. Some AI model tradeoffs include:

- Accuracy vs. interpretability
- Speed vs. cost
- Fairness vs. personalization
- Data volume vs. privacy
- Data volume vs. security
- Risk minimization vs. innovation

For example, a bank selects a random forest AI model for credit scoring because it provides a good balance between accuracy and explainability, which is critical for regulatory compliance.

Human Oversight and Feedback Loops

AI systems in socio-technical environments require human intervention to play key roles in monitoring, intervening, and improving outcomes. Sound system and process design anticipate and support these roles.

Designing Human-in-the-loop Processes

Organizations must determine the role humans will play and at what stage(s) in the process workflow they will intervene. Options for roles include:

- **Approver:** Worker must sign off on AI output before action commences.
- **Reviewer:** Worker reviews audit system output periodically.
- **Escalation point:** System or affected party flags uncertain cases for human review.

For example, a healthcare diagnostic AI system flags cases with low confidence scores for radiologist review, improving safety, trust, and confidence in results.

Designing for Usability and Understanding

Human oversight is only effective if they understand what the system is doing and their role(s) in ongoing system operations. Considerations include:

- Use of clear language, not black box scores
- Provide clear, contextual explanations
- Allow easy overrides with rationale logging

Feedback Integration

User feedback is a valuable signal that should be incorporated into system improvement. Organizations should seek feedback from the users of their AI systems in several ways, including:

- Online surveys
- Focus groups
- Explicit feedback ("This decision was incorrect")
- Implicit signals (for example, user overrides or workflow abandonment)
- Formal review cycles with business stakeholders

Feedback from users should trigger retraining, bug fixes, or policy review and not just be logged and ignored.

Metrics, Thresholds, and Controls

Design decisions determine how success and failure can be measured. Metrics must reflect business goals, user and/or societal expectations, and operational constraints. Performance thresholds should be adjustable and defendable, and a part of the design, rather than determined after deployment.

Metric Selection

Different problems require different metrics. Poorly chosen metrics can lead to unintended behavior or the organization being unaware of significant developments.

Metrics is a big topic. I suggest you purchase a copy of *How to Measure Anything in Cybersecurity Risk*, 2nd Edition; *Operational Risk: Measurement and Modelling*; or *Key Performance Indicators: Developing, Implementing, and Using Winning KPIs*, 4th Edition, all from Wiley Publishing, for a comprehensive review of metrics, to assist in metrics development.

Threshold Setting and Adjustment

Thresholds in metrics determine when actions should be taken. These must be aligned with risk tolerance, regulatory requirements, and ethical norms. Essential design practices concerning thresholds include:

- Calibrate thresholds based on business impact and harm potential
- Require formal approval for threshold changes
- Document rationale for chosen settings

Operational Controls

Organizations need to implement AI-specific controls as a part of responsible AI use. Controls should be a part of an AI system design for monitoring, failover, and rollback from the start. Some AI-related controls may include:

- Alerting for unusual model behavior
- Failsafe defaults (for example, deny by default)
- *Shadow mode* deployment before production
- "Kill switch" to disable faulty models

Controls are another big topic: they must also be implemented across the entire tech and business operational stack, to ensure sound management of information systems, personnel, and third parties. Books written by this author on the CISM (Certified Information Security Manager) and CRISC (Certified in Risk and Information Systems Control) certifications contain comprehensive discussions, models, and practices on controls.

Controls in the EU AI Act

High-risk AI systems must include human oversight mechanisms, logging, and post-deployment monitoring—all of which depend on decisions made during the design phase.

Identifying and Managing Design Risks

AI system design introduces a unique set of risks—some inherited from traditional software development, while others stem from the probabilistic and adaptive nature of machine learning. These risks can manifest internally (for example, design flaws, poor data quality, or integration failures) or externally (for example, misuse, bias, or systemic harm). Proactively identifying and managing these risks during the design phase is essential to building systems that are trustworthy, resilient, and aligned with organizational values.

Organizations that wait until AI systems are built and deployed risk disruption through expensive rework that could have been avoided if risks had been identified in the design process or earlier.

Framing AI Design Risk

Unlike deterministic systems, AI systems introduce *nondeterminism*, *opacity*, and *emergent behavior*, complicating risk identification and assessment. AI systems introduce new risks, including:

- Unpredictable behavior in unseen data conditions
- Difficulty tracing decisions to specific inputs or logic
- Bias or harm embedded in training data, particularly by proxy variables
- Drift in model behavior over time
- Overreliance on automation without proper oversight

An example AI system trained to predict employee attrition may inadvertently learn to penalize candidates from smaller universities or underrepresented backgrounds due to historical data bias.

Distinguishing Between Internal and External Risk Sources

Internal risks (those in the organization and in its control) arise from within the system's design, architecture, development process, technology environment, and related business processes. These risks may include:

- Faulty assumptions
- Inadequate documentation
- Inadequate training data
- Poor integration
- Incomplete testing
- Model performance gaps
- Lack of explainability or transparency
- Inadequate monitoring for drift or misuse

External risks (those outside the organization and largely out of its control) may include:

- Adversarial use and cyberattacks
- Legal or regulatory changes
- Supply chain risks and breaches
- Cultural backlash or ethical misalignment
- User misuse or automation complacency

Internal ≠ Controllable

Even though the organization creates internal risks, they are not always entirely within its control, especially when using third-party components or pretrained models.

Systematic Risk Identification Techniques

A standard and structured approach to risk identification reduces blind spots. This section covers methods and tools that help teams systematically surface design risks.

Risk Matrix

One of the most widely used tools in risk management is the *risk matrix*, which helps prioritize risks based on their probability of occurrence and potential impact or harm. This technique is used in cybersecurity and cyber risk management. Table 8.1 depicts a typical risk matrix.

For example, suppose an AI model used in medical imaging has a low probability of misclassifying rare conditions but high severity when it does. In that case, it may require layered oversight, even if errors are rare.

Risk Mitigation Hierarchy

A hierarchy-based approach prioritizes interventions from most to least effective:

1. **Eliminate the risk** (for example, don't use AI in high-risk contexts)
2. **Substitute with a lower-risk design** (for example, use simpler, interpretable models)
3. **Engineer safeguards** (for example, include confidence thresholds, human-in-the-loop controls)
4. **Warn and train users** (for example, provide interpretability dashboards)
5. **Accept with monitoring** (for example, include audit logs, post-market review)

TABLE 8.1 Typical Risk Matrix

Risk Impact	Low Probability	Medium Probability	High Probability
Low severity	Monitor	Low priority	Medium priority
Medium severity	Low priority	Medium priority	High priority
High severity	Medium priority	High priority	Critical

This is a simple depiction of risk treatment, the formal process of determining what to do about an identified risk. Risk treatment is a component of risk management, which is discussed at length in other books written by this author on the CISM (Certified Information Security Manager) and CRISC (Certified in Risk and Information Systems Control) certifications.

Tradeoffs in the Real World

Elimination or substitution are ideal concepts but often impractical due to business demands. Risk mitigation in AI frequently involves strategic layering of controls.

Stakeholder Mapping and Risk Perception

Across the organization, different stakeholders perceive risk differently. Mapping stakeholder roles, incentives, and exposure helps identify overlooked risk vectors. The types of risks that are relevant to stakeholders include:

- **Users:** risk of error, confusion, overreliance, hallucination
- **Subjects:** risk of unfair decisions, privacy harm
- **Operators:** risk of workload mismatch, liability, costs
- **Regulators:** risk of non-compliance, lack of transparency
- **Public:** reputational or ethical concerns

Use Case Evaluation and Scenario Planning

The specific use case(s) and environment(s) where an AI system will be deployed significantly influence the types and severity of design risks. Scenario analysis and edge case testing help reveal context-specific risks.

Use Case Classification and Risk Profiling

Not all use cases are created equal. The frequency and rigor of risk assessments should be proportionate to the level of risk. Some of the factors to consider include:

- Decision impact (low, medium, high)
- User exposure (internal vs. external)
- Affected rights or opportunities
- Automation level (advisory vs. autonomous)
- Regulation (for example, EU AI Act, various privacy laws)
- Contextual factors (cultural norms, societal scrutiny, or prior incidents in the same domain)

For instance, an AI used for Netflix recommendations may warrant lighter-touch risk controls than one used for unemployment eligibility decisions.

Scenario Planning and Adverse Events

AI system design teams and other stakeholders should honestly ask: "What could go wrong?" Scenario planning includes both expected failures and edge cases. Techniques include:

- Red teaming and adversarial testing
- "Premortem" exercises (envisioning failure and tracing causes)
- Misuse and attack role-play with diverse personas

Designers of a predictive policing tool for an example school district create scenarios in which the model over-targets specific neighborhoods due to biased historical data, leading to design changes and an external review.

Benchmarking and Comparative Risk Analysis

Organizations may want to compare the AI systems they plan to develop to existing benchmarks or similar implementations. Potential sources include academic benchmarks such as ImageNet or GLUE, or peer organizations using similar AI systems.

Benchmarking is another tool that helps inform design and can also help defend decisions during audits or litigation.

Testing, Piloting, and Validation

AI system testing is not a single event: it is a design tool. AI system testing is not conducted all at once but rather in stages throughout the development cycle. Structured testing and controlled pilots help validate assumptions, reveal hidden risks, and generate the evidence base needed for deployment decisions.

Unit Testing

If the AI system is part of an ecosystem, organizations should consider unit testing, in which individual components are tested in isolation, to determine whether they operate as expected and as designed.

For example, a credit union plans to integrate an AI loan evaluation tool into its back-office loan application system. Before performing any end-to-end testing, the credit union opts to test the portions of the loan application system to ensure it creates the correct data to be passed to the AI system. Similarly, the credit union also simulates AI system outputs to test the loan application system to see whether it responds properly.

Pre-deployment Pilots

A *pilot*, which is the release of a production system to a limited set of users for final readiness testing, simulates real-world deployment in a controlled, limited environment. Considerations for pilot testing include:

- Start with "friendly" users
- Define success and failure criteria

- Include diverse users and contexts
- Monitor for unexpected behaviors
- Plan for post-pilot review to capture issues
- Determine which, if any, issues require changes to the AI system before general release

For example, a chatbot for customer support is tested with employees and volunteers before full rollout to identify gaps in response quality, escalation logic, and tone.

Shadow Mode Testing

Some organizations opt to deploy their AI systems in shadow mode, *integrating* them into existing business processes to capture and analyze outputs. This technique is similar to *parallel testing* performed in IT disaster recovery planning, where a backup information system is fed the same inputs as live production systems to verify that the backup systems function correctly.

Some of the benefits of shadow mode testing include obtaining live performance data, verifying that all *integrations* function correctly, and determining whether AI systems yield the same results as their human counterparts.

For example, a bank deploying an AI loan review system runs it in shadow mode testing mode for 60 days, comparing AI suggestions with actual underwriter decisions. Discrepancies inform model retraining and explainability improvements.

Stress Testing and Red Teaming

Stress testing and red teaming tests are essential for understanding an AI system's response to adversarial, edge case, and malicious inputs. Types of stress tests may include:

- Injection of out-of-distribution data
- Adversarial inputs crafted to trick the model
- Simulated outages or data loss
- Drift simulations to assess how the model handles changing input distributions

Ideally, organizations will have performed stress testing and red teaming, such as *threat modeling*, prior to development. Threat modeling is a common activity in cybersecurity, where analysts examine a system's requirements, architecture, and/or design to identify potential vulnerabilities and threats.

Risk Governance Integration

Design-phase risk management must connect to the organization's broader risk and compliance infrastructure. Without this integration, even well-identified risks may go unmitigated. Organizations will fare better when their various risk management disciplines are integrated into a single, broad process.

Alignment with Enterprise Risk Frameworks

Organizations should track AI design risks in the same systems and formats used for other operational risks. This includes the use of a single risk register, common scoring rubrics, assignment of risk ownership, and escalation paths.

For instance, a telecommunications firm incorporates AI-related risk into its ISO/IEC 27005-based information security risk process, ensuring consistent treatment across projects.

Documenting Residual Risk and Justifications

Not all risks can be eliminated. When an organization identifies a risk and decides to mitigate the risk, there is often "leftover" risk (like change from your dollar) known as *residual risk*. Residual risk should be documented. If the magnitude of the residual risk is great enough, it should be treated as yet another risk, like any other, and analyzed, and a decision made on what to do about it.

Don't Hide Residual Risk

Residual risk should not be swept under the rug. Regulators and stakeholders are often more concerned with undisclosed risks than with the existence of risk itself. Transparency builds trust.

Preparing for Post-deployment Risk Monitoring

Risk assessments are not only appropriate (daresay, essential) in the design and development phase of an AI system but equally necessary throughout the AI system's service life. Hence, design risk controls must extend beyond deployment. How to prepare:

- Drift detection
- Retraining thresholds
- Monitoring dashboards
- Incident response workflows, just in case
- Feedback loops from users or regulators

For example, an organization's e-commerce platform uses AI model anomaly detection to flag changes in recommendation quality or user engagement after deploying a new ranking algorithm.

Proprietary Model Considerations

An organization that chooses to develop and deploy a proprietary AI model enjoys greater control but also must accept greater responsibility. Unlike the use of third-party AI solutions, proprietary models involve end-to-end ownership of requirements, design,

training, deployment, and all operational and legal outcomes. A proprietary AI model, if you can find the talent to build it, can offer competitive advantage and customizability, but it also introduces heightened obligations, regulatory scrutiny, and liability exposure.

Ownership and Accountability

Organizations that create proprietary AI models and place them into production use are accountable for the entire model lifecycle, from initial design to long-term maintenance. Accountability includes:

- Ethical alignment and governance
- Data sourcing and legality
- Staff to design, build, and operate the model
- Model performance, fairness measurement, and maintenance
- Transparency and documentation for use in audits, legal proceedings, and other purposes
- Post-deployment behavior and monitoring
- System disposal at the end of its useful life

For example, a healthcare provider that builds its own diagnostic model cannot shift blame to a vendor if the model misdiagnoses a condition. The organization is solely accountable for outcomes and auditability. However, upon closer analysis, even if an organization utilizes a third-party AI system, it must still accept a considerable share of accountability; after all, the organization vetted and selected the third party. In short, accountability cannot be outsourced.

Legal Liability and Civil Exposure

Proprietary models can increase exposure to civil and regulatory liability, especially if the model's use causes harm. There is simply no one else to blame.

Sources of risk include:

- Algorithmic bias or discrimination
- Security vulnerabilities
- Decisions affecting protected rights (for example, housing, employment, and credit)
- Failure to meet industry-specific legal obligations
- Deficiencies in explainability or documentation

The EU AI Act and Proprietary Models

Under the EU AI Act, developers of high-risk proprietary models face stricter obligations, including:

- Conformity assessments and demonstration
- Technical documentation
- Logging and traceability
- Quality management
- Human oversight mechanisms
- Incident reporting

In many cases, these cannot be delegated.

Regulatory Classification as "Provider"

In regulatory frameworks such as the EU AI Act, an organization that develops and deploys its own AI model is typically considered the "provider," not just the "deployer" (a person or legal body using an AI system under its own authority). This shifts the burden of compliance upstream to the organization running its proprietary AI model.

Responsibilities as a provider include:

- Performing risk classification
- Conducting impact assessments
- Maintaining documentation
- Ensuring cybersecurity and robustness
- Providing disclosures and notices

An example insurance company develops a proprietary underwriting model. It must ensure it complies with both sectoral regulations and AI-specific legal requirements, even if the model is not sold to or used by others.

Data Rights, Quality, and Governance

Data is a part of the foundation of every proprietary AI model. When organizations develop their own AI systems, they must assume full control and legal authority over the data they use, along with sound governance processes for its handling.

Data Sourcing and Legal Entitlement

Using internal or scraped data without a proper legal basis can expose the company to significant risk, particularly when such data is subject to privacy regulations, such as the GDPR. Organizations should ensure that data is collected with appropriate consent, and usage rights include AI training; they should also document the lawful basis for its use.

Internal Data ≠ Safe Data

Just because data is held internally doesn't mean it's legally usable. Repurposing data without proper notice or consent may violate the principles of data minimization and purpose limitation.

Data Quality and Labeling Integrity

Low-quality or biased data can undermine performance and increase the downstream risk of complaints from citizens and potential regulatory action.

Data-related controls for AI system design include:

- Enforce data validation
- Document labeling methodology and *inter-rater reliability*
- Log data transformations and preprocessing steps

For example, an organization implemented a proprietary AI model designed to detect defects in industrial parts. The AI model failed in production due to inconsistent labeling across training datasets. Documentation of labeler guidelines and audits was missing.

Data Governance Integration

As mentioned elsewhere in this book, data governance is an essential prerequisite for organizations that intend to implement AI systems. Data governance practices ensure safe, responsible, and compliant uses of data. Essential practices include:

- Defining AI-specific data stewards
- Tagging training and validation datasets in data catalogs
- Applying data retention and deletion policies
- Aligning with broader data ethics frameworks

Privacy is a critical component, closely tied to data governance. Compliance with applicable privacy laws (for example, GDPR, CPRA) imposes requirements on organizations that their AI systems must support.

Model Explainability, Testing, and Monitoring

"We don't know how or why our AI system unfairly declined a loan application for our customer," is a statement that no organization wants to make publicly. But if a proprietary model involved in decision-making is not explainable, sooner or later an organization is going to get into trouble. Proprietary models must be explainable to users, defendable to regulators, and maintainable over time. Organizations that build their own systems are responsible for embedding these qualities from the start.

Explainability Obligations

Many regulatory frameworks (for example, GDPR, CPRA, EU AI Act) require that automated decisions affecting individuals be explainable. It is reasonable to expect that new regulations in other jurisdictions will follow suit.

Approaches for explainability include:

- Use inherently interpretable models (for example, decision trees, linear models)
- Implement post hoc explanations (for instance, SHAP, LIME)
- Offer counterfactual or case-based reasoning
- Provide user-facing justifications in plain, friendly, but defendable language

For instance, a financial services firm uses a proprietary credit scoring AI model and must provide applicants with reasons for denial in a manner that is both accurate and understandable.

Testing and Validation Requirements

Unlike third-party systems, which are tested and validated by the vendor, organizations developing proprietary models must rely on their internal teams to design and execute testing regimes. Testing regimes should include:

- Bias audits across protected attributes
- Performance validation on real-world scenarios
- Adversarial robustness testing (red teaming)
- Human-in-the-loop failover testing

Monitoring and Change Control

Proprietary models require ongoing *incident management* and a formal process for *change management*. Indeed, the entire practice of *IT service management* (which includes change and incident management, among several other practices) is essential for every organization that builds and operates a proprietary AI model. Monitoring practices include:

- Setting alert thresholds for performance degradation
- Logging inputs and outputs for traceability
- Detecting data drift
- Retraining models regularly

Change control practices include:

- Formal written change requests with stated business reasons
- Approvals by stakeholders
- Approved implementation and backout plans
- Implementation of *configuration management* tools to record change details

For example, a retailer retrains its AI-powered recommendation engine monthly but fails to document hyperparameter changes. Unexpected customer behavior results, and regulators request evidence that changes were risk-assessed. The documentation gap exposes the company's negligence.

ISO Certification for IT Service Management

Organizations in regulated industries, as well as those that want to demonstrate competence in their technical operations, may wish to pursue *ISO/IEC 20000* certification to attest to their IT service management. Note that ISO/IEC 20000 is not AI-specific but applies generally to IT.

Intellectual Property and Competitive Risk

Proprietary AI models may provide strategic advantages, but they also introduce new risks around *intellectual property (IP)*, secrecy, and potential *reverse engineering* or misuse.

IP Ownership and Protectability

Organizations should, as a part of their overall data governance, maintain an accurate catalog of their intellectual property. This practice enables organizations to handle such information properly with minimal risk of compromise. Organizations should establish clear ownership of AI assets and associated data.

Steps to consider on the protection of IP include:

- Whether the AI model was developed in-house, or through contractors or consultants
- Whether the training data is proprietary, licensed, or open-source
- Whether model weights and code are documented and secured
- Whether trade secrets are adequately protected

Patents vs. Trade Secrets

AI models are often protected as trade secrets rather than patented, especially when model transparency might expose security or privacy vulnerabilities. This approach limits legal remedies in some cases.

Reverse Engineering and Model Theft

Proprietary AI models may be vulnerable to reverse engineering by competitors or malicious actors. In the author's opinion, commercial software and SaaS companies do a somewhat better job of ensuring their products are resilient than organizations that develop systems

in-house. The point here is that it's not so much that the AI model is proprietary but that any in-house developed system is apt to be less resilient than one produced by a software vendor.

Mitigation strategies include:

- Deploy only model APIs, not source code.
- Employ robust access control on all APIs.
- Use authentication, encryption, and obfuscation for model artifacts and API traffic where possible.
- Monitor for unauthorized access, abuse, and attack.
- Red team and threat model often.
- Audit usage logs and limit the rate of queries.
- Implement privacy-preserving techniques (differential privacy, response randomization) and output watermarking to deter model theft.

An example fintech company discovered that a competitor was able to infer the logic of its AI-powered, proprietary fraud detection system through repeated API calls and replicate its model with minor modifications.

Model Misuse and Access Control

Organizations need to control who can use or modify a proprietary model. Robust access control (itself a broad, critical practice) is essential to preventing abuse or accidental damage. Considerations include:

- Implement "need to know" and "least privilege" concepts.
- Apply role-based permissions for development and deployment.
- Log changes to model weights or logic.
- Utilize internal approval workflows for model updates.
- Implement revocation procedures for unused or compromised models.
- Periodically review privileged access to the AI system.

Resourcing and Organizational Readiness

Like any internal systems development effort, the development of proprietary models is a resource-intensive process that requires sustained investment. Organizations must ensure they have the infrastructure, talent, and governance to support proprietary AI at scale.

Many organizations underestimate the level of effort required to build and sustain proprietary AI models. AI is not unique in this respect, as the art of estimating software projects has been a challenge for organizations since the 1960s. We're no better at estimating the level of effort for large, complex projects using new technology than before.

Technical Infrastructure

Custom AI models typically require more demanding infrastructure than off-the-shelf solutions. Organizations must exercise caution and closely manage project expenditures. And it takes a special type of courage to stop an AI project if it has gone off the rails.

Infrastructure needs include:

- GPU/TPU compute clusters
- Scalable data pipelines
- Secure model hosting environments
- Sandboxes for development and testing
- Separate development and testing environments

Talent and Skill Development

To successfully undertake a large-scale system development project, organizations require senior-level team members to lead the effort. Internal development increases reliance on specialized talent, including data scientists, ML engineers, and model governance professionals. Because of the high demand for AI talent, these skills are in short supply and are priced accordingly. Training the right internal team members can help fill this gap.

Talent management ideas include:

- Retention and compensation strategies for key technical staff
- Cross-training between AI and compliance teams
- Internal knowledge transfer and documentation
- Establishment of an AI "center of excellence" to foster collaboration and learning

The Risk of a Single Point of Failure

When only one or two individuals understand how a proprietary model (and any other internally developed system) works, the organization is exposed to continuity and risk management challenges if those individuals leave.

Governance and Ethical Review Capacity

Organizations that develop proprietary AI models will need to establish additional governance workflows. Given that traditional software development and AI model development share DNA, a mature SDLC may be able to be adapted to AI model development, as opposed to building a separate, similar development lifecycle process from scratch. Keep in mind these guardrails:

- Implement approval boards for internal models.
- Use expanded ethical review criteria.
- Integrate AI governance with legal, risk, and audit functions.
- Designate "model owners" as responsible for ongoing oversight.

For example, a tech startup adds a formal AI Governance Board to review proprietary models through each stage of development, including signoffs from legal, security, and ethics leads.

Summary

Designing sound, resilient AI systems responsibly requires comprehensive documentation and contextual understanding across the entire lifecycle. From initial concept to post-deployment oversight, each phase introduces decisions that shape performance, compliance posture, and societal impact. A structured approach to designing AI systems, emphasizing lifecycle documentation, business alignment, risk awareness, and tailored controls for proprietary models, helps ensure favorable outcomes.

Documenting the AI lifecycle is critical for traceability, compliance, and future auditing. The design and build stages should include clear statements of the AI system's purpose, stakeholder roles, data sources, model architecture, and decision rationales. Prior to training and testing, documentation must capture dataset provenance, preprocessing steps, performance metrics, and risk mitigation efforts. When training data includes privacy data, organizations must formally establish a lawful basis for processing. Thorough records not only help validate the system's outcomes but also provide the basis for explainability and regulatory defensibility throughout the model's operational life. The outcome of an AI system should, to the greatest extent possible, be determined prior to development.

Effective design also begins with a clear understanding of the business context and intended use case(s). Organizations must identify what problem(s) the AI system is solving, who it serves, and how its outputs integrate into operational workflows. Aligning the system's capabilities with business goals, performance thresholds, data availability, and ethical considerations is critical for success. Stakeholder engagement is vital, as users, impacted populations, and oversight bodies each bring valuable insights about constraints, risks, and unintended consequences. A well-articulated use case ensures that technical choices align with real-world constraints and deliver measurable outcomes.

Impact assessments serve as a vital milestone in the design process, helping organizations uncover potential harms, compliance gaps, or ethical blind spots before the AI system is implemented. These assessments evaluate factors such as bias, explainability, autonomy, and fairness across use cases. A structured impact assessment process should involve diverse stakeholders, evaluate mitigation strategies, and document accepted, resolved, and residual

risks. Findings from these assessments should inform design revisions, user interface choices, human-in-the-loop safeguards, and deployment policies. In jurisdictions like the EU and Canada, these assessments are increasingly becoming a legal requirement. All these activities should be documented.

Designing for fairness, safety, trustworthiness, and resilience requires more than good intentions; it involves specific policies and best practices throughout the design phase. These practices include thorough requirements gathering, model architecture selection, metrics definition, human oversight design, and feedback loop integration. Systems should be evaluated not just for technical performance but for alignment with organizational values and operational sustainability. Design should anticipate the need for continuous monitoring, retraining, and adjustment, embedding those controls into the system from the outset. As design moves through its phases, formal approval should be utilized to keep the project on track.

Organizations developing proprietary models face an elevated set of obligations and risks. End-to-end ownership means they are solely accountable for ensuring legal compliance, achieving ethical outcomes, and ensuring model explainability. Proprietary development also introduces concerns around data rights, intellectual property, infrastructure, and workforce capabilities. AI governance must be more robust, testing more thorough, and documentation more complete and precise. In many regulatory regimes, including the EU AI Act, proprietary model developers are considered "providers," and must satisfy stricter reporting, traceability, and oversight requirements. Organizations with mature SDLCs may be able to extend them to include the development of proprietary AI models.

Exam Essentials

Documenting the AI lifecycle ensures traceability, accountability, compliance, and audit readiness. From system design to testing, comprehensive documentation provides the foundation for governance, compliance, and future maintenance. It supports explainability, allows teams to replicate decisions, and enables post-incident analysis when systems fail or behave unexpectedly.

Defining the business context and use case(s) aligns AI development with real-world needs and constraints. AI systems must solve clearly and formally defined problems, serve specific user groups, and operate within set performance boundaries. Understanding business objectives, stakeholder roles, and the operational environment ensures that design choices are relevant and accountable.

Impact assessments are not optional for high-risk AI systems but are a legal and ethical necessity. Impact assessments identify potential risks, harms, fairness concerns, and gaps in oversight. They are required under frameworks like the EU AI Act and Canada's DADM and serve as a central governance tool to ensure that the AI system aligns with legal, ethical, and organizational expectations.

Design decisions must balance technical performance with ethical, human-centered, operational considerations. Choosing a model architecture, defining metrics, and establishing thresholds often requires tradeoffs. These choices should be informed not just by accuracy but by explainability, fairness, robustness, and usability. Impactful design decisions should be made by appropriate leadership and documented. Design also includes creating effective human-in-the-loop mechanisms and feedback paths.

Organizations deploying proprietary models assume greater liability and governance responsibilities. Unlike off-the-shelf tools, proprietary AI systems place full lifecycle accountability on the development organization. This includes legal exposure, regulatory classification as a provider, and the burden of maintaining documentation, oversight, and audit readiness internally.

Review Questions

The following questions are designed to test your understanding of this chapter's material. For more information on how to obtain additional questions, see this book's Introduction. You can find the answers in Appendix A.

1. What is the primary purpose of documenting the AI system lifecycle?
   - **A.** To accelerate deployment timelines
   - **B.** To reduce storage costs
   - **C.** To enable traceability and support governance
   - **D.** To comply with open-source licensing requirements
2. Which factor is most important when defining the business context for an AI system?
   - **A.** The organizational objective(s) the system supports
   - **B.** The number of employees on the data team
   - **C.** The vendor's market share
   - **D.** Whether the system uses neural networks
3. An AI impact assessment should be conducted primarily to:
   - **A.** Document proprietary model architectures for publication
   - **B.** Increase investor visibility
   - **C.** Certify the model for ISO 27001
   - **D.** Identify and mitigate risks associated with AI system use
4. What is a typical reason for integrating human-in-the-loop (HITL) mechanisms into AI design?
   - **A.** To speed up model inference
   - **B.** To eliminate the need for documentation
   - **C.** To allow for human oversight and intervention in critical decisions
   - **D.** To reduce power consumption
5. Which of the following is a disadvantage of developing a proprietary AI model?
   - **A.** Increased liability and regulatory obligations
   - **B.** Reduced system accuracy
   - **C.** Increased dependency on open-source data
   - **D.** Inability to access APIs

6. What does a probability/severity matrix help teams accomplish during design risk assessments?
   - **A.** Optimize GPU usage
   - **B.** Score and prioritize potential risks
   - **C.** Determine intellectual property value
   - **D.** Choose a programming language
7. Which of the following best supports explainability in AI systems?
   - **A.** Obfuscation of model weights
   - **B.** Publishing training code without documentation
   - **C.** Feature importance scoring and user-facing explanations
   - **D.** Encrypting data pipelines
8. Why is stakeholder mapping critical during risk identification?
   - **A.** It replaces the need for testing
   - **B.** It ensures that all server endpoints are documented
   - **C.** It helps classify model types by algorithm
   - **D.** It identifies how different groups perceive and are affected by AI system risks
9. Which of the following would best demonstrate readiness to deploy a proprietary AI model?
   - **A.** Having a high-performing model with no documented metrics
   - **B.** Completing shadow mode testing with documented outcomes
   - **C.** Outsourcing testing to a third party without internal review
   - **D.** Publishing a press release
10. Under the EU AI Act, a company that develops and deploys its own AI model is considered which of the following?
    - **A.** Consumer
    - **B.** Deployer
    - **C.** Provider
    - **D.** Tester



Chapter

9

Data Governance and Model Training

THIS CHAPTER COVERS AIGP DOMAIN III, "UNDERSTANDING HOW TO GOVERN AI DEVELOPMENT," SUBDOMAIN III.B, "GOVERN THE COLLECTION AND USE OF DATA IN TRAINING AND TESTING THE AI MODEL." THE FOLLOWING TOPICS ARE COVERED:

- ✔ Data governance concepts and practices
- ✔ Data provenance and data lineage
- ✔ AI system testing
- ✔ Identifying and managing testing risks

THE AI GOVERNANCE PROFESSIONAL (AIGP) PERFORMANCE INDICATORS IN THIS SUBDOMAIN INCLUDE:

- ✔ Establish and follow the requirements for data governance (for example, assess and document lawful rights to collect and use data, and to assess data quality, quantity, integrity and fit-for-purpose).
- ✔ Establish and document data lineage and provenance.
- ✔ Plan and perform training and testing of the AI model (for example, unit, integration, validation, performance, security, bias and interpretability).
- ✔ Identify and manage issues and risks during training and testing of the AI model.



The quality, legality, and trustworthiness of an AI system depend heavily on training data and the processes that shape it. This chapter explores the foundations of data governance, data lineage, model training, and testing, the core functions that determine whether AI systems are safe, fair, and fit for purpose. You will learn how to assess data quality and lawful usage, trace the origin and transformation of data, and implement structured testing across both technical and ethical dimensions. The chapter also highlights how to identify and manage risks that emerge during model development, including bias, performance failures, and explainability gaps. Whether an organization is building its first AI system or refining an existing one, these practices are a guide for aligning AI outputs with responsible development goals.

Data Governance

AI systems are only as reliable and resilient as the data on which they are trained. For this reason, data governance stands as a critical enabler (even a prerequisite) of responsible AI development. While an AI system may technically function without formal governance structures, its trustworthiness, accountability, integrity, and effectiveness will suffer. And, many organizations either do not practice data governance at all or they implement it carelessly. These shortcomings create vulnerabilities, including biased or inaccurate training sets, non-compliance with legal obligations, and data integrity failures, among others. A strong, effective data governance framework helps reduce these risks and sets a more solid foundation for ethical, legal, and high-performing AI.

Key components of data governance include:

- *Data classification*: A scheme of assigning risk levels to datasets, together with handling and governance requirements for each level
- *Data flow diagrams (DFDs)*: Developed by data or system architects, these are visual representations of the flow of data within and between systems and entities
- *Input controls*: To ensure that data flowing into a system is approved by management and also is formatted correctly
- *Output controls*: To ensure that data flowing out of a system is approved by management and formatted correctly

- *Data loss prevention (DLP)*: Software that monitors the movement of data, to alert or block certain movement if it violates policy

This section provides a brief review of data governance and its application to AI systems. For a comprehensive review of corporate data governance, refer to *Practitioner's Guide to Operationalizing Data Governance* or *Data Governance for Dummies* from Wiley Publishing.

Defining Data Governance for AI

Data governance refers to the framework of managing enterprise data through policies, processes, tools, roles, and responsibilities to ensure data is accurate, usable, secure, and compliant with relevant requirements and applicable regulations. In AI development, governance ensures that training, testing, and operational datasets are not only technically suitable but are legally and ethically sound and will be used appropriately.

Data governance should encompass issues such as who owns the data, how it was obtained, whether it can be used for AI development, how it should be protected, and what applicable laws, regulations, and requirements stipulate for proper protection and use. Without these guardrails, organizations risk building AI systems on faulty or unlawful foundations.

Legal Authority and Lawful Use of Data

The first step in data governance in the context of AI is the confirmation that the organization has a defendable legal right to use it.

Confirming Lawful Rights to Use Data

Organizations must formally verify that they are authorized to collect, process, store, and analyze the data that will feed into AI systems. These rights may be based on one or more of the following:

- Direct user consent (for example, checkbox acceptance in an app)
- Implicit user consent (included in terms of use; implicit consent is insufficient in some jurisdictions, including the EU via GDPR)
- Contractual provisions (for example, data-sharing agreements)
- Privacy policy or notice of privacy practices declarations
- Regulatory allowances (for example, HIPAA-compliant health data sharing, which often requires deidentification and *business associate agreements (BAAs)*)

Formal legal counsel review of all of these, together with a written assertion, is considered essential for an organization to move forward with the use of data to train an AI system.

Documentation Requirements

Once lawful access is established, organizations must document it with as much detail as required by regulation, policy, or organizational culture. Data governance policies should include:

- Data source registration
- Consent records (if applicable)
- Contract references
- Privacy policy/notice of privacy practices references
- Jurisdiction-specific legal justifications

For example, a med-tech company purchases a dataset containing anonymized patient records for predictive modeling. Months later, a government investigation reveals the data was re-identifiable under specific conditions. The company had not verified the vendor's anonymization process or legal basis for selling the data. The project is shut down, and a substantial fine is imposed.

Secondary Use Risks

Some organizations repurpose data beyond its original collection and consent purpose. Organizations lacking data governance will likely experience *function creep*, leading to violations of privacy policies, contracts, or regulations. And without data governance, this can all occur with little or no awareness.

Review processes and impact assessments performed as part of data governance should identify high-risk secondary uses.

When Using Third-party AI Systems

Organizations planning to utilize third-party AI systems must consider several additional factors. In particular, when an organization trains a third-party AI system with personal information, it must explicitly understand whether that training data will be used elsewhere. Depending on the level of risk, an organization in this position may want to impose specific terms in the legal agreement, including:

- Training data may not be used for any other system, customer, or purpose.
- The organization retains the *right to audit* the third-party organization and related systems.
- Audits and cybersecurity assessments must be performed by qualified third parties.
- The organization reserves the right to terminate the agreement without cause.
- The vendor must agree to destroy data upon request and at the termination of the agreement.

This topic is discussed in more detail in Chapter 11.

Assessing Data Quality and Integrity

While establishing a lawful basis for data is essential, lawfully acquired data can still undermine an AI system if it lacks quality or integrity.

The Dimensions of Data Quality

A data governance policy should define data quality and outline the process for evaluating it. Key dimensions include:

- **Accuracy:** Are the values contained in the dataset correct and free from error?
- **Completeness:** Are all required fields populated?
- **Integrity:** Are relationships across datasets correctly maintained?
- **Timeliness:** Is the data up to date?
- **Consistency:** Do values follow expected formats and rules?

Fit for Purpose Means More Than Accuracy

An AI model can be technically excellent but still fail if it does not align with its intended use. Fitness for purpose includes compliance with regulations, integration with workflows, and suitability for the operational environment. An AI system must deliver value in the context of its use, not just in test results.

Detecting and Preventing Data Drift

For models that update or learn over time, governance must include mechanisms to detect and respond to data drift, which occurs when new output data deviates significantly from the original distributions. Monitoring protocols should be mandated as part of AI model lifecycle governance.

Ensuring Sufficient and Representative Data

Data sufficiency refers to having a sufficient volume of data to train a model effectively. *Representativeness* refers to ensuring that the data accurately captures the relevant diversity of the environment in which the AI will operate.

Avoiding Underfitting and Overfitting

An insufficient volume of training data can lead to *underfitting*, where models are unable to generalize effectively. Too much data that is either narrow or repetitive may cause *overfitting*, where models learn irrelevant patterns. Effective data governance helps find the right balance by defining sufficiency thresholds and reuse limits.

Representing Real-world Populations

To avoid the appearance of bias, AI governance frameworks should incorporate fairness checks to detect and address representation gaps. Examples include demographic audits, data distribution checks (such as geographical distribution), and detection of outliers.

For example, an organization implemented an AI résumé screening tool and trained it on historical hiring data; the AI system heavily favored male candidates for technical roles. A post-deployment audit reveals that training data underrepresented women applicants. Governance could have caught the bias earlier had it looked for it.

Fit-for-purpose Evaluation

For an AI system to succeed, the data used for its training must not only be clean and representative, but it must also be appropriate for the AI model's task and context.

Governance reviews should evaluate whether the dataset supports the AI model's objective. For example, using social media sentiment data to predict voting behavior may be inappropriate without cross-validation against polling or demographic controls.

Taking data intended for one purpose and using it for a different purpose should be approached with caution and be subject to review and approval.

AI Training ≠ General Use

A dataset used to train an AI system on language translation may not be suitable for training a chatbot in a medical advice setting. Even technically similar data can diverge in ethical implications and legal exposure.

Data Stewardship and Roles

Effective data governance requires clearly defined roles and responsibilities. Two primary roles are:

- **Data stewards:** These stewards are operational custodians responsible for ensuring data quality, documentation, and resolving issues. They ensure that governance policies are consistently applied throughout the data lifecycle.
- **Data owners:** Owners are accountable for the strategic use of data. They approve use cases, certify lawful use, and sign off on risk assessments. A governance framework should identify data owners by domain or dataset.

These activities should all be part of documented and recorded business processes. Data owners and stewards should be identified wisely, and each should have a clear understanding of their responsibilities.

In an example insurance company, the underwriting department owns actuarial data, while the marketing department owns customer segmentation datasets. Without clear ownership, governance reviews will stall or be ineffective, and risk approvals will become bottlenecks.

Security, Privacy, and Ethical Use

To be effective, a data governance program must be integrated with cybersecurity and privacy, so that they operate as a cohesive structure to protect data and ensure its responsible, legal use. The introduction of AI intensifies this need, with its novel uses of data, where breaches and misuse can scale quickly.

Privacy Risk Assessments

A data governance program should mandate privacy impact assessments (PIAs) for all AI systems that will process personal data. These assessments evaluate risks of compliance, re-identification, inference, and discriminatory outcomes.

Security Controls

A wide variety of controls should be implemented to ensure adequate protection and proper use of AI-related datasets. At the data layer, controls such as encryption, access management, and audit logging should be required for any AI-related datasets, including training data. At the infrastructure layer, a comprehensive set of controls should be required to protect facilities, networks, systems, and database management systems from attack and misuse.

Books by this author on cybersecurity certifications, including CISSP, CISM, CISA, and CRISC, delve into the topic of controls in substantial detail.

Ethical Considerations

Due to the novel and emerging uses of data in AI systems, organizations need to incorporate ethical considerations into their data governance programs. An ethical review should address issues beyond those related to policies, requirements, and regulations. An ethical review may be considered the final approval step, as it is essential to consider that while a particular use of data may be allowed and legal, it may still be unwise. Examples of things to look for in an ethical review include:

- Training data that embeds racial, gender, or socioeconomic bias via proxy variables
- Training data that was collected under exploitative or disingenuous conditions
- Training data that invades privacy, even if it's technically anonymized or pseudonymized

For example, consider a controversial legal issue that involved an AI model that was trained on scraped online therapy forums. The data was public, but its use violated user expectations and sparked backlash. The moral of the story: the use of any particular dataset can be lawful and yet unethical.

Organizational Maturity and Implementation

Data governance is not a one-time compliance task but a lifecycle process that continues throughout the lifespan of an AI system—indeed, for the lifespan of the organization's use of data. To be effective and serve the needs of all parties, data governance requires leadership support and institutional buy-in, and may even precipitate cultural change.

Organizations may opt to centralize data governance under a single body or federate it across departments. While models will vary, they all require executive sponsorship, training and awareness, metrics, documented procedures, and official business records.

For organizations establishing new data governance programs where none previously existed, it is recommended that they emulate existing governance structures, borrowing what works well, and consider adjustments as needed to ensure success. For organizations using GRC platforms or similar systems to manage governance processes, it is recommended that data governance also utilize these systems for consistency and familiarity.

Organizations that focus on process maturity should consider adopting a maturity model, to help them assess and improve their governance posture. Levels typically include:

- **Ad hoc:** Minimal documentation or oversight
- **Emerging:** Some formal roles and basic policies
- **Defined:** Clear responsibilities and standardized procedures
- **Managed:** Monitoring, measurement, and risk-based decision-making
- **Optimized:** Continuous monitoring, automation, metrics, and risk-adjusted decision-making

An example shipping company with ad hoc governance launches an AI-powered delivery routing system. When delivery failures rise, it takes weeks for the organization to trace the issue to corrupt location data. A mature governance program could have flagged the issue more quickly, even in hours.

Common Pitfalls and Failure Modes

Even well-organized, mature organizations can struggle with implementation. Less mature organizations will be unable to escape a parade of problems. Here are some examples:

- **Treating governance as a compliance exercise:** Governance that focuses solely on ticking boxes, rather than enabling trustworthy data use, often leads to resistance and ineffective controls.
- **Overcentralization:** Excessively centralized governance structures, particularly in larger organizations, can create bottlenecks and lead to the disintermediation of data governance. The right balance must be struck between control and agility.
- **Tool overload:** Excessive use of governance tools and dashboards can overwhelm teams, leading to inconsistent usage. Simplicity and training are the keys to workable data governance practices.

- **Using shadow copies of training data:** An organization extracts production data from business applications and performs ad hoc filtering offline without recording the procedure. Data governance (if it exists at all) will be blind to this use of data.

Benefits of Governance in AI Development

When well designed and implemented, data governance does more than reduce risk: it adds value through improvements in:

- **Improved model performance:** Clean, representative data leads to better outcomes.
- **Faster audits and investigations:** Good documentation accelerates regulatory response.
- **Cross-team alignment:** Clear roles reduce conflict and rework.
- **Reputational advantage:** Demonstrated governance maturity builds stakeholder trust.

While not necessarily required by law, data governance should be considered a strategic imperative for organizations adopting AI systems. Organizations that forego governance often experience more and larger issues, including rework, litigation, and brand value issues.

An example tech firm that adopted *ISO/IEC 38505-1* data governance standards experienced reduced model training time, faster stakeholder signoff, and greater external confidence in its AI systems.

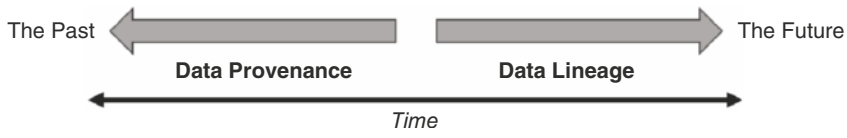
Data Lineage and Provenance

To build a trustworthy, responsible, and ethical AI system, organizations need to understand where data comes from, how it moves, and how it changes. The concepts of data lineage and data provenance form the backbone of this understanding. Although most organizations do not follow these practices, documenting lineage and provenance is crucial for accountability, regulatory compliance, and model explainability.

Without robust lineage and provenance tracking, organizations may struggle to explain how a model arrived at its output, assess the trustworthiness of the training data, or detect inappropriate or corrupted inputs. As AI systems increasingly influence high-stakes decisions, the need for traceability becomes more urgent. Lacking this tracking may result in peril.

Figure 9.1 depicts data provenance and data lineage.

FIGURE 9.1 Data provenance and data lineage.



Lineage and provenance tracking may feel like behind-the-scenes work (and so they are), but they are essential to building AI systems that are credible, defensible, and ethical. Organizations that neglect this domain expose themselves to failure, while those who prioritize it create AI systems worthy of trust, creating a competitive advantage in the private sector.

Defining Data Lineage and Data Provenance

It is essential to clearly distinguish data lineage and data provenance, as they are often used interchangeably but refer to distinct (though related) concepts.

Data lineage refers to the lifecycle of data within a system. Data lineage documents how data flows through pipelines, from ingestion, storage, transformation, consumption, maintenance, and disposal. Lineage tracks the "how and where" of data including:

- How it was generated (or from where it was obtained—this is where data lineage and data provenance meet)
- Where it was stored
- What processes transformed it
- Which models or systems consumed it
- What processes carried out modifications, corrections, and removals
- What processes carried out data retention operations
- How and when the data was disposed of when no longer needed

Data provenance refers to the origin and original characteristics of a dataset, including:

- From where the data was obtained
- Who created or collected the data
- Under what conditions was it collected
- What was its intended purpose

For example, an organization's customer behavior dataset used in an AI recommendation engine may have a clear lineage within the company's data lake. Still, its provenance may be unknown if it was initially acquired from a third party with insufficient documentation. Lineage ≠ provenance.

When viewed together, data lineage and provenance support a complete picture of data trustworthiness. They help organizations identify the source(s) of errors and bias, reproduce outcomes (for explainable AI models, at least), and demonstrate compliance with applicable regulations and other legal obligations.

The Role of Lineage and Provenance in AI

In AI systems, the consequences of poor lineage and provenance tracking can be significant. As AI systems are themselves force multipliers, the data-driven models they use amplify the quality, origin, and transformation history of the inputs they consume.

Or, put another way, garbage in, garbage out. Or, further, if you don't track the origin and transformations of AI system training data, how do you expect to understand or troubleshoot its results?

- **Explainability and trust:** When stakeholders question an AI system's output or decision, such as why a job applicant was passed over, its developers must be able to trace the model's input data. Without lineage and provenance, explanations become speculative, or worse.
- **Data security and integrity:** A lack of lineage can conceal data tampering, poisoning, and unauthorized modifications.
- **Compliance and auditability:** Many regulatory frameworks require that organizations document their data-handling practices. For instance, the EU AI Act requires traceability for both source data and its transformations.

Reproducing model behavior for debugging, retraining, or litigation requires a clear understanding of the data's journey and origin. Lineage and provenance support reproducibility and, therefore, accountability.

For example, a transportation agency deploys a predictive model for optimizing bus routes. Six months later, it wants to retrain the model but cannot locate the original dataset or reconstruct its feature transformations. The project is suspended until the agency can determine a way forward.

Capturing Data Provenance

Effective provenance tracking begins with *metadata* collection and documentation. Such metadata should include:

- Source entity (owner, organization, system)
- Collection method and date
- Version information
- Data licensing (terms of use)
- Access rights
- Intended use or original context
- Lineage
- Jurisdictional considerations (where collected, applicable regulations)
- Context (intended use, assumptions, limitations)

Metadata should be a required element that is a part of the overall documentation package for an AI system. AI governance and data governance can enforce this through policy.

Data obtained from third parties, including vendors, brokers, partners, or public repositories, may include additional risks. Data governance policies should require that

the organization obtain data provenance declarations from data sources, as well as proof of lawful collection and consent, and documentation describing any modifications or aggregations performed by the source.

For example, a system developer in an organization uses an open-source image dataset to train a facial recognition model. After deployment, the organization learns that the dataset included copyrighted images that were obtained without permission. As a result, the organization now faces legal exposure and potential damage to its brand.

Standards, including *W3C PROV* or *ISO/IEC 5181 (Artificial Intelligence—Data and AI system provenance)*, define formal models for provenance tracking. Adoption is limited, and ISO/IEC 5181 is still in the development stage. Still, these standards may provide direction for organizations that want to adopt a standard approach for collecting and documenting data provenance.

Tracking Data Lineage

While provenance looks backward to origin, lineage tracks the path forward in time. Lineage records should include:

- Data source and format
- Storage location
- Version history
- Transformation scripts or ETL (extract, transform, load) processes
- Data pipeline paths
- Systems and models using datasets
- Data retention and disposal processes

Backward lineage, sometimes confused with data provenance, traces each data point back to its origin. *Forward lineage* tracks how data elements flow through systems to influence downstream processes.

Modern data observability platforms, such as Monte Carlo, Collibra, or OpenLineage, offer visualization tools to represent lineage graphically. These tools are especially valuable when multiple teams interact with shared datasets.

An example company noticed a spike in false positives from an AI model, which were traced to a malformed transformation script introduced two weeks prior. Lineage tracking tools help isolate the change, enabling a rollback of the affected pipeline.

Common Failure Modes

Many organizations fail to implement effective lineage and provenance tracking due to structural and cultural barriers. Organizations either decide that lineage is unimportant or are unaware of the practice altogether. Some of the failure modes include:

- **Data and organizational silos:** Data silos and organizational silos inhibit effective lineage and provenance. When datasets are transferred between departments without clear ownership or established standards, lineage and provenance can both become unclear. Governance frameworks should require consistent documentation across silos.
- **Informal data use:** *Shadow AI* development, where data scientists implement AI models and manipulate datasets outside sanctioned pipelines, creates potential blind spots. Without governance, these ad hoc projects can feed unreliable data into unapproved production models, leading to inaccurate results.
- **Manual and fragmented processes:** If documentation is manual or scattered across multiple tools, tasks, and procedures, it is often overlooked. Automation can help provide consistent lineage and provenance capture.

As an example, a large retailer manually documents transformations in spreadsheets. An intern renames a feature column and forgets to update the spreadsheet. Months later, the error causes flawed inventory predictions during the holiday season.

Building a Provenance and Lineage Framework

Traceability doesn't occur in a vacuum; organizations need to be intentional about provenance and lineage, establishing policies, tools, and processes to operationalize traceability.

Establishing provenance and lineage begins with policy, which should define:

- Mandatory metadata fields for new datasets
- When and how provenance must be verified and approved
- Ownership responsibilities for provenance and lineage documentation
- Requirements for automated capture and audit trails
- Required granularity of tracking (for example, dataset-level, record-level, or feature-level lineage)

In the tech stack, a provenance and lineage framework may include:

- Metadata repositories (for example, data catalogs)
- Data lineage visualizers
- ETL tools with integrated tracking and logging
- Version control systems for datasets and transformation scripts
- Security controls (for example, encryption and access management) to protect lineage and provenance records themselves

Provenance and lineage data can be integrated with MLOps tools, such as MLflow, DVC (Data Version Control), or Amazon SageMaker, enabling traceability from raw data through training to deployment.

Many organizations use model cards (first discussed in Chapter 2) to describe model behavior in detail. Model cards can include references to lineage and provenance data, thereby increasing transparency for stakeholders and regulators, and linking AI models to their corresponding datasets.

Use Cases and Industry Examples

Traceability practices may vary by industry, but the principles of traceability are broadly applicable. Some examples include:

- **Financial services:** Regulators may require a clear lineage for all data used in credit scoring, fraud detection, and algorithmic trading. Documentation should indicate how decisions are made and what data influenced them.
- **Healthcare:** HIPAA compliance and patient data safety concerns require that all health-related AI systems can trace data back to validated sources. The provenance of clinical trial or electronic health record data is often scrutinized.
- **Public sector:** Transparency is critical in government AI projects. The auditability of data sources ensures democratic accountability, particularly in criminal justice or welfare automation.

An example city agency uses AI to flag at-risk youth based on school performance and neighborhood data. After concerns are raised, officials use documented lineage to demonstrate that the model did not utilize protected attributes, such as race or income level.

Benefits of Strong Traceability

Organizations that invest in lineage and provenance capabilities reap benefits across technical, legal, and business domains, including:

- **Reduced debugging time** and faster model issue resolution for issues related to training data
- **Greater confidence in model outputs** for stakeholders
- **Improved audit response** and regulatory posture
- **Higher data quality** results in fewer data-related issues, and more rapid issue resolution

Traceability Challenges and Tradeoffs

Building a comprehensive provenance and lineage program presents technical and organizational hurdles. The tradeoff comes down to creating the overhead of additional processes and tooling that provides vital data, or foregoing traceability to improve agility.

Automated lineage tracking can add latency to data pipelines and increase storage requirements. Governance teams must work with engineering to find the right balance between traceability and performance.

Organizations using third-party or open-source models may not be able to obtain full provenance or lineage data, as this could be considered intellectual property. Organizations should evaluate traceability when selecting external tools to ensure optimal performance and reliability.

Organizations need to determine the level of investment required to build infrastructure for metadata management, auditing, and visualization. Like other initiatives, traceability itself may warrant the development of a business case. Like similar initiatives, many organizations will under-resource these functions until forced to do so by an incident or investigation.

AI System Testing

Like any new or updated system or application, testing an AI system is essential to ensuring model quality, reliability, and trustworthiness. Unlike traditional software testing, which focuses on verifying expected outputs for given inputs, AI system testing encompasses a broader array of testing methods. These include statistical performance, fairness, security, explainability, and integration within larger systems.

And, like other business applications, AI systems also need to be tested iteratively and continuously, as models often evolve through retraining, fine-tuning, or adaptation to new data. Thorough testing during the training and deployment lifecycle verifies model behavior, and it uncovers risks and weaknesses that may otherwise remain hidden until post-deployment.

AI Testing Is a Discipline

AI testing is not a one-size-fits-all activity but a cross-disciplinary practice that blends business objectives, engineering, statistics, ethics, and domain expertise. When performed correctly, AI testing helps ensure that AI systems are not only functional but also fair, secure, compliant, and aligned with their intended business purpose.

As AI adoption continues to accelerate, organizations that use AI must elevate their testing practices from ad hoc validation to formal, documented, and auditable processes. Testing is not just a technical task but a governance responsibility and a driver of long-term trust.

Understanding AI Testing Objectives

AI system testing serves multiple objectives, including assessing accuracy, ensuring safety, evaluating fairness, and establishing stakeholder trust.

AI testing differs from traditional application testing because it involves *stochastic* behavior, data dependency, and shifting performance boundaries. AI system testing must address:

- Variability in model output
- Performance on unseen or edge cases
- Tradeoffs between performance and interpretability
- Compliance with evolving regulatory standards

Model testing that passes accuracy benchmarks may still produce biased or harmful outputs in production, if edge cases or fairness were not identified or adequately tested.

Testing must be performed throughout the AI lifecycle:

- During **training** (to monitor convergence and behavior)
- In **validation** (to verify outputs and decisions)
- **Pre-deployment** (to stress-test against adversarial or unexpected inputs)
- After **deployment** (to detect data drift or performance decay)

Each of these phases requires distinct test strategies and metrics.

Unit and Component Testing

Unit testing in AI serves to validate the functionality of individual components within the overall system, including data processing logic and model functions.

Input data transformations such as normalization, encoding, or imputation must be tested for correctness and consistency. Bugs in these steps can introduce problems in the downstream model.

Model wrappers (software that encapsulates an AI model), scoring functions, and evaluation logic should be isolated and tested, using synthetic inputs to verify correct handling of edge cases (for example, null values, extreme values, malformed input data).

Unit tests can be integrated into *continuous integration (CI)* tooling and workflows using testing frameworks (such as PyTest or unittest in Python), ensuring consistent behavior across code changes.

Integration Testing

Integration testing validates the interactions between components: data ingestion, model inference, storage, and downstream systems. Any point in the overall system where one component communicates with another is an item that should be tested, to ensure that each element works correctly and that the data or instructions passed between them are structured correctly.

Input data should be scrubbed to ensure that incoming data conforms to expected formats, values, and character sets, and that APIs or interfaces function as intended under

load. For instance, a model trained on numeric ages may fail when a live system provides age as a string (for example, "35 years"). Integration testing would identify type mismatches before the production launch. Interface testing is particularly critical when interfaces connect systems of different types (for example, between UNIX or Windows and mainframes).

Integration testing ensures that the model's predictions are correctly handled by applications, dashboards, or decision systems.

Trust but Verify

Integration tests should simulate full-system, end-to-end workflows, including failure conditions (for example, dropped records, unexpected input, or unavailable services), to validate system resilience under real-world conditions.

Validation Testing

The purpose of validation testing is to confirm whether the AI system performs as intended in real-world applications. Validation testing is performed after unit and integration testing, which check whether the system works technically. Validation testing goes further by comparing AI system outputs against business expectations, ethical requirements, and compliance standards.

Put another way, validation testing confirms whether the AI system solves the stated business problem(s).

Like other types of testing, success criteria should be defined before testing begins. Otherwise, it can be difficult to determine whether validation testing is considered successful.

Metrics associated with validation testing include *accuracy*, *precision*, *recall*, and *F1 score*.

For example, in a credit card issuer's fraud detection system, precision is more important than recall to avoid false accusations. AI model validation testing focuses on optimizing this balance.

Performance Testing

Organizations utilize performance testing to ensure the AI system can meet speed, scalability, and availability targets in its operational environment. Some of the methods include:

- **Inference latency:** The time it takes an AI system to return a prediction from an input is known as *inference latency*. Real-time systems require low-latency inference. Performance testing under simulated load reveals bottlenecks in model execution or I/O.
- **Throughput and scalability:** For batch-processing tasks, organizations test how many inputs the system can handle per unit time, known as *throughput*. Tools like LoadRunner, JMeter, or Locust can simulate load.

- **Resource utilization:** It's important to know how much CPU, GPU, memory, storage, and network resources an AI system requires. Organizations often perform average load and peak load tests, to ensure the system runs within infrastructure limits.

For example, a retail organization's product recommendation engine fails to deliver results under heavy holiday traffic. Load testing reveals the need for improved infrastructure that can operate under peak load.

Test for the Worst, Not the Average

A model that averages 50 ms inference time but spikes to 2 s under certain loads can still degrade the user experience.

Security Testing

While they have some unique properties, AI systems are information systems, which are inherently vulnerable to many forms of attack. In addition to conventional vulnerabilities, AI models are vulnerable to a range of security threats, including adversarial attacks and model inversion.

Some of the AI-specific testing that organizations should perform includes:

- **Adversarial testing:** Inject inputs specifically crafted to cause incorrect predictions. For example, adding subtle noise to an image can cause misclassification.
- **Data poisoning and backdoors:** Assess the integrity of training data to detect malicious inputs that can influence model behavior under specific conditions.
- **Model extraction and inversion:** Organizations can test whether attackers can reconstruct training data or reverse-engineer model behavior via repeated querying.

Organizations also need to perform comprehensive security testing, including:

- Vulnerability scans at the application, system, and network levels
- Penetration tests at the application, system, and network levels
- Supply chain security testing for third-party models and dependencies
- Physical penetration tests for on-premises systems
- Social engineering testing to test workforce resilience

An example facial recognition model is attacked with thousands of probing images. The attacker reconstructs training faces with high fidelity. Security testing could have identified this vulnerability, prompting a reduction in model exposure via rate limiting or differential privacy.

Bias and Fairness Testing

As discussed throughout this book, bias is a potential weakness inherent in any AI system, where its outputs systematically disadvantage specific individuals or groups. Some of the types of testing include:

- **Performance parity:** Measure how performance metrics (for example, accuracy, false positive rate) vary across demographic groups. Identify unjustified disparities that may result from proxy bias.
- **Counterfactual fairness:** Evaluate whether changing a protected attribute (for example, gender, religious affiliation, political affiliation) changes the model's output inappropriately.

Tools for Bias Testing include AIF360, Fairlearn, and the What-If Tool from TensorFlow.

For example, a bank's credit model rejects applicants of color at twice the rate of white applicants, despite equivalent credit profiles. Fairness testing reveals the gap and prompts a retraining effort.

Legal Exposure

In many jurisdictions, biased algorithms are not only unethical but also illegal. Fairness testing helps organizations identify fairness issues, enabling them to avoid regulatory fines and public backlash.

Interpretability and Explainability Testing

Interpretability testing is used to determine whether an AI model's behavior can be meaningfully understood by the organization, and hence can be explained to users, regulators, and developers. Types of testing include:

- **Model-agnostic methods:** Tools like SHAP or LIME can help explain individual predictions, which can aid in identifying spurious correlations or data artifacts.
- **Global interpretability:** It's essential to be able to summarize what an AI model generally values. Feature importance analysis, decision trees, or coefficient inspection (in linear models) can provide insight.
- **Simulatability:** Whether a human can replicate or understand model predictions is known as *simulatability*. Simpler models (for example, *logistic regression* or rule-based systems) may be favored in high-risk use cases for this reason.

An example healthcare AI system predicts heart disease risk. SHAP values show that postal code is a major driver, suggesting a potential proxy for race or socioeconomic status.

Human-in-the-loop and Scenario Testing

Many organizations implement AI systems that interact with humans in critical ways, including the review of an AI system's decision or other type of output. Testing should include scenario-based evaluations that simulate real-world user behavior and decision pathways. The types of testing include:

- **End-to-end user flows:** This test evaluates the system's behavior when interacting with real users. This testing can surface unexpected results due to UX design, miscommunication, or unexpected edge cases.
- **Interactive testing and feedback loops:** For AI systems that perform recommendations or assist in decision-making, the user's response may influence future system behavior. System testing should account for long-term dynamics and feedback effects.

An example of a human-AI medical decision support tool provides recommendations that doctors sometimes disregard. Post-deployment testing reveals that confidence scores were poorly calibrated, meaning that users did not trust the AI system.

Documentation and Traceability

All types of testing discussed in this section should be documented, including both testing procedures, test plans, and results. These testing artifacts support transparency, auditability, and reproducibility.

Every AI system's document library should include a test plan that defines:

- Objectives
- Scope and responsibilities
- Test scenarios
- Metrics and thresholds
- Pass/fail criteria

The results of all tests should be documented, including:

- Test results
- Issues encountered
- Remediation steps identified
- After-action review
- References to datasets, code versions, and environments used

Discussed elsewhere in this book, model cards and model fact sheets summarize model behavior, training/testing datasets, performance, intended use cases, and known limitations. They also provide stakeholders with a high-level understanding of what the model is—and isn't—designed to do.

For example, a model card for an organization's AI-powered resume screener discloses that performance degrades for applicants without college degrees. The organization flags this for analysis and potential mitigation in the next release.

Organizational Practices for AI Testing

AI testing must be routine, repeatable, documented, and embedded into team process workflows, accountability structures, and governance frameworks. Essential practices include:

- **Clearly designate responsibilities for testing.** In regulated sectors, this may involve third-party validation or separate risk teams.
- **Establish formal review gates in lifecycle development processes,** where testing evidence is evaluated before a model can proceed to deployment.
- **Organizations need to foster a culture** where testing is not just a formality but a key quality and risk control mechanism. Incentives, including rewards for issue discovery and transparency, help improve testing results.

Challenges and Tradeoffs

Organizations find that testing AI systems is challenging due to the probabilistic nature and real-world complexity of these systems. It's not always easy to design a test to address a specific concern.

For instance, in some AI use cases, there is no definitive "correct" answer, making validation inherently subjective and fuzzy. Also, fairness can be defined in multiple, often conflicting ways (for example, equal opportunity vs. equal outcomes), prompting organizations to decide which principles to prioritize. Finally, some high-performing models, such as deep neural networks, may be less explainable, making integrity and accuracy testing more challenging. At times, testing must strike a balance between performance, stakeholder trust, and regulatory requirements.

Identifying and Managing Testing Risks

As stated throughout this chapter, AI systems must be thoroughly tested before deployment; however, testing is not just about verifying functionality. Testing is also about identifying risks that emerge during training and validation; these risks may arise from flawed data, inappropriate metrics, poor test coverage, or misaligned performance expectations.

AI models differ distinctly from traditional software applications: their behavior is learned, not explicitly programmed. Furthermore, the behavior of an AI system is learned not only from its software code but also from its training data. These factors make testing more

probabilistic and less deterministic. An AI system that performs well in ordinary situations may still produce harmful or unfair outputs in some scenarios.

These uncertainties should be identified and assessed during the training and testing phases of AI system development as a part of an overall AI system development process. Ideally, these training and testing activities specifically call out risk identification. There are technical, ethical, and organizational risks that surface during these stages, and all should be addressed proactively.

Managing risk during training and testing should be viewed as a lifestyle, even a part of the organizational culture, rather than just a one-time activity. Testing is less about performance and more about responsibility. The risks encompass technical, legal, and ethical aspects. By adopting rigorous, structured, and collaborative approaches to AI testing, organizations can build systems that are not only smart but also safe, fair, and trustworthy.

The Unique Nature of AI Testing

As stated earlier in this chapter, AI testing differs from traditional software testing in both objective and approach. AI testing focuses on behavior across distributions rather than deterministic correctness, unlike conventional software applications.

While conventional application software testing verifies that specific inputs yield expected outputs, AI system testing evaluates performance distributions, such as precision, recall, or fairness across subgroups.

For many AI applications, the "correct" answer (that is, the expected output) may be subjective or unavailable. For example, in content moderation or risk prediction, there may not be a single definitive label. In other words, there is no guarantee of *ground truth* in some AI systems.

Training itself is a dynamic process; organizations must carefully monitor training data and methods for convergence issues, instability, overfitting, and other risks.

Why "It Passed Testing" Isn't Enough

An AI model may perform well on a validation set but poorly in the real world, due to dataset shifts, new edge cases, and emerging adversarial techniques. AI innovation translates into a growing attack surface. AI system testing must extend beyond accuracy scores to consider emerging and broadening risk dimensions.

Types of Testing Risks in AI

AI system testing itself has risks that span multiple dimensions: technical, ethical, and operational. This section examines the key types of testing risks and explore how they arise.

These data-related risks are associated with the data used to train an AI system:

- **Biased training data:** Uneven representation leads to unfair or inaccurate outcomes.
- **Label noise:** Inconsistent or incorrect labels reduce model reliability.
- **Outliers and anomalies:** Extreme values or rare cases can distort training or lead to unstable model behavior.
- **Imbalanced datasets:** One class or group is underrepresented, leading to skewed predictions.
- **Concept drift:** Training data becomes outdated or misaligned with deployment conditions.

For example, an AI system trained on pre-pandemic medical data failed to identify common COVID-19 symptoms, due to temporal concept drift.

Several model-related risks are associated with the selected AI model:

- **Overfitting:** The model learns the noise in the training set rather than the general patterns.
- **Underfitting:** The model is too simple to capture real-world complexities in the context of the AI model.
- **Lack of robustness:** The AI model performs poorly on edge cases or on minor input variations.
- **Poor calibration:** A model's confidence scores don't match actual probabilities (for example, a model is "90% confident" but correct only 60% of the time).

Evaluation risks are associated with the evaluation of an AI system:

- **Inappropriate metrics:** When the wrong metrics are selected for evaluating an AI system, it can be challenging to interpret the testing results accurately.
- **Test set leakage:** Overlap between training and test data leads to inflated performance.
- **Insufficient coverage:** Test data may not include critical subgroups or use cases likely to occur in production.

Deployment and integration risks are associated with differences between test and production environments:

- **Environment mismatch:** Training assumptions don't hold up in production due to differences in data, operating systems, and software.
- **Shadow AI use:** Unvetted models are deployed without formal testing, documentation, or approval.
- **Feedback loops:** The AI model's predictions influence future data collection, reinforcing biases.
- **Security and access differences:** Production environments often introduce new attack surfaces (for example, exposed APIs) that weren't present in testing.

Testing Risk Identification Techniques

Identifying AI testing risks requires quantitative analysis and structured review processes. These risks include:

- **Exploratory Data Analysis (EDA):** EDA uncovers data distribution issues, imbalances, missing values, and potential outliers. It's often the first step in identifying risks before AI model training begins.
- **Fairness and bias audit:** These audits examine performance across various subgroups (for example, gender, age, or race). Differential performance may reveal systemic bias that wasn't obvious before.
- **Error analysis:** Manual reviews of misclassified output can reveal systematic patterns of failure. This approach is beneficial for high-impact applications, such as healthcare or criminal justice.

For example, a financial services firm conducts an EDA and finds that 80% of loan data comes from urban postal codes, while rural areas are underrepresented. This insight prompts them to collect more balanced data before training their AI model.

Risk Mitigation Strategies

Once risks have been identified, organizations must apply targeted and intentional interventions to mitigate them. Some common mitigation types are discussed here.

Three techniques for improving data quality and diversity are as follows:

- **Data augmentation:** Generate additional training data to fill gaps.
- **Resampling:** Balance class distributions using *upsampling* or *downsampling*.
- **Synthetic data:** Use carefully constructed artificial data to improve representativeness. When using synthetic data, ensure that it cannot be reidentified as real data subjects.

For example, an organization's recruiting platform augments its dataset with profiles of underrepresented job seekers, to reduce gender bias in its AI-powered recommendation engine.

In situations where an AI model is overfitting, techniques such as *L1 regularization* (adding a penalty based on the absolute value of model parameters, thereby simplifying the model) or *L2 regularization* (adding a penalty based on the square of model parameters, thereby promoting stability) can be employed. Both techniques discourage a model from fitting noise or irrelevant details in training data. *Dropout* is a similar technique used in neural networks.

Using ensemble models (discussed in Chapter 3) or performing *model stacking* can mitigate the weaknesses of any single approach and improve overall performance and resilience.

Accuracy Isn't Always Enough

In a medical diagnostic system, 95% accuracy may sound impressive, but if disease prevalence is only 5%, this could mean the model is classifying almost every case as "healthy." Precision and recall matter in healthcare.

Managing Risk Throughout the Training Lifecycle

Testing doesn't end when the AI system is deployed. Instead, testing occurs throughout the AI model training lifecycle, and even after training is complete. Even as AI systems can be explainable, they are not always predictable. Hence, whenever there is any change in operating parameters or changes in training data, AI systems should be regression tested to see whether their behavior still falls within acceptable bounds. The types of testing that should occur are outlined here.

Early Training Phase

- Monitor convergence behavior
- Detect vanishing gradients or poor learning rates
- Validate model architecture decisions

Mid-training Monitoring

- Validate loss curves
- Track performance metrics on validation sets
- Watch for instability or sudden metric spikes

Late-stage Tuning

- Perform final hyperparameter optimization
- Use large-scale validation sets
- Conduct scenario-based or adversarial tests

Post-training Testing

- Use real-world simulations where possible
- Evaluate explainability and counterfactual behavior
- Audit logs and documentation for transparency

In an example shipping and logistics company, during post-training testing, an AI system for supply chain optimization begins, where all goods are processed through a single hub. A scenario test reveals a learned cost shortcut that ignores system fragility, which was identified just in time.

Testing Tools and Infrastructure

Organizations benefit from specialized tools to support testing and risk tracking for their AI systems.

For MLOps integrations, tools such as MLflow, SageMaker, and Vertex AI offer pipelines for managing training, testing, and deployment with auditability. Testing platforms, including Great Expectations (GX), Checklist (at checklist.gg), and TensorFlow Model Analysis (TFMA), facilitate data validation, testing across task dimensions, and performance evaluation, respectively.

Production monitoring tools, such as Arize and Fiddler, allow post-deployment testing by comparing real-world results to expected outcomes.

Post-deployment monitoring is an extension of testing in AI systems. Post-deployment testing identifies new risks and feeds them back into the model improvement cycle. Organizations that do not perform post-deployment monitoring are at risk of model drift and other issues.

Common Testing Pitfalls

Even experienced and mature AI development teams can encounter issues during testing and risk management. Examples include:

- **Overreliance on test accuracy:** Accuracy alone can mask numerous problems, including false positives, false negatives, and the unfair treatment of subgroups.
- **Lack of real-world scenarios:** While synthetic tests are helpful, real-world context is essential. Edge cases and environmental variability must be included in testing.
- **Siloed responsibility:** When data scientists test their own models without outside review, bias and blind spots go unchecked. The practice of *peer review* as an essential software development technique applies to AI systems, and for the same reason: developers sometimes don't recognize their own errors.
- **Metric misalignment with business goals:** A model may optimize for technical metrics (for example, F1 score) but fail to deliver real business value, such as reducing customer churn or improving efficiency.
- **Neglecting explainability:** High performance is meaningless if stakeholders cannot understand or trust the model's behavior. Explainability testing (for example, with SHAP or LIME) should be included.

Case Studies and Examples

This section discusses some additional examples of AI system testing risk.

Hiring Algorithm Bias

A tech firm trains a hiring model that favors resumes using masculine-coded language. Post hoc analysis reveals that word embeddings used during training introduced subtle bias. The company retrains with debiased embeddings and updates its evaluation metrics.

Autonomous Vehicle Simulation

An AV company discovers that its vehicles perform poorly in tunnels, a scenario that was not included in the test set. They built a synthetic tunnel-driving simulator to address the gap and improve edge case handling.

Financial Services Outlier

A fraud detection system begins triggering false alarms on large but legitimate transactions. Further testing reveals that high-volume data was underrepresented in training. The model is updated with stratified sampling to resolve the issue.

Summary

Developing responsible, effective, and trustworthy AI systems begins with managing the data on which they are trained. Data governance plays a foundational role by ensuring that data usage is compliant, lawful, accurate, sufficient, and fit for purpose. Without governance structures in place, organizations expose themselves to significant risks, ranging from biased models and compliance failures to eroded stakeholder trust and regulator sanctions. While governance is not always mandated by law, it is widely regarded as a prerequisite for the scalable and high-quality deployment of AI. Sound governance practices encompass lawful data use, integrity controls, and the assignment of roles such as data stewards and owners to ensure accountability throughout the AI lifecycle.

Documentation of data lineage and provenance is critical. Lineage tracks the flow of data through systems, while provenance focuses on its origin and collection context. These disciplines support traceability, explainability, and reproducibility in AI development, enabling organizations to understand how data has been transformed, utilized, or repurposed, and whether those uses remain appropriate. Often lacking data governance capabilities, many organizations struggle to implement traceability frameworks, especially when working with third-party or historical data. Without clear lineage and provenance, organizations may struggle to defend their model outputs, meet regulatory requirements, or pinpoint the root cause of performance issues.

When data governance and traceability are in place, another challenge is rigorous testing of the AI system itself. Testing is not a single event but a structured process that spans unit testing, integration testing, validation, and performance evaluation throughout the AI system's lifespan. Unlike traditional software, AI systems require statistical and behavioral testing under a range of conditions. Developers must confirm that the AI model performs well on standard benchmarks and also that it is secure, interpretable, and fair. Testing infrastructure must include appropriate metrics, real-world simulations, and tools for continuous evaluation as models and training data evolve.

Closely linked to AI model testing is the identification and management of risks that surface during training and testing. Risks may stem from data quality issues, biased labels,

inappropriate evaluation metrics, or the failure to test across demographic subgroups. Effective risk management requires a culture of transparency, structured review gates, and the use of MLOps tools to detect and remediate issues before models are deployed to production. Risk identification should not be a passive process; instead, personnel should examine and review test data and results to proactively identify risks. While some risks can be mitigated through technical means, others demand organizational awareness and cross-functional collaboration. Some may require model redesign.

Exam Essentials

Data governance is a foundational enabler for the development of trustworthy AI. While not always a strict legal requirement, data governance ensures that the data used in AI systems is lawful, reliable, and appropriately managed. Poor or no governance can lead to biased outcomes, regulatory exposure, and a loss of stakeholder trust. Effective governance includes lawful data use, role assignment, data quality monitoring, and alignment with business purpose.

Data provenance and lineage must be documented to support accountability, traceability, and reproducibility. Provenance explains where the data originated and under what conditions it was collected, while lineage tracks how it is transmitted and modified across systems. Both are essential for auditability and understanding model behavior, particularly when facing regulatory reviews or real-world performance issues.

AI testing is broader than traditional software testing and must address fairness, robustness, and interpretability. AI system testing encompasses unit, integration, validation, and performance testing, and must also include ethical dimensions, such as bias, explainability, and security. Standard metrics are insufficient; testing must also account for edge cases, subgroup disparities, and evolving risks.

Training and testing introduce unique risks that must be proactively identified and mitigated. AI systems can fail silently if risks during training and testing are ignored. These include label noise, data drift, test leakage, and the inappropriate selection of metrics. Organizations should apply structured review processes, approval gates, validation tools, and continuous monitoring to stay ahead of potential failures.

A risk-aware testing culture requires organizational maturity and interdisciplinary collaboration. AI testing and risk management are not solely technical activities. They require buy-in from governance teams, legal counsel, domain experts, and business leaders. Formal ownership, clear review gates, and robust documentation help ensure that AI systems are effective and accountable.

Review Questions

The following questions are designed to test your understanding of this chapter's material. For more information on how to obtain additional questions, see this book's Introduction. You can find the answers in Appendix A.

1. What is the primary purpose of data governance in AI development?
   - **A.** To improve marketing performance
   - **B.** To reduce software development costs
   - **C.** To ensure lawful, high-quality, and appropriate data use
   - **D.** To eliminate the need for human review
2. Which of the following best describes data provenance?
   - **A.** The pipeline that transforms and stores data
   - **B.** The real-time path data takes through memory
   - **C.** The statistical summary of model inputs
   - **D.** The origin and collection context of a dataset
3. In AI testing, which metric is most appropriate for an imbalanced classification problem?
   - **A.** Accuracy
   - **B.** Recall
   - **C.** Precision
   - **D.** F1 score
4. What is a common risk when model testing lacks subgroup validation?
   - **A.** Decreased model size
   - **B.** Fairness or bias issues in specific populations
   - **C.** Overuse of memory resources
   - **D.** Reduced training time
5. What is the main distinction between data lineage and data provenance?
   - **A.** Lineage tracks flow and transformations; provenance identifies origin and context
   - **B.** Lineage refers to external sources; provenance refers to internal flows
   - **C.** Lineage is legal; provenance is technical
   - **D.** There is no practical difference

6. Which of the following is an example of an adversarial risk in AI systems?
   - **A.** Long response times under peak load
   - **B.** Poor recall on minority class samples
   - **C.** Inputs crafted to cause model misclassification
   - **D.** Model convergence failure during training

7. What is the goal of counterfactual fairness testing?
   - **A.** To test model performance on future datasets
   - **B.** To simulate random inputs
   - **C.** To evaluate if changing a protected attribute affects model output
   - **D.** To measure runtime performance under varying loads

8. Why is integration testing important for AI systems?
   - **A.** It reduces GPU costs
   - **B.** It ensures that model predictions are legally compliant
   - **C.** It confirms that different system components work together correctly
   - **D.** It guarantees interpretability

9. Which activity is most critical during the training phase for identifying model learning problems?
   - **A.** Reviewing deployment logs
   - **B.** Inspecting feature importance
   - **C.** Monitoring convergence and loss curves
   - **D.** Testing production APIs

10. What is the benefit of using model cards or fact sheets in AI testing and deployment?
    - **A.** They eliminate the need for validation testing
    - **B.** They automate model tuning
    - **C.** They increase the learning rate during training
    - **D.** They provide structured transparency about model behavior and limitations



Chapter

10

Deployment and Monitoring

THIS CHAPTER COVERS PORTIONS OF AIGP DOMAIN III, "UNDERSTANDING HOW TO GOVERN AI DEVELOPMENT," SUBDOMAIN III.C, "GOVERN THE RELEASE, MONITORING, AND MAINTENANCE OF THE AI MODEL," AS WELL AS PORTIONS OF AIGP DOMAIN IV, "UNDERSTANDING HOW TO GOVERN AI DEPLOYMENT AND USE," SUBDOMAIN IV.A, "EVALUATE KEY FACTORS AND RISKS RELEVANT TO THE DECISION TO DEPLOY THE AI MODEL," AND SUBDOMAIN IV.C, "GOVERN THE DEPLOYMENT AND USE OF THE AI MODEL." THE FOLLOWING TOPICS ARE COVERED:

- ✔ Preparing for deploying an AI system to production
- ✔ Deploying options and considerations
- ✔ Implementing continuous monitoring

THE AI GOVERNANCE PROFESSIONAL (AIGP) PERFORMANCE INDICATORS IN THIS SUBDOMAIN INCLUDE:

- ✔ Assess readiness and prepare for release into production (for example, creating the model card and satisfying conformity requirements).
- ✔ Understand the differences in AI deployment options (for example, cloud vs. on-premise vs. edge, and using the AI model as-is or with fine-tuning, retrieval augmented generation, or other techniques to improve performance and fit).



- ✔ **Apply the policies, procedures, best practices, and ethical considerations to the deployment of an AI model (for example, data governance, risk management, issue management, user training).**
- ✔ **Conduct continuous monitoring of the AI model and establish a regular schedule for maintenance, updates, and retraining.**



Deploying an AI system into production is a significant milestone in the lifecycle of an AI model. This chapter explores the governance, technical, and operational requirements necessary to ensure that AI systems are deployed safely, ethically, and effectively. The chapter begins with pre-release readiness, including documentation, model validation, and environment selection. Next, the chapter covers the policies, risk controls, and user enablement strategies required during deployment. Finally, the chapter emphasizes the importance of continuous monitoring practices to ensure performance, fairness, and compliance over time. Topics discussed include managing drift, scheduling retraining, assigning oversight responsibilities, and aligning monitoring with governance frameworks. These practices ensure that AI systems function as intended and serve their intended purpose in dynamic and often unpredictable environments.

About the Coverage of AIGP Domains in This Chapter

This chapter covers elements from the AIGP body of knowledge from Parts III and IV, to better align the subject matter and avoid repetition.

Refer to the AIGP domain chart in this book's Introduction to better understand how the AIGP domains are covered by the chapters in this book.

Pre-release Readiness

Preparing an AI system for production deployment requires structured documentation, regulatory conformity, performance benchmarking, architectural decisions, and operational planning. Here, organizations transition from experimentation to execution, ensuring the model is production-worthy, well-governed, and responsibly and ethically designed. This section explores readiness criteria, deployment models, and optimization strategies that ensure a successful release.



It is assumed that organizations have mature IT service management practices, including *release management*. This chapter focuses on AI-specific considerations, which can be considered an overlay to standard release management practices.

Assessing AI System Readiness for Production

Before an AI system can be promoted into production, organizations should undertake a readiness evaluation to ensure its safety, accountability, and suitability for its intended context. This evaluation begins with robust documentation, including the creation of a model card.

Initially discussed in Chapter 2, model cards provide structured documentation about the AI model's intended use, training data, evaluation metrics, known limitations, and ethical considerations. Creating a model card encourages the development team to think critically about the model's assumptions, limitations, and performance in real-world settings. Model cards are widely recommended in AI governance practice, and they align with principles in frameworks such as NIST AI RMF and ISO/IEC 42001.

Many jurisdictions impose legal obligations on AI systems before they are released to production. While these obligations vary, they may include:

- **Conformity assessments:** Required under the EU AI Act for high-risk systems, including documented quality controls, risk assessments, and human oversight provisions. Requirements will vary in other jurisdictions, although these are sound practices.
- **Third-party audits or internal validation:** Internal governance boards may require internal or external validation before approving production use.
- **Data provenance and licensing compliance:** Ensuring all training and fine-tuning datasets are lawfully sourced, appropriately licensed, and, where personal data is involved, processed in compliance with relevant privacy regulations (for example, the GDPR).
- **Insurance-imposed requirements:** A *cyber insurance* policy may require covered organizations to implement various application-related controls and safeguards, such as change management, code reviews, gate reviews, and periodic risk assessments.

AI governance professionals must be familiar with regional and sector-specific requirements, especially when deploying systems that interact with the public, make consequential decisions, or use personal data. Legal agreements may also include some or all of these same requirements.

Operational Readiness and Risk Evaluation

Before deployment, models should undergo stress testing under expected and edge case workloads. This testing includes:

- Latency and throughput benchmarks (especially for real-time use cases)
- Resource (memory, compute) usage under load
- Failover behavior during outages or degraded performance
- Evaluation against adversarial and malicious prompts

These pre-deployment tests ensure that the AI system will behave predictably and reliably in production, under both ideal and abnormal or hostile conditions.

In addition to technical readiness, AI governance should require a formal review of the system's potential harms and benefits. This may take the form of:

- **AI impact assessments (AIIAs):** These evaluate the social, ethical, and operational impacts of the model. Chapter 8 covers them in depth.
- **Risk assessments:** Teams should articulate potential harm events (for example, discriminatory outcomes or misinformation propagation) and define corresponding mitigation strategies. This is covered in Chapters 7 and 8.

AI Model Deployment Considerations

There are several ways to implement an AI system. Deployment model decisions influence performance, latency, control, and regulatory posture. Each option has tradeoffs, as shown in Table 10.1.

TABLE 10.1 Comparison of AI System Deployment Options

Deployment Mode	Pros	Cons
Cloud	Scalable, cost-effective, easy to integrate with APIs	Less control over data locality, possible regulatory concerns
On-premises	High control, good for sensitive data, meets strict compliance	Costly infrastructure, lower scalability
Edge	Ultra-low latency, no dependency on network connection, user privacy	Limited compute power, model size constraints
Hybrid	Balances performance and compliance by combining edge/on-prem inference with cloud training or updates	Added architectural complexity, higher integration overhead

The selected deployment architecture must align with business needs, user expectations, compliance requirements, and technical constraints. Pragmatically, the deployment model should have been decided in the design phase of the AI system project.

Production-ready AI may require *containerization* (for example, using Kubernetes or Docker) to enable consistency across environments. Containerization tools enable auto-scaling and deployment orchestration in multi-cloud or hybrid environments.

For example, a smart home assistant runs inference on the edge to improve responsiveness and protect user privacy, but its language model is updated periodically via a cloud sync process.

AI Model Optimization for Production Fit

Organizations need to decide whether to deploy the AI model in its original state, fine-tune it for specific tasks, or augment it with retrieval mechanisms. Each path has distinct implications, as shown in Table 10.2.

Fine-tuning enhances performance in specific domains but may also increase maintenance costs and retraining requirements. RAG strategies reduce the risk of hallucinations and improve factual grounding, especially in enterprise settings. These options could be considered in the design phase.

When deploying on edge or resource-constrained devices, models must often be *quantized* (for example, reducing mathematical precision from 32-bit floating point to 8-bit integers) or *distilled* (training a smaller "student" model to approximate the behavior of a larger "teacher" model) into smaller models. These techniques may reduce model size and inference latency but sometimes at the cost of modest reductions in accuracy.

For example, a vision model deployed on a drone utilizes quantization to reduce power consumption, albeit at the slight expense of object detection confidence.

TABLE 10.2 AI Model Adaptation Options

Approach	Description	Use Case Example
As is	Use base model without modification	General-purpose chatbot or summarization tool
Fine-tuned	Adapt model to domain-specific data	Legal document classifier for contract terms
Retrieval augmented generation (RAG)	Use external knowledge base during inference	Search assistant that cites internal policy documents

In another example, a financial services firm uses a RAG-based chatbot to assist employees with regulatory compliance questions. The AI model queries internal policy documents in real time and generates grounded responses, reducing legal exposure from hallucinated or outdated answers.

Human-in-the-loop and Release Protocols

Some AI systems, particularly those in high-risk classifications, must include human review or other types of human-centric override capabilities. This may involve:

- **Human-in-the-loop (HITL):** A human must approve or confirm some or all model decisions (for example, for loan applications or medical diagnoses).
- **Human-on-the-loop (HOTL):** A human monitors outputs and can intervene if needed (for example, in content moderation systems).
- **Human-in-command:** Strategic decisions, such as incident response, system deactivation, or escalation protocols, remain in human hands.

The AI governance team must define and test these controls before launch, to ensure they are practical, scalable, and effective. Key parties should be aware of their responsibilities and receive proper training.

Pre-release planning requires not only a release procedure but also a verification procedure that confirms the successful implementation of the AI model. Also, a *rollback* procedure must be developed, in case the release is unsuccessful. Essential practices include:

- Release procedure, verification procedure, and rollback procedure submitted through *change management* for review and approval
- Staged rollouts with feature flags
- Canary deployments for performance testing
- Monitoring for KPIs and leading risk indicators
- Clear communication channels for status reporting and issue escalation

For example, an insurance company releases its underwriting model to 5% of users first, monitoring for policy approval rate discrepancies before expanding the release to more users.

Final Gate Reviews and Stakeholder Signoff

Before an AI system is cleared for release to production, it must pass through one or more final gate reviews, typically conducted by a multidisciplinary governance team. These reviews ensure that:

- All documentation (including model cards, design artifacts, audit logs, licensing records, and test reports) is complete
- Risk mitigations are implemented, or deferrals are approved

- Security hardening and incident response procedures are validated
- Operational readiness and monitoring plans are in place
- Regulatory and policy compliance is verified
- Version control and traceability of models, data, and configurations are maintained
- All users and operations personnel are trained

This step culminates in stakeholder signoff, where AI governance, legal, engineering, and business leaders formally approve the system's release.

Example roles for a final gate review include:

- **Legal:** Data use, licensing, and contracts
- **Engineering:** Technical validation and scalability
- **Business:** Value alignment and readiness for customers
- **Cybersecurity:** Vulnerabilities and risks addressed
- **Privacy:** Data use review
- **Governance:** Fairness, accountability, and transparency

Deployment Considerations

Deployment is not only a technical milestone but also a governance transition. Deployment shifts responsibility from the development team to operational stakeholders, which calls for a broader framework that includes risk management, ethical accountability, user enablement, and continuous monitoring and oversight. This section examines the policies, procedures, and essential practices that govern AI deployment, to ensure the AI system operates responsibly, securely, and in alignment with its intended purpose.

Enforcing Deployment Policies and Procedures

Well-defined deployment protocols bridge the gap between development and operations. These protocols ensure that deployment is orderly, repeatable, and governed. Key components of an effective deployment policy include:

- **Approval gates:** Deployment should be contingent on prior signoffs (for example, from compliance, legal, and security).
- **Strict release procedures:** AI system developers should not have administrative access to production environments, even for deployment. Instead, developers should package the AI system and hand it over to operations personnel, who then install the AI system's components according to the provided installation procedures.

- **Environment controls:** Policies should mandate isolation between development, staging, and production environments, preventing unvetted code from reaching end users, as well as keeping developers out of the production environment.
- **Rollback procedures:** Each deployment plan should include defined verification procedures, rollback steps, escalation contacts, and fail-safe procedures.

An example financial institution uses an internal "go/no-go" checklist to ensure that every AI model entering production has passed bias testing, cybersecurity vetting, and executive signoff.

Deployment must be traceable, repeatable, and reversible. Development tools like Git, CI/CD pipelines, and automated logging help ensure that every change is logged, tracked by a *version control* system, auditable, and reversible. *Model registries* (which resemble *configuration management databases*) track versions, training data lineage, and metadata, supporting both reproducibility and governance.

What's in a Model Registry?

- Model version and hash
- Training dataset identifiers
- Evaluation results
- Deployment status (for example, staging, live)
- Ownership and contact info

Transparent deployment practices support troubleshooting, incident response, legal defensibility, and auditability.

Data Governance in the Deployment Phase

AI models are often deployed alongside application pipelines or data feeds that feed them data in real time. Organizations must apply data governance policies to the input data (used at inference time) and any data generated during AI operations. Key data governance activities include:

- **Data minimization:** Limit data collection to only what is necessary for the task.
- **Purpose limitation:** Use data only for formally defined business purposes.
- **Retention policies:** Define the duration for which various datasets are stored, particularly user or behavioral data collected during AI system processing.
- **Access controls:** Protect real-time data inputs with fine-grained access permissions.
- **Data quality monitoring:** Validate inputs and detect anomalies to prevent cascading errors in production.

Deployment frequently triggers legal obligations, including cross-border data flow stipulations (systems deployed globally are subject to laws like the GDPR or LGPD, which regulate international data transfers), as well as consent and notice requirements.

Deployment and AI governance teams should work closely with legal counsel to confirm that data-handling practices remain compliant when AI systems go live.

Data governance is discussed in depth in Chapter 9.

Risk Management During Deployment

AI system deployment introduces new risks that extend beyond AI model performance, which include:

- **Infrastructure risks:** Network failures, cloud outages, and resource exhaustion can impair production systems.
- **Security risks:** AI systems may become targets for attacks, making them vulnerable to prompt injection, model inversion, data exfiltration, and *denial-of-service (DoS)* attacks.
- **Compliance and legal risks:** Violations of privacy, intellectual property, or sector-specific regulations.
- **Reputational risks:** Flawed outputs and ethical failures can damage user trust and brand credibility.

As is done earlier in the development lifecycle, AI governance and cybersecurity teams must inventory these risks and process them using risk treatment to determine appropriate risk response (one of: accept, mitigate, transfer, or avoid). Mitigation may involve business continuity plans, red teaming, and proactive testing.

For example, a healthcare AI tool undergoes adversarial testing before release to ensure it resists inputs that could lead to unsafe treatment suggestions.

Mitigating deployment risk requires that organizations monitor AI behavior after release. These monitoring techniques include:

- **Drift detection:** Alerting when inputs or predictions deviate significantly from training data or historical trends.
- **Bias detection:** Surfacing inequities that emerge as the system interacts with real-world populations.
- **Misuse detection:** Monitoring for unauthorized or unintended uses of the AI system.
- **Explainability monitoring:** Ensuring generated explanations remain accurate, consistent, and understandable to end users.

Arguably, other monitoring techniques, such as cybersecurity and IT infrastructure monitoring, should also be employed. These techniques include vulnerability scanning, penetration testing, audit log ingestion and alerting, and threat intelligence analysis.

Organizations that establish dashboards, alert thresholds, and escalation workflows ensure that AI governance continues after launch.

Drift in Deployment

An AI-powered chatbot trained on pre-pandemic language patterns begins misinterpreting user queries in a post-COVID context. Drift detection monitoring reveals a sharp rise in unrecognized intents, prompting the retraining of the AI model with recent data.

Issue Management and Incident Response

Organizations implementing IT infrastructure, business applications, and AI systems should implement a structured issue management process that aligns with existing IT service management (ITSM) frameworks. AI-specific nuances are essential for monitoring and responding to incidents in their AI systems. Core elements of AI-specific incident response include:

- **Triage categories:** Classify issues by severity (for example, service interruption, ethical violation, user harm, as well as lower-severity issues).
- **Ownership and escalation paths:** Define which team is responsible for each class of issues, such as data, model, UX, and so forth, and outline the process for escalating problems.
- **Tracking tools:** Utilize systems such as JIRA, ServiceNow, or custom dashboards to track model-related tickets.

AI-specific incidents, including hallucinations, unexpected outputs, or bias events, require appropriate, tailored responses. AI governance should be embedded in the broader incident response plan and include:

- Predefined incident categories (for example, ethical, legal, and security)
- Pre-written response procedures and playbooks
- Prompt notification to cyber insurance carrier for covered events
- Stakeholder notification and escalation trees
- Postmortem procedures that include updates to training, documentation, or policy

AI incident response should be integrated into an organization's overall cybersecurity and privacy incident response frameworks, ensuring a consistent and understandable overall incident response approach.

Incident response is discussed fully in Chapter 11.

Anatomy of an AI Incident

Here are the steps performed in a typical AI incident scenario:

Trigger: User/customer submits a complaint about unfair treatment.

Analysis: Team reviews model behavior and decision path.

Impact: System found to misclassify based on dialect.

Communication: Appropriate level of management and leadership informed and updated periodically.

Action: Hotfix deployed; long-term retraining scheduled; user/customer notified; incident response record keeping performed.

Governance: Model card updated, incident logged for audit, post-incident AAR performed.

User Training and Enablement

For many AI systems, users are typically employees, such as analysts, customer service agents, and claims processors, who interact with the model within some tool-assisted workflow. Successful outcomes with new AI systems require:

- **Role-based training:** Tailored instructions for different groups of users (for example, frontline agents vs. supervisors).
- **Administrative training:** IT and others whose responsibilities include managing the AI system and its underlying infrastructure.
- **Confidence thresholding:** Teaching users when to trust the system and when to intervene.
- **Documentation and job aids:** Creating cheat sheets, workflows, and escalation guides can help users use the AI system early on in their learning curves.
- **Ongoing enablement:** Refreshing training and updating job aids as models evolve and workflows change.

For example, a shipping and logistics company deploys a delivery routing optimization model. Dispatchers are trained to review AI suggestions and apply overrides in cases of local disruptions (for example, road closures or driver illness).

For customer-facing AI systems, deployment should include transparency and guidance for external users. This might involve:

- Disclosure of the presence of AI capabilities
- Plain-language descriptions of how the AI works
- Notice-and-consent mechanisms, including links to a privacy policy or notice of privacy practices (NOPP)
- Tutorials, FAQs, and customer support integration

Cross-functional Governance Alignment

AI governance is not an island but depends on cross-functional collaboration with other governance bodies. Typical supporting participants are shown in Table 10.3.

Coordination mechanisms used among governance bodies may include standing responsible AI councils, deployment war rooms, governance cross-membership, and shared dashboards.

TABLE 10.3 Cross-functional Governance for AI System Deployment

Role	Responsibility
Data science	Model deployment, monitoring metrics
IT/DevOps	Infrastructure, CI/CD, scalability
Legal	Compliance with laws and contracts
Security	Threat detection and incident response
Governance	Alignment with organizational values and policies
Business units	Value realization, user adoption

The governance and management of AI systems should not be isolated; instead, they should be integrated into overarching governance, risk, and compliance (GRC) frameworks. This means that AI risks should be included in the enterprise risk register, AI controls should be included in audits, and deployment review gates should be aligned with the product governance lifecycle workflows.

For example, a bank's AI-powered credit scoring system is reviewed as part of quarterly SOX compliance audits. It must meet model governance criteria set by the board-level risk committee.

Continuous Monitoring

Deployment is not the end of the AI system lifecycle but the beginning of a new operational phase that introduces new activities. Continuous monitoring ensures that the AI system remains accurate, fair, performant, and secure. Monitoring also confirms alignment with the AI system's intended purpose, even as it encounters new data, changing environments, and evolving expectations. This section discusses monitoring strategies, governance frameworks, retraining schedules, and maintenance plans necessary to ensure the sustained effectiveness and trust of AI.

Foundations of AI System Monitoring

AI systems are probabilistic and data-dependent, prone to degradation over time as real-world data changes. Without continuous monitoring, AI model predictions may become stale, biased, or outright harmful. Ongoing monitoring ensures that:

- Performance remains within defined thresholds
- Drift is detected early
- New risks and edge cases are identified and dealt with
- Emerging compliance obligations are met

The monitoring of AI systems is both a technical necessity and a governance imperative. Common monitoring areas include:

- **Prediction quality:** Accuracy, precision, recall, and confidence
- **Data inputs:** Volume, format, anomalies, missing values, and malicious inputs such as *prompt injection* and model poisoning attacks
- **Latency and performance:** Response times under varying load
- **Resource consumption:** CPU/GPU, memory, storage, and network consumption
- **Fairness:** Disparity across subgroups over time
- **Drift:** Shifts in data distribution or model output
- **Usage metrics:** Adoption rates, abandonment, and user behavior

For example, an organization's AI-powered hiring assistant monitors and tracks rejection rates by gender and race to detect and mitigate disparate impact over time. Anomalies are reported to management.

Refer to Chapter 6 for information regarding the stated responsibilities of developers, deployers, and operators, as outlined in the EU AI Act.

Technical Monitoring Frameworks

Modern monitoring infrastructure and tooling typically includes:

- **Transaction logging:** Capture inputs, outputs, metadata, and errors
- **Dashboards:** Provide real-time and historical KPIs and visualizations
- **Alerts:** Triggered when KPIs exceed predefined thresholds
- **Audit logs:** Record both system and model lifecycle events for forensic analysis, incident alerting, and compliance audits

Standard monitoring tools include Prometheus (metrics), Grafana (dashboards), ELK stack (logs), and AI/ML-specific tools such as Evidently, WhyLabs, or AWS SageMaker Model Monitor.

Observability of an AI system means an organization can understand *why* an AI model produces a given output or fails under certain conditions. Observability aims to provide traceability of the prediction path, logging of model versions and configuration, and correlation of outputs with data inputs. Observability is particularly important for regulated industries and for systems that make high-impact decisions.

Detecting Drift and Performance Degradation

As discussed throughout this book, model drift occurs when the data a model sees in production differs significantly from the data it was trained on. The types of model drift include:

- **Covariate drift:** Input distribution changes (for example, new customer behaviors)

- **Label drift:** Changes in the distribution of outcomes (for example, loan defaults rise)
- **Concept drift:** The relationship between inputs and outputs changes (for example, fraud tactics evolve)

Drift detection techniques include:

- **Statistical tests:** Kolmogorov-Smirnov (K-S) and Population Stability Index (PSI)
- **Prediction monitoring:** Sharp drops in confidence or consistency
- **Shadow models:** Parallel models trained on recent data for comparison to models in production

Detection tools should support both batch and streaming environments.

For example, a credit card company's financial fraud detection model becomes less effective as fraudsters adapt to its decision patterns—a form of concept drift.

Fairness and Ethical Monitoring

After deployment, AI models can exhibit bias even if they pass fairness checks during development and at the time of release. Continuous fairness monitoring should include:

- Tracking prediction patterns across sensitive attributes and watching for trends that could indicate bias
- Watching for subgroup degradation over time
- AI model transparency
- Reviewing user complaints and appeal outcomes—both what they are complaining about and the number of complaints

Ethical monitoring of an AI system can be more difficult to automate, but monitoring and alerting can include:

- Surges in user opt-outs or complaints
- Increases in flagged content (for content generation systems)
- Escalation logs involving model overrides or rollbacks

AI governance should receive regular summaries and have the authority to declare incidents, which could result in suspending or retraining problematic systems.

For example, an AI model for email autocomplete is found to consistently suggest gendered language: "nurse" with "she" and "engineer" with "he." Ongoing audits reveal the pattern and prompt retraining with balanced data.



Organizations should monitor all aspects of any business process enabled by AI and not only the AI system itself. Issues and anomalies can occur anywhere, including outside information systems.

Defining Maintenance Schedules

While monitoring AI systems is essential, alone, it is not sufficient. Periodic maintenance is also needed for AI systems, including:

- **Regular evaluation:** Quarterly or monthly model quality assessments
- **Scheduled retraining:** Rebuilding models with updated datasets; frequency depends on model use cases and volatility of the operating environment and may range from weekly to annually
- **Hotfix intervals:** Immediate patches for urgent failures
- **Business reviews:** Quarterly business reviews to study metrics, KPIs, and incidents
- **Dependency and compliance updates:** Updating libraries, frameworks, and container images and performing periodic compliance or regulatory reassessments

Retraining AI models is potentially expensive, requiring data curation, validation, testing, and redeployment. Organizations will need to balance:

- Accuracy improvements vs. infrastructure costs
- Need for freshness vs. risk of instability
- Legal obligations vs. business priorities

Tools like automated model pipelines (for example, MLFlow and TensorFlow Extended—TFX) can help reduce overhead.

For example, an insurance chatbot is retrained monthly with newly tagged intents and customer feedback logs. A staging environment tests each new model before deployment.

Handling Emerging Risks

Because there are so many dynamic factors within and surrounding AI systems, continuous monitoring should include leading indicators of systemic risks, including:

- Rapid increase in edge case failures
- Repeated human overrides of predictions
- Decline in user trust or engagement metrics
- Increase in output toxicity, misinformation, and hallucinations
- Rapid increase in abuse and misuse cases

Governance teams should maintain "tripwires" that trigger model reviews or suspensions.

Occurrences in any of these areas should be considered an incident, with appropriate levels of incident response as directed by incident response playbook procedures.



I am not calling for additional monitoring but for vigilance concerning the essential areas to monitor.

For example, a product recommendation engine experiences a spike in customer complaints about irrelevant or offensive suggestions. A monitoring dashboard flags the anomaly and freezes further model updates pending review.

Human Oversight in Monitoring

Because monitoring involves numerous disciplines, monitoring responsibilities should be assigned across roles, including:

- **AI/ML engineering:** Performance, drift, and accuracy
- **Data governance:** Fairness, data lineage, and usage compliance
- **Cybersecurity:** Threat intelligence, vulnerabilities, and attack attempts
- **Operations/IT:** Latency, uptime, and resource utilization
- **Ethics/risk:** Ethical violations, bias, and hallucinations
- **Internal audits:** Results and actions taken
- **Regulation developments:** New laws, changes in enforcement, and case law precedents
- **Maintenance:** Operations, logs, and retraining

Ownership maps should include escalation trees and support 24/7 coverage if needed for critical systems. Even with automation, humans must remain "on the loop" to validate anomalies, respond to alerts, and exercise judgment.

As stated earlier, monitoring will, at times, compel monitoring personnel to declare an incident, triggering incident response proceedings.

For example, an AI-powered content moderation flags a high number of false positives for satire. Human reviewers step in to recalibrate classification thresholds and improve labeling.

Monitoring Third Parties

Thus far, discussions of deployment and monitoring assume the AI model is operated by the organization that deployed it. However, because many organizations employ AI systems operated by third parties (often in SaaS environments), some aspects of monitoring may not be available to the organization. Instead, only the vendor will have visibility into certain aspects of the AI system, particularly its underlying infrastructure, which is often concealed in SaaS environments.

The reduction in visibility and control of the AI system and its environment can be partially offset in some ways, including:

- Specific terms and conditions in the legal agreement between the customer organization and the AI system provider
- Periodic audits of the vendor's operating environment, including development, deployment, monitoring, incident response, and cybersecurity operations

It is assumed that organizations have third-party risk management programs in place, and such programs are not described here in detail. For a comprehensive review of third-party risk management practices, refer to any of the author's published books on the CISSP, CISA, CISM, or CRISC certifications.

Summary

Deployment of an AI system begins with pre-release readiness, which includes validating that the model is well-documented, compliant, and operationally fit. Development teams update model cards, conduct conformity assessments, and finalize documentation that describes use cases, limitations, data flows, and monitoring. Production readiness also requires architectural decisions regarding deployment environments, whether cloud-based, on-premises, or at the edge. The organization then considers model adaptation strategies such as fine-tuning or retrieval augmented generation (RAG). The release process also readies human oversight roles and rollback mechanisms in case unexpected outcomes emerge post-launch.

Once the system is formally approved for production, deployment practices must align with organizational policies, IT service management procedures, risk management frameworks, and ethical standards. This phase is subject to deployment protocols, approval gates, rollback plans, and version control procedures. Data governance remains critical, particularly in managing real-time data inputs, enforcing retention and access policies, and ensuring compliance with applicable regulations. Deployment risk management encompasses technical failures and reputational and ethical exposure. Organizations must be prepared to detect and respond to emerging threats, including adversarial exploitation, model misuse, fairness violations, and threat intelligence. User training ensures that users understand how to interact with and interpret AI outputs appropriately.

Continuous monitoring ensures the integrity of AI systems over time, enabling organizations to detect data or performance drift, latency spikes, fairness degradation, hallucinations, and unexpected user inputs and behaviors. Monitoring systems track input distributions, model confidence, subgroup impacts, and system health in real time. Governance frameworks define who is responsible for reviewing and monitoring data, which thresholds trigger alerts and incidents, and how retraining or suspension decisions are made. Maintenance schedules include regular evaluation cycles, hotfix protocols, and cost-benefit analyses for retraining. Human oversight remains essential, both as a real-time safety mechanism and a strategic control and governance function.

Monitoring logs and transparency reports provide the auditability and accountability required for responsible and compliant AI use. These artifacts capture governance proceedings and AI system activities, including key events, model decisions, and system behaviors, enabling traceability for both internal stakeholders and regulators. Ultimately, monitoring is not just about technical continuity but also preserves public trust, minimizes harm, and ensures that deployed AI systems remain aligned with their original purpose.

AI governance extends into production environments with the same rigor applied during development. By embedding oversight, adaptability, and transparency into every stage of deployment and monitoring, organizations can manage the full operational life of AI systems with integrity and confidence.

Exam Essentials

Model readiness requires more than accuracy: it requires documentation, governance, and conformity. Before deployment, AI systems must meet both technical and governance criteria, including the creation of model cards, risk assessments, and documentation of limitations and intended use. In many regulatory contexts, such as under the EU AI Act, formal conformity assessments are required for high-risk systems.

Deployment must align with organizational policies, legal obligations, and operational safeguards. The deployment phase of an AI system is the intersection of technical execution with legal and ethical responsibilities. Organizations must employ version control, enforce environment separation, and ensure data governance policies are followed. Deployments to production should be auditable, reversible, and compliant with cross-border data regulations and consent requirements.

Risk management in production includes monitoring for drift, fairness degradation, and emerging threats. Once deployed, an AI system is exposed to real-world data and usage patterns that can deviate from its training context. Continuous monitoring helps detect input drift, bias, hallucinations, and attacks and should be supported by alerting tools and defined incident response procedures.

Retraining and maintenance must be scheduled based on system volatility and risk exposure. AI models are not static: they require retraining to remain effective. Retraining frequency depends on how rapidly the operating environment changes and the criticality of the model's outputs. High-risk or high-volume systems (for example, healthcare or customer service chatbots) may require monthly retraining with human validation.

Monitoring is both a technical function and a governance responsibility. Effective monitoring includes logging, alerting, fairness analysis, human oversight, incidents, and regulations. These outputs feed into compliance efforts, risk registers, and responsible AI programs. Monitoring also provides critical evidence for audits, transparency reporting, and ethical assurance across an AI system's lifecycle.

Review Questions

The following questions are designed to test your understanding of this chapter's material. For more information on how to obtain additional questions, see this book's Introduction. You can find the answers in Appendix A.

1. What is the primary purpose of a model card in AI governance?
   - **A.** To document source code for internal reuse
   - **B.** To certify the model's performance metrics in court
   - **C.** To summarize key information about the model, including purpose, performance, and limitations
   - **D.** To encrypt the model for secure deployment
2. Which deployment environment is most suitable for low-latency, privacy-sensitive AI applications?
   - **A.** Public cloud
   - **B.** On-premises data center
   - **C.** Shared hosting platform
   - **D.** Edge device
3. What is the role of version control in AI deployment?
   - **A.** It tracks changes to model code, configurations, and data over time
   - **B.** It stores user feedback from production systems
   - **C.** It prevents users from accessing outdated documentation
   - **D.** It encrypts the model before it goes live
4. Which of the following best describes concept drift?
   - **A.** An update that changes model hyperparameters
   - **B.** A shift in the relationship between inputs and outputs
   - **C.** A temporary system outage
   - **D.** A scheduled retraining event
5. What is a key feature of retrieval augmented generation (RAG)?
   - **A.** The model automatically self-retrains based on user input
   - **B.** It compresses the model to reduce latency
   - **C.** It queries external data sources during inference to enhance accuracy
   - **D.** It anonymizes training data for compliance

6. Which team typically owns the responsibility for monitoring data lineage and access compliance?
   - **A.** IT operations
   - **B.** Data governance
   - **C.** Marketing
   - **D.** End users
7. Why is it essential to include a rollback plan in AI deployment procedures?
   - **A.** To avoid needing a model card
   - **B.** To meet encryption requirements
   - **C.** To quickly restore a prior state in case of failure or harm
   - **D.** To reduce licensing costs
8. Which of the following is a good indicator that a deployed AI model may be exhibiting fairness issues?
   - **A.** Reduced GPU usage
   - **B.** Consistent latency across regions
   - **C.** Fewer API requests from users
   - **D.** Disparities in outcomes for different demographic groups
9. What is the primary function of a "shadow model" in AI operations?
   - **A.** To simulate attacks for penetration testing
   - **B.** To generate synthetic training data
   - **C.** To operate alongside the live model for comparison and drift detection
   - **D.** To block unauthorized access attempts
10. Why should AI retraining schedules be based on model volatility and risk exposure?
    - **A.** To match software patch cycles
    - **B.** To fulfill entertainment value in user experiences
    - **C.** To optimize system cooling requirements
    - **D.** To balance model freshness with resource constraints and potential harm



Chapter

11

AI Risk Management and Assurance

THIS CHAPTER COVERS PORTIONS OF AIGP DOMAIN III, "UNDERSTANDING HOW TO GOVERN AI DEVELOPMENT," SUBDOMAIN III.C, "GOVERN THE RELEASE, MONITORING AND MAINTENANCE OF THE AI MODEL," AS WELL AS PORTIONS OF AIGP DOMAIN IV, "UNDERSTANDING HOW TO GOVERN AI DEPLOYMENT AND USE," SUBDOMAIN IV.B, "PERFORM KEY ACTIVITIES TO ASSESS THE AI MODEL," AND SUBDOMAIN IV.C, "GOVERN THE DEPLOYMENT AND USE OF THE AI MODEL." THE FOLLOWING TOPICS ARE COVERED:

- ✔ AI risk management and risk forecasting
- ✔ AI system testing and audits
- ✔ AI system incident management

THE AI GOVERNANCE PROFESSIONAL (AIGP) PERFORMANCE INDICATORS IN THESE SUBDOMAINS INCLUDE:

- ✔ Identify and evaluate key terms and risks in the vendor or open-source agreement.
- ✔ Forecast and reduce risks of secondary or unintended uses and downstream harms.
- ✔ Conduct periodic activities to assess the AI model's performance, reliability and safety (for example, audits, red teaming, threat modeling and security testing).



- ✔ **Manage and document incidents, issues and risks.**
- ✔ **Collaborate with cross-functional stakeholders to understand why incidents arise from AI models (for example, brittleness, lack of robustness, lack of quality data, insufficient testing, and model or data drift).**
- ✔ **Make public disclosures with to meet transparency obligations (for example, technical documentation, instructions for use to deployers, and post-market monitoring plans).**



As AI systems become embedded in critical business operations, organizations must be equipped to identify, forecast, and mitigate a broad range of technical, legal, and ethical risks that may impact ongoing business operations. This chapter examines the principles of building resilient AI governance through effective risk management, audits, and incident response. Resilient AI governance begins with strategies for forecasting AI-specific risks, and evaluating key terms and clauses in vendor or open-source agreements. The chapter then covers the tools and techniques for assessing AI model performance and safety, including audits, red teaming, and security testing. Finally, the chapter describes how to manage AI-related incidents and fulfill transparency and disclosure obligations. These practices support an end-to-end assurance program, enabling organizations to anticipate harm, respond effectively to failures, and meet both regulatory and stakeholder expectations.

About the Coverage of AIGP Domains in This Chapter

This chapter covers elements from the AIGP body of knowledge from Parts III and IV to better align the subject matter and avoid repetition.

Refer to the AIGP domain chart in this book's Introduction to better understand how the AIGP domains are covered by the chapters in this book.

AI Risk Management and Forecasting

While AI systems introduce powerful new capabilities, they pose novel risks that are not easily captured by traditional enterprise risk management (ERM) approaches. While conventional IT risk frameworks focus on infrastructure availability, data loss, or system failure, AI risks arise from probabilistic behavior, evolving outputs, dependency on third-party models, and societal-level impacts.

Cybersecurity professionals would say that AI-related risks exist at or above Layer 7 (of the OSI interconnect model), meaning these risks are related to the socio-technical uses of data, rather than the logic of information systems and their supporting infrastructure.



It is assumed that organizations utilizing AI have mature cybersecurity and enterprise risk management (ERM) practices in place. This chapter focuses on AI-specific considerations, which can be considered an overlay to standard risk management practices.

Proactively Identifying and Categorizing AI Risks

Effective AI risk management begins with clear *risk* identification. Unlike static IT systems, AI models evolve based on training data, feedback loops, environmental inputs, and retraining schedules, which means new risks can emerge even after deployment.

In terms of current risk management thinking, this is not so different from risk management of traditional applications and information systems. A system that is secure (free from exploitable vulnerabilities) today will likely not remain secure tomorrow, as researchers will continue to discover new vulnerabilities. Thus, any claim that a system is risk-free is fleeting at best.

AI risks can be classified into the following domains:

- **Data-related risks:** Biases, poor quality, leakage, prohibited cross-border data transfers, or illegal/unethical use of personal data.
- **Model-related risks:** Overfitting, lack of transparency, the use of black box models, or *vulnerabilities* that may permit prompt injection attacks and model poisoning attacks.
- **Deployment risks:** Misuse in contexts not intended initially, automation surprises, and performance drift.
- **Operational risks:** Resource consumption, integration failure, and cascading failures.
- **Ethical and social risks:** Discrimination, loss of agency, and manipulation of public opinion.
- **Regulatory and legal risks:** Non-compliance with emerging AI-specific or general data protection laws.

The uncertainties of AI systems, combined with the speed and sophistication with which cybercriminals attack and exploit information systems, should compel every organization to proactively and aggressively identify and mitigate risks. There is simply more *inherent risk* associated with AI than with many other emerging technologies. Instead of sitting back and waiting for risks to manifest themselves, go out there and look for them.

As an example, a financial institution adopted a third-party fraud detection AI model trained on historical claims data. Post-deployment audits revealed that the model had a disproportionately high false-positive rate for foreign-born customers, traced to bias in the training set. The reputational and legal risks led to a system recall and a public apology.

Risk Management in AI Supply Chains

Most organizations do not build AI systems from scratch. Instead, they rely on vendors, open-source models, or cloud AI services. These options may be easier to implement, but they create dependencies on third parties that must be proactively managed.

When procuring or deploying third-party AI tools, key terms in agreements can have direct implications for risk. The essential practice in third-party risk management (TPRM) is to conduct as thorough an upfront assessment of a prospective vendor and its tools as possible, and then use that assessment to inform contract negotiations. Governance teams should identify and evaluate:

- **Licensing terms:** Open-source AI components (for example, models, datasets, and frameworks) may come with restrictive or permissive licensing requirements (for example, GPL, AGPL), which can impose obligations such as code disclosure or usage limitations.
- **Data use restrictions:** Vendor agreements should clearly address data ownership, permitted uses, and handling of sensitive data. This includes whether the vendor can retain or use customer data for retraining.
- **Model transparency and explainability:** Governance professionals should request disclosures on training data sources, model lineage, and performance evaluation methods, as well as on mechanisms for auditing and challenging model decisions.
- **Indemnity and liability:** AI vendors often disclaim responsibility for any harmful outcomes resulting from model outputs. Organizations should carefully negotiate liability clauses to ensure recourse in the event of legal, financial, or reputational damage.
- **Cyber liability insurance:** Organizations should require vendors to carry cyber liability insurance, particularly for smaller vendors, where breaches may impose severe financial hardship.
- **Security and compliance posture:** Request evidence of certifications or attestations (for example, ISO/IEC 27001, ISO/IEC 42001, and SOC 2 Type II) and documentation of secure model development practices (for example, data sanitization, vulnerability testing, and model hardening).

For example, an open-source language model released under a non-commercial license was inadvertently integrated into a for-profit chatbot platform, exposing the company to potential claims of intellectual property infringement.

Common Contract Clauses to Scrutinize

Vendors are known for asserting one-sided contract terms that provide few rights to customers. Here are some examples, the likes of which should be considered red flags:

- Data sent to the model becomes the property of the vendor.
- Data may be retained for model improvement purposes.
- Data may be sent to third parties to perform services.
- Output generated by the model is not warranted to be accurate or lawful.
- Customer grants irrevocable license to use feedback for future products.

Forecasting Secondary and Downstream Harms

A key component of risk management is the proactive anticipation of future risks and unintended consequences. This concept is particularly relevant to AI, as the range of AI-related consequences remains largely unexplored. Many consequences become apparent only when the system is used at scale or in unanticipated contexts.

Secondary use refers to applications of AI outputs that differ from their original intended purpose, often without official institutional awareness or control. These can include:

- **Input repurposing:** Using AI-generated content as training data elsewhere
- **System coupling:** Chaining outputs from one model into another system
- **Malicious adaptation:** Repurposing benign tools (for example, image upscalers) for harmful uses (for example, deepfakes)

Downstream harms occur when the consequences of AI outputs are felt by individuals, communities, or systems not initially considered by the developer or deployer. These harms may include:

- **Societal:** For example, discriminatory hiring practices from algorithmic resume screening
- **Environmental:** For instance, massive compute demands for AI training leading to excessive carbon emissions
- **Political:** For example, LLMs unintentionally generating misinformation during elections

Tools for mapping downstream harm may include stakeholder impact mapping, recursive scenario modeling, external consultation panels, focus groups, misuse testing, and red teaming.

Governance frameworks should extend risk reviews beyond direct users and customers to include stakeholders affected by second- and third-order effects.

For example, an AI model trained to generate academic essays was later used by students to cheat on exams, triggering plagiarism scandals and questions of institutional integrity.

Risk Forecasting Techniques

The practice of forecasting AI risk involves utilizing predictive modeling to see how systems might behave under various internal and external pressures. These techniques include:

- **Model drift detection:** AI model performance can degrade over time due to *data drift* (input distribution shift) or *concept drift* (changes in the target variable). These drifts may increase bias, error rates, or harm.
- **Anticipating feedback loops:** AI systems that influence the environment from which they learn (for example, recommendation engines and predictive policing) can create self-reinforcing *feedback loops* that reinforce bias or lead to runaway behavior.
- **Risk forecasting through simulation:** Digital twins, synthetic data, or simulated user environments can help organizations anticipate how AI systems will behave in real-world and adversarial scenarios.

Consider using statistical drift detection tools such as KS tests and KL divergence to trigger retraining or alerts when thresholds are crossed.

For example, one community's predictive policing algorithm sent more officers to certain neighborhoods, resulting in more arrests, which in turn reinforced the location as high-risk in the retrained data.

Embedding Risk Management into Governance Structures

AI risk management is a lifecycle process that is a key component of AI governance. AI risk management should be integrated into each phase of the development lifecycle:

- **Initiation:** Define use case risk level; consult cross-functional stakeholders
- **Design:** Perform model risk assessments, including harm forecasting; identify training data provenance
- **Development:** Use explainable, testable algorithms; track lineage
- **Testing:** Include adversarial testing, red teaming, and misuse potential evaluation
- **Deployment:** Establish rollback procedures and incident thresholds
- **Post-deployment:** Monitor drift, *audit logs*, user feedback, and incident reports

Include AI-related risks in the ERM risk register and utilize the same risk management process to process all risks, including AI risks.

Assigning clear risk-related roles improves governance and enables traceability. Consider appointing these roles:

- **AI risk officer or chief risk officer:** Accountable for enterprise-level risk posture
- **Business owners/senior executives:** Responsible for risk treatment decisions
- **System owners:** Responsible for individual models
- **Audit leads:** Oversee testing and review cadence
- **Data protection officer (DPO)/legal counsel:** Oversee compliance with data protection, intellectual property, and AI-specific regulatory requirements

AI governance programs often utilize tooling to organize their proceedings, including:

- NIST AI Risk Management Framework (AI RMF)
- ISO/IEC 42001 (AI Management System Standard)
- Risk-focused model cards
- Harm-forecasting templates
- Automated risk detection in MLOps pipelines

As AI capabilities and ecosystems mature, governance professionals should prioritize tooling that enables ongoing risk detection, transparency, and remediation. In environments with layered AI components, risk forecasting must become more granular, continuous, and adaptive.

Documentation and Assurance

Complete and accurate documentation supports internal risk communication, external audits, incident investigation, model troubleshooting, and regulatory reviews. Documentation and *records* artifacts should include:

- AI risk register (a dynamic, living business record)
- Third-party model assessments
- Data provenance and consent records
- Data lineage records
- Impact assessments (for example, Algorithmic Impact Assessment, DPIA)
- Risk management records, including risk treatment decisions, remediation plans, and residual risk justifications
- Model and system cards (concise documentation describing purpose, inputs, limitations, and evaluation results)

Points to cover for AI model assurance reviews should include:

- Known failure modes of an AI system
- How risks are mitigated, and what residual risks remain
- Who is accountable for each class of risk
- How frequently risk controls are tested
- How discovered failures or harms are investigated, communicated, and corrected

Regulatory Considerations

Global and regional AI laws are being enacted faster than can be addressed in this book. Many of these global AI laws require formal risk management processes. Example requirements include:

- **EU AI Act:** Requires providers of high-risk AI systems to perform risk assessments, maintain logs, and submit systems for conformity assessments.
- **Canada Artificial Intelligence and Data Act (AIDA):** Requires impact assessments and public disclosure for high-impact AI systems (as of the time of writing, this law has not yet been enacted).
- **U.S. Executive Order on AI (2025):** Encourages agencies to ensure that AI systems are "free from ideological bias or engineered social agendas."
- **China Interim Measures for the Management of Generative AI Services:** Requires the classification of AI systems and oversight based on risk and category, as well as the mandatory labeling of AI-generated content.

Organizations that fail to forecast and manage risk properly may be subject to penalties, injunctions, or civil liability.

AI System Audits and Testing

To ensure responsible, resilient, and compliant AI, organizations must go beyond model development and actively assess how AI systems perform in the real world. These assessments require a deliberate, recurring process of audits, red teaming, threat modeling, and security testing. Each provides distinct insights into the performance, reliability, safety, and vulnerability of AI systems. These activities help uncover unintended behavior, regulatory compliance gaps, adversarial weaknesses, and ethical concerns that might otherwise go undetected until harm occurs.



For a comprehensive study of audits and audit techniques, refer to this author's books on the Certified Information Systems Auditor (CISA) certification.

Performance Audits and Model Validation

Performance audits focus on measuring how effectively an AI system fulfills its intended purpose across various dimensions, including accuracy, fairness, robustness, performance, and generalization. These audits validate or refute claims made during development and detect degradation over time. Arguably, continuous monitoring should detect degradation earlier than an audit would.

Audits begin with the definition of acceptable performance criteria that were established in the design phase of the AI system. These criteria should include:

- **Business alignment:** For example, fraud detection accuracy
- **Context-sensitive:** For example, false negative tolerance in medical diagnosis
- **Stakeholder-informed:** For example, transparent and explicit opt-in
- **Fairness and lack of bias:** For instance, fairness across demographic groups
- **Precision and confidence:** For example, accuracy of classification and consistent decision-making
- **Robustness and resilience:** Acceptable performance even under load and with edge cases

For example, a credit scoring model met accuracy thresholds during development. However, after deployment, an audit revealed performance disparities when applied to rural applicants, prompting a reevaluation of feature inputs and model retraining.

Auditors and system owners must account for the possibility that AI models often behave differently in production environments than in controlled development and testing environments. When production systems fail to meet stated criteria, auditors may consider testing the same AI models in test environments to determine whether they behave differently. If so, this should prompt an inquiry to understand the difference.

Red Teaming for Adversarial Robustness

Red teaming is a structured process and a set of techniques for simulating adversarial attacks and misuse scenarios against a target system. Red teaming is essential for identifying vulnerabilities that internal teams may have overlooked, as well as for ensuring resilience under worst-case conditions.

There are several types of red team assessments targeting AI systems, including:

- Prompt injection and *jailbreak* testing for large language models
- Evasion attacks (modifying inputs to cause misclassification)
- Poisoning attacks (injecting corrupt data into training or fine-tuning pipelines)
- Functionality misuse (using tools for unintended or unethical purposes)
- *Social engineering* testing (for example, phishing via chatbots)

Red teaming should also evaluate the broader AI ecosystem, including supporting infrastructure, APIs, integration layers, and data pipelines. Many attacks exploit infrastructure, interfaces, and dependencies, not just model parameters.

The members of a red team should include domain experts who understand the AI system and common *threats*, policy experts who will understand misuse, and external testers who bring objectivity. Red teaming should target both the AI systems themselves and the personnel who build, manage, and use them.

Threat Modeling for AI-specific Attack Surfaces

Traditional threat modeling techniques must be extended to account for AI-specific attack surfaces, such as model inversion, training data leakage, and adversarial attacks.

A modified version of *Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, Elevation of Privilege (STRIDE)* can be used to identify threats to an AI system, including:

- Data poisoning (tampering)
- Model theft (information disclosure)
- Unauthorized API chaining (elevation of privilege)
- Prompt injection (spoofing and tampering)

Other threat modeling methodologies include:

- **MITRE ATLAS (Adversarial Threat Landscape for Artificial-intelligence Systems):** A knowledge base for adversarial threats against AI
- **Kill Chain Analysis:** Mapping attacker steps from reconnaissance to exploitation
- **Attack Trees:** Visualizing paths to compromise
- **TARA (Threat Analysis and Risk Assessment):** Adapted for AI systems
- **PASTA (Process for Attack Simulation and Threat Analysis):** Can be extended to model business-impact scenarios for AI misuse

In addition to threat modeling, *threat hunting* is an essential activity that is used to proactively search for signs of intrusions, intruders, abuse, and *indicators of compromise*. Instead of waiting for incidents to occur, threat hunting proactively looks for signs of attack, which often leads to earlier detection and can help reduce damage and harm.

For example, a public-facing generative AI model was threatened by a coordinated attempt to extract its training data. Threat modeling revealed that repeated sampling with targeted prompts could yield training data fragments, which prompted the organization to implement additional throttling and watermarking controls.

Threat Modeling Across the Lifecycle

Threat modeling should be performed at various points in the development of an AI system, including:

- **Design phase:** Identify vulnerable features and components
- **Pre-deployment phase:** Test implemented defenses
- **Post-deployment phase:** Adapt to evolving threats and contexts

Security Testing for Confidentiality and Integrity

Audits of AI systems should extend beyond performance and output quality to include how effectively they safeguard data integrity, user privacy, and system confidentiality. Security testing evaluates how AI systems behave under malicious and abnormal conditions. AI-specific security techniques include:

- **Penetration testing:** Simulate intrusion attempts into AI systems, pipelines, and APIs
- **Model extraction testing:** Assess the ease of reverse-engineering or copying an AI model
- **Inference attacks:** Determine how easily training data can be inferred from outputs
- **Membership inference:** Test whether an attacker can guess whether specific data was used to train the model
- **Re-identification testing:** Validate anonymization of training data to determine whether it can be re-identified
- **Adversarial robustness testing:** Examine how the model behaves under adversarial perturbations or input manipulations

Organizations that utilize continuous integration/continuous deployment (CI/CD) pipelines must secure these environments to prevent tampering, unauthorized updates, or the deployment of shadow AI models. Key CI/CD safeguards include:

- Monitoring of check-outs and check-ins
- Restricting access to sensitive code on a need-to-know basis

- Code signing and version control of model artifacts
- Containerization and sandboxing during model training
- Developing and monitoring policy-based access control for retraining triggers
- Audit logging for data access and model deployment events

For example, an e-commerce company discovered that a rogue employee had deployed an unapproved AI model into the pipeline. Security testing revealed missing integrity checks and weak access controls in the CI/CD chain.

Full-stack Security Testing Is Essential

Security testing must encompass the entire AI tech stack, including physical security, networks, operating systems, database management systems, application servers, storage systems, network services, and other relevant components.

Regulatory and Compliance Audits

Periodic audits of AI systems help confirm their compliance with applicable laws, legal agreements, policies, and ethical standards. These audits provide traceability, accountability, and defensibility in case of regulatory scrutiny. Key objectives of a compliance audit include:

- **Verify adherence to AI risk management frameworks:** For example, NIST AI RMF and ISO/IEC 42001
- **Evaluate required controls:** For instance, bias mitigation and impact assessments
- **Confirm documentation existence and quality:** For example, model cards, data sheets, and usage policies
- **Confirm business records existence and quality:** For example, risk register, risk treatment decisions, periodic maintenance, and opt-out requests
- **Confirm proper application of opt-out rights**
- **Confirm the presence and effectiveness of human-in-the-loop safeguards**

Selected audit requirements by jurisdiction include:

- **EU AI Act:** Requires high-risk AI systems to include a quality management system and undergo conformity assessments and post-market monitoring.
- **Canada AIDA:** Requires impact assessments and transparency measures (as of the time of writing, this law has not yet been enacted).
- **U.S. federal agencies:** Encouraged to adopt AI governance frameworks via applicable Executive Orders and OMB guidance.

Steps to prepare for an audit of an AI system include:

- Maintain a **single source of truth** for documentation.
- Ensure **audit logs** are immutable and timestamped.
- Practice **dry runs** or internal audit simulations.
- Assign **responsible owners** for each compliance area.
- Coach responsible owners on **proper conduct** with auditors.



Specific regulations or contractual stipulations may require that audits be performed by qualified, external parties rather than internal audit staff. For instance, the EU AI Act requires that an external Notified Body conduct audits in some cases.

Real-time Testing and Continuous Monitoring

Periodic assessments are not only essential but also insufficient. Organizations should implement continuous assurance mechanisms to detect emerging risks and performance degradation in real time. Continuous monitoring practices include:

- **Latency and throughput metrics** to detect performance issues
- **Drift detection** in input and output distributions
- **Anomaly detection** for rare or unexpected behaviors
- **User feedback loops** to surface safety and ethical issues
- **Incident alerts** to be aware of incidents as they occur
- **Incident alerting thresholds** for automated remediation triggers
- **Security scans** to proactively identify newly discovered vulnerabilities
- **Threat hunting** to proactively detect signs of abuse and intrusion

For example, a healthcare AI system monitoring vital signs began flagging numerous high false positives. Continuous monitoring revealed gradual input drift resulting from a firmware update in a medical device, prompting the organization to initiate a retraining cycle.

As stated elsewhere in this chapter, testing and monitoring the entire technology stack is not only essential but also considered an essential practice for any information system.

Integrating Testing into the AI Lifecycle

Audits and testing are not only performed on production AI system environments but must be embedded into every stage of the AI system's design, development, and deployment pipeline. Audits test not only the AI system itself but also the policies, business processes, and tools used to design, build, test, deploy, and monitor it.

Key points for AI lifecycle audits include:

- **Design phase:** Design reviews and approvals, risk assessments, threat modeling, and red teaming scenario planning
- **Development phase:** Unit tests for model code, adversarial test sets, repository access controls, and feature change requests and approvals
- **Pre-deployment phase:** End-to-end performance audits and regulatory reviews
- **Deployment phase:** Canary testing and rollback safeguards
- **Post-deployment phase:** Continuous monitoring, incident response planning, and incident response instances
- **Retirement phase:** Data archival, model decommissioning, disposal, and migration to other systems (AI system retirement is discussed in Chapter 12)

Some organizations implement a gate model to ensure that testing is completed before projects can proceed. Such gates might include:

- **Gate 1:** Risk assessment and design review complete, approving the start of development
- **Gate 2:** Adversarial testing and bias audit passed, approving the promotion to production use
- **Gate 3:** Operational monitoring and rollback plan validated

AI Audit Roles and Responsibilities

AI governance artifacts, or a documented audit program charter, should clearly define ownership and accountability for audits and testing activities. Common roles include:

- **AI risk officer:** Oversees assurance strategy across systems
- **Model owner:** Responsible for system-specific testing results
- **Internal audit lead:** Coordinates internal and third-party audit activities
- **Red team coordinator:** Designs and executes adversarial evaluations
- **Security engineer:** Conducts penetration, integrity testing, and threat modeling
- **Ethics advisor:** Guides social harm evaluations

Assigning these roles ensures testing efforts are not siloed within technical teams alone but instead reflect cross-functional governance.

Audit Documentation and Evidence Retention

Audit charters, plans, results, test reports, and monitoring data must be retained in a manner that supports accountability and future inquiries. An organization's records retention schedule should specify the retention period for each type of business record, after which

it should be destroyed in accordance with the organization's policy. Audit documentation artifacts include:

- Audit charter that defines the audit program and audit roles and responsibilities
- Audit reports with findings, severity levels, and recommendations
- Red teaming logs and adversarial scenarios tested
- Threat models with attack paths and mitigations
- Threat modeling records
- Testing protocols and reproducible test sets
- Model version histories and deployment metadata
- Evidence of remediation actions taken

Storage must meet regulatory, legal, and ethical obligations, including access controls, retention schedules, encryption, and data minimization. Organizations should consult with their legal counsel to determine the applicability of retention requirements.

Audit and Testing Tooling and Automation

Manual testing does not scale well across a growing portfolio of AI systems, nor is it effective for non-AI traditional business applications. Organizations should invest in tooling to automate key parts of the audit and testing workflow. Examples of AI-specific testing tooling include:

- **Bias audit frameworks:** For instance, Aequitas and Fairlearn
- **Red teaming platforms:** For example, Microsoft's Azure AI Red Team tooling
- **Security scanners:** For testing AI development, testing, and production environments
- **Model card and data sheet generators**
- **Monitoring dashboards:** For example, Evidently AI, Arize, and WhyLabs

Tool selection should align with system complexity, risk level, and internal capabilities.

Incident Management and Disclosures

In the cybersecurity world, it is generally understood that, despite the best efforts to prevent events and *incidents*, they will still occur—requiring effective incident detection and response capabilities. Likewise, despite rigorous risk controls and testing, AI systems may still generate harmful, misleading, or otherwise unintended outputs after deployment, making incident management a critical component of AI governance. A robust approach not only limits harm when incidents occur but also supports organizational learning, transparency obligations, and compliance with existing and emerging regulations.

Identifying and Managing AI Incidents

An *AI incident* refers to any unexpected behavior, outcome, or interaction involving an AI system that causes, or has the potential to cause, harm, operational disruption, or regulatory concern.

Examples of AI incidents include:

- A recommendation system surfacing harmful or illegal content
- A chatbot providing incorrect medical advice
- A classification model mislabeling sensitive demographic attributes
- An autonomous system failing to operate safely under edge case conditions
- A generative model producing IP-infringing, toxic, or factually incorrect outputs
- Unexpected performance degradation due to concept drift or environmental changes

AI incident response should be modeled after established IT and cybersecurity incident management processes and adapted for the probabilistic and opaque nature of AI systems. Organizations should consider enacting a single incident response framework that encompasses IT, cybersecurity, privacy, and AI-related incidents.

Key steps of incident response are as follows:

1. **Detection:** Enabled by various monitoring activities, detection identifies anomalous or harmful behavior through alerts, monitoring, or user reports
2. **Triage:** Assessment of an incident's severity, scope, and potential impact
3. **Communication:** Informing internal and external stakeholders
4. **Investigation:** Determining root cause, including model and data lineage
5. **Containment:** Isolating or disabling the affected system or component, preventing its spread or persistence
6. **Eradication:** Implementing corrective actions, such as model retraining, data updates, or policy changes
7. **Recovery:** Restoring the system to its pre-incident state (but with changes enacted to prevent a similar incident)
8. ***After-action review (AAR)***: Conducting a *root cause analysis* to better understand the real cause of the incident, enabling effective long-term fixes
9. **Documentation:** Updating policies, logs, and audit trails

For example, a generative AI-powered customer support assistant began giving customers incorrect return instructions due to a prompt leak from a test environment. Triage identified the source as a misrouted input stream. The model was temporarily suspended, retrained, and revalidated with updated input filters.

Not All Incidents Are Failures

Some incidents stem not from bugs but from misaligned objectives. For example, a model trained to maximize engagement may inadvertently learn to promote sensationalist or polarizing content. Incident response must therefore consider *intent vs. outcome.*

Categorizing and Escalating Incidents

Incidents vary widely in their severity and impact; thus, not all incidents warrant the same level of response or disclosure. Clear categorization facilitates triage efforts, determines communication and escalation, and prioritizes remediation. A typical severity scale includes these levels:

- **Level 1—Critical:** Widespread harm, regulatory violation, or systemic failure
- **Level 2—Major:** Significant user impact or model breakdown in core function
- **Level 3—Moderate:** Isolated issue with limited exposure
- **Level 4—Minor:** No external impact; handled by internal monitoring

Escalation options for higher-severity incidents include:

- Executive notifications and periodic updates
- External regulator notification
- Legal hold of related artifacts
- Protection of records through *attorney-client privilege*
- Model suspension or rollback

Organizations that utilize automated workflows can configure them to trigger escalation rules based on incident metadata.

Documenting Incidents, Issues, and Risks

While required by some regulations, it's an essential practice to document all events, incidents, issues, and risks associated with both AI systems and non-AI systems. Effective documentation ensures traceability, regulatory readiness, and improved organizational memory. Good recordkeeping also enables trend analysis and proactive mitigation of recurring patterns.

Essential practices for documenting incidents include:

- Date/time of detection
- Detection source (monitoring system, user report, audit)

- System, service, or model affected
- Description of behavior and observed consequences
- Severity classification
- Immediate response actions taken (and by whom)
- Root cause and contributing factors
- Mitigation measures and verification steps
- Status and resolution date

The use of standardized taxonomies (for example, incident type, system component, and harm type) helps to facilitate analysis, cross-team understanding, metrics, and reporting.

It is common for organizations to record one or more risks in the risk register when an incident reveals risks that were not previously considered. This works in the opposite direction as well: each incident should be traced to the relevant risks in the organization's AI risk register that already existed when the incident occurred.

Residual Risk After Remediation

Risks are rarely eliminated entirely. When documenting an incident, note whether the applied fix reduces or eliminates the risk, and whether residual risk remains within acceptable thresholds.

Collaborating with Stakeholders to Understand Root Causes

Understanding the reasons why AI incidents occur generally requires a multidisciplinary approach. Technical root causes may be related to design decisions, policy choices, and contextual misunderstandings.

Common root causes of AI-related incidents include:

- **Model brittleness:** Inability to handle edge cases or unanticipated inputs
- **Lack of robustness:** Small perturbations cause unpredictable outputs
- **Insufficient or poor-quality training data**
- **Misaligned objectives:** For example, optimizing engagement vs. user safety
- **Overreliance on automation or absence of human oversight:** Inadequate human supervision or human-in-the-loop
- **Inadequate testing, especially under adversarial or real-world conditions**
- **Model or data drift**
- **Failures in prompt design for generative AI**

Post-incident reviews should be formalized and include many, if not all, AI system stakeholders. Suggested roles are as follows:

- **Data scientists:** Explain model behavior and training data lineage
- **Software engineers:** Investigate system logs and integration layers
- **Cybersecurity:** Understand and explain related vulnerabilities and threats
- **Privacy:** Evaluate data subject rights failures
- **Ethics and legal advisors:** Evaluate regulatory and societal impacts
- **Operations staff:** Identify downstream business or customer effects
- **End users (where appropriate):** Provide contextual understanding of harm
- **External experts:** Bring objectivity to bear on the evaluation

For example, a predictive equipment maintenance model failed to flag a major component failure in time. A cross-functional review revealed that the model was trained exclusively on normal operating conditions, resulting in a blind spot for rare but critical failure modes.

Avoiding a Culture of Blame

The goal of incident analysis is to learn and improve, not to identify and punish the guilty. Organizations that foster psychological safety in postmortem reviews will encourage open sharing and process improvement.

Transparency and Public Disclosures

Increasingly, security and privacy incidents can no longer be swept under the rug. Regulations in many jurisdictions require the disclosure of privacy incidents, and increasingly, of AI incidents. Transparency is a foundational principle of AI governance, and a regulatory requirement. Certain incidents or risks must be disclosed to users, regulators, or the general public to ensure accountability and informed decision-making.

Disclosure requirements vary by law and context and may include:

- High-risk AI system deployment
- Material performance degradation
- AI system safety failures
- Significant data breaches affecting AI inputs or outputs
- Inaccurate or misleading outputs that affect public health or safety
- System behavior outside intended use cases

Some of the regulations requiring transparency include:

- **EU AI Act:** Requires deployers of high-risk AI systems to maintain technical documentation, usage instructions, user complaints, performance, risks and remediation, and post-market monitoring plans
- **Canada AIDA:** Mandates public disclosure of the purpose and function of high-impact AI systems (as of the time of writing, this law has not yet been enacted)
- **California CPRA and U.S. FTC guidance:** Require companies to explain automated decision-making in specific contexts
- **Brazil LGPD:** Gives individuals the right to explanation and review of AI-driven decisions

Disclosures may need to be tailored to the needs and rights of each audience, including:

- **Users and customers:** Simple, actionable explanations of what occurred and what is being done
- **Regulators:** Technical documentation, timelines, impact assessments, and remedial actions
- **Partners and vendors:** System-level implications and interface updates
- **Internal stakeholders:** Root cause analysis, revised policies, mitigation details, and training updates

Organizations can be proactive and plan for incidents, including preparing disclosures well in advance of any incident. Examples include:

- Public statement and FAQ describing an incident and the organization's response
- Timeline graphic of the incident
- Plain-language explanation of model behavior
- Contact information for inquiries

For example, a social media company publicly disclosed that its content moderation AI failed to detect and block a widely disseminated hate speech campaign, triggering reputational damage and increased regulatory scrutiny.

Organizational Preparedness and Readiness

Establishing an AI incident response program requires careful planning and coordination, thorough training, and robust infrastructure. The main components of such a program include:

- Documented AI *Incident Response Plan (IRP)* as a standalone document or (ideally) a part of an overall IT, cybersecurity, and privacy IRP
- *Playbooks* for common AI-related incident types
- Testing and simulation of response scenarios
- Integrated response with cybersecurity and legal teams
- Training programs for employees on incident detection and reporting

Training for incident responders is often combined with a *simulation* exercise, where a moderator guides incident responders through a developing incident scenario, allowing them to discuss their approach to the incident and their response procedures. Simulations like this help build "muscle memory" so that responders are better prepared when a real incident occurs.

Potential simulation scenarios include:

- A generative model outputs copyrighted material
- A predictive model fails due to a drifted input feature
- Red teaming reveals a security vulnerability in a foundation model
- Threat hunters find that an outside attacker has recently compromised an AI system
- A regulator requests all logs from a recent incident



For organizations with cybersecurity insurance, policies may stipulate specific terms and conditions related to incident response, including the implementation of controls such as multifactor authentication, detailed audit logging, and others.

Summary

AI risk management and assurance is a foundational aspect of responsible AI governance, addressing how risks are identified, evaluated, and managed throughout the AI system lifecycle. The chapter begins with a discussion of risk forecasting and categorization, including new risk domains that extend beyond traditional IT frameworks. Practitioners and risk managers must consider AI-specific risks, including bias, drift, lack of transparency, and downstream harms, especially when third-party or open-source components are used. Understanding vendor agreements, licensing terms, data use restrictions, and liability clauses is crucial for comprehending the full extent of the exposure an organization assumes when adopting external AI tools. Forecasting also plays a critical role, which requires governance teams to anticipate secondary uses, misuse, and emerging threats through simulation, drift monitoring, and stakeholder impact mapping.

The nature of AI systems underscores the importance of performance audits, adversarial red teaming, threat modeling, and security testing. Audits validate real-world functionality and uncover issues that may not emerge in development environments or during testing. Red teaming enables proactive exposure of vulnerabilities by simulating malicious use or prompt manipulation, particularly relevant for foundation models and LLMs. Threat modeling introduces AI-specific concerns such as model theft or poisoning attacks, while security testing helps ensure the confidentiality and integrity of training data, model weights, and deployment environments. Threat hunting takes the fight to adversaries by proactively searching for indicators of compromise. These and other activities underpin an assurance program that enables traceability, detectability, and resilience. They also support compliance

with regulatory mandates, which increasingly require formal documents and records, including audit records, usage logs, and third-party reviews.

When issues do arise, robust, documented practices for incident management and disclosure ensure minimal impact and compliance with regulations. AI incidents may include unexpected outputs, model hallucinations, harmful recommendations, systems behaving outside their intended boundaries, and a range of incidents in the underlying infrastructure that may compromise the integrity and confidentiality of affected AI systems. Managing these events involves clear detection, triage, communication, investigation, containment, recovery, post-incident review, and documentation. Root causes of AI incidents may stem from brittle models, poor data quality, insufficient testing, or model drift. Effective incident response requires collaboration across all stakeholders. Documentation feeds back into risk registers and compliance reports, while well-planned disclosures promote transparency and regulatory alignment. Public-facing models, particularly those used in high-risk domains, must be accompanied by clear instructions, post-market monitoring plans, and tailored disclosures for different audiences.

All of these practices should be considered an overlay of existing programs and practices, including risk management, audits, and incident response.

Exam Essentials

AI risk must be assessed across the entire lifecycle, not just at deployment. AI systems evolve over time through retraining, user interactions, and environmental changes. A robust risk management program must include early-stage risk identification during design and data collection, mid-stage validation and testing, and post-deployment monitoring and forecasting. This continuous view ensures risks are not missed as the system matures or is repurposed.

Vendor and open-source AI components may obscure or introduce hidden contractual and technical risks. Organizations often integrate third-party AI tools, which may include restrictive licenses, data usage and ownership clauses, and disclaimers of liability. Exam candidates should understand the importance of identifying and vetting vendors, reviewing and revising vendor agreements, assessing transparency, and evaluating whether licensing or support terms align with the organization's risk tolerance and regulatory obligations.

Auditing and testing AI models requires a blend of technical and adversarial methods. Effective AI system assurance requires performance audits, red teaming, threat modeling, threat hunting, and security testing. These activities reveal vulnerabilities such as bias, prompt injection, model poisoning, data leakage, or system misuse. Candidates should be familiar with these methods and know when each is most appropriate, especially in safety-critical or regulated domains.

AI incident response must include documentation, cross-functional review, and a feedback loop into governance processes. When an AI system causes harm or fails unexpectedly, organizations must detect the issue early, respond quickly, contain the impact, and identify the root cause. Exam-takers should be familiar with the steps of AI-specific incident management, including detection, triage, communication, investigation, containment, eradication, recovery, post-incident review, and documentation, and know how these align with compliance requirements and stakeholder expectations.

Public disclosure of AI risks and incidents is increasingly required by law. Frameworks, such as the EU AI Act and Canada's AIDA (as of the time of writing, AIDA has not yet been enacted), require transparency for high-risk systems, including technical documentation, usage instructions, and post-market monitoring. Candidates should understand the scope of disclosure obligations, how to tailor communications to different audiences, and the documentation artifacts that support compliance with these obligations.

Review Questions

The following questions are designed to test your understanding of this chapter's material. For more information on how to obtain additional questions, see this book's Introduction. You can find the answers in Appendix A.

1. Which of the following is most critical when forecasting downstream harm from an AI system?
   - **A.** Stakeholder impact mapping
   - **B.** Model inference speed
   - **C.** Data labeling automation
   - **D.** Algorithm version numbering
2. What is the primary goal of red teaming in AI assurance?
   - **A.** Increasing model accuracy
   - **B.** Identifying adversarial vulnerabilities and misuse scenarios
   - **C.** Reducing training costs
   - **D.** Automating model updates
3. Which clause in a vendor agreement could most directly affect the organization's data governance obligations?
   - **A.** Termination clause
   - **B.** Model refresh schedule
   - **C.** Data usage rights
   - **D.** Warranty exclusions
4. What is the best method for detecting model drift after deployment?
   - **A.** Comparing audit logs with the risk register
   - **B.** Rereading the model documentation
   - **C.** Conducting monthly user surveys
   - **D.** Monitoring shifts in input or output data distributions
5. An AI chatbot begins generating biased or toxic responses. What is the most appropriate first step?
   - **A.** Increase its training epochs
   - **B.** Escalate the incident and disable the model temporarily
   - **C.** Ignore the issue and monitor it over time
   - **D.** Retrain the model on different hardware

6. Which of the following best describes a "shadow deployment"?
   - **A.** A deployment of an untested model to the public
   - **B.** A model released under a different product name
   - **C.** A model run in parallel without impacting end users
   - **D.** An outdated model left running without documentation

7. Which regulatory framework explicitly requires post-market monitoring for high-risk AI systems?
   - **A.** EU AI Act
   - **B.** California CPRA
   - **C.** Brazil LGPD
   - **D.** ISO/IEC 27001

8. What is the key difference between a performance audit and a threat model?
   - **A.** Threat modeling is conducted after deployment
   - **B.** Performance audits assess legal compliance
   - **C.** Threat modeling identifies attack vectors; audits evaluate functional behavior
   - **D.** Performance audits focus on infrastructure, not on AI

9. Why is prompt injection a significant concern for generative AI systems?
   - **A.** It slows inference times
   - **B.** It may bypass safety mechanisms and produce harmful outputs
   - **C.** It can overwrite hyperparameters
   - **D.** It only affects offline training models

10. What is the purpose of linking incident records to the AI risk register?
    - **A.** To consolidate training logs
    - **B.** To simplify contract negotiation
    - **C.** To ensure discovered risks are tracked, evaluated, and managed
    - **D.** To reduce model latency



Chapter

12

Ongoing AI Operations

THIS CHAPTER COVERS PORTIONS OF AIGP DOMAIN IV, "UNDERSTANDING HOW TO GOVERN AI DEPLOYMENT AND USE," SUBDOMAIN IV.C, "GOVERN THE DEPLOYMENT AND USE OF THE AI MODEL." THE FOLLOWING TOPICS ARE COVERED:

- ✔ Managing business records for AI systems
- ✔ External communications
- ✔ Retiring AI systems

THE AI GOVERNANCE PROFESSIONAL (AIGP) PERFORMANCE INDICATORS IN THIS SUBDOMAIN THAT ARE COVERED BY THIS CHAPTER INCLUDE:

- ✔ Document incidents, issues, risks and post-market monitoring plans.
- ✔ Establish external communication plans.
- ✔ Create and implement a policy and controls to deactivate or localize an AI model as necessary (for example, due to regulatory requirements or performance issues).



As AI systems transition from development to deployment, their impact is determined by design decisions, as well as how they are operated, maintained, and eventually retired. Governance is required to ensure AI systems remain effective, compliant, and aligned with organizational values over time. This chapter begins with a discussion on business records, including the documentation of incidents, issues, risks, and post-market monitoring results. Next, external communication plans help provide transparency, build trust, and meet legal obligations. This chapter concludes with AI system retirement, the responsible deactivation or localization of models in response to performance, ethical, or regulatory triggers. These practices define the operational backbone of AI governance, enabling adaptation to various circumstances while maintaining accountability throughout the AI system lifecycle.

Managing Business Records

Business records tell the story of an organization's operations and proceedings. They are essential in the context of AI systems, to ensure organizational accountability, regulatory compliance, and continuous system improvement. The systematic documentation of incidents, issues, risks, decisions, and post-market monitoring activities provides a foundation for the responsible management of AI systems across all the activities in AI development and governance discussed throughout this book. This section discusses essential practices for maintaining these records and the governance structures that support their long-term integrity.



Many of the topics in this section are described in greater detail in other chapters but are brought together here in an overarching discussion of business records.

Incident Documentation and Tracking

Incident records are crucial for AI operations and governance, as they enable organizations to document events that deviate from intended system behavior, whether caused by software bugs, model drift, data issues, or unexpected user interactions. Incident management is covered fully in Chapter 11.

The types of incidents that should be recorded in detail include:

- **Model performance failures:** Degradation in accuracy or reliability
- **Ethical/regulatory violations:** Biased outputs, unfair decisions, and harmful recommendations
- **Security breaches:** *Adversarial inputs*, model extraction, model poisoning, and prompt injection attacks
- **System downtime:** AI system service disruptions, particularly those caused by AI components
- **Unexpected use cases:** Users applying the AI system in unintended ways

What Constitutes an Incident in AI?

A poorly ranked search result may be a mere annoyance in a content recommendation engine, but a misclassification in a healthcare diagnostic tool could be life-threatening. Classifying an event as an "incident" depends on both severity and context.

For example, an organization's employment applicant tracking system uses an AI tool to filter job candidates. The organization discovered that the system systematically ranks older applicants lower, thereby jeopardizing compliance with NYC Local Law 144, which requires annual bias audits of applicant tracking systems. The incident record includes logs from affected queries, analysis of training data distribution, internal review findings, and notes on retraining efforts with corrected data.

Issue Reporting and Lifecycle Management

Issues and events may not meet the threshold of incidents, but they still reflect undesirable system behavior, user complaints, or technical shortcomings that require investigation. They represent matters that need to be formally addressed and decisions made on next steps.

Examples of issues include:

- Minor UI/UX inconsistencies
- Sporadic misclassifications
- Feedback from internal QA testers
- Suggestions and complaints from user community forums

Organizations should implement an issue-reporting system using ticketing or workflow tools. The lifecycle of an issue typically includes:

1. **Intake:** Submission of an issue through designated channels.
2. **Categorization:** Tagging based on severity, area, and AI component.
3. **Assignment:** Routing to responsible teams.

4. **Escalation:** Routing to others when SLAs are violated.
5. **Resolution:** Implementation of fixes or improvements.
6. **Closure:** Documenting resolution and notifying stakeholders.

Mature organizations track issues within their CI/CD pipelines and retraining workflows, enabling automated alerts when repeated issues signal model degradation.

Risk Documentation and Mitigation Tracking

Because AI systems are subject to unique risks, documenting risks, potential harms, treatment decisions, and mitigation efforts are critical for risk governance. AI-related risks include:

- Model drift
- Data drift
- Bias and fairness risks
- Security vulnerabilities
- Legal and compliance risks
- Misuse or dual-use concerns

Each risk should be documented in a *risk register*, ideally within an AI governance platform or GRC (governance, risk, and compliance) tool. A typical risk register record usually includes:

- System or application name
- Risk description
- Associated use cases
- Likelihood and impact scores
- Risk treatment decision, including the person who made the decision
- Mitigation strategy (when applicable)
- Residual risk post-mitigation
- Risk owner and review schedule

For example, a bank's AI-powered loan application system includes a risk entry for racial bias due to proxy variables in historical data. The record consists of mitigation steps (for example, fairness audits and post-processing techniques), as well as periodic review dates to reassess effectiveness.

Post-market Monitoring Plans

Monitoring an AI system after its deployment, referred to as *post-market monitoring*, is crucial for identifying real-world issues that internal, pre-deployment testing and validation may have overlooked.

The goals of post-market monitoring generally include:

- Validation of continued model accuracy, robustness, and fairness

- Detection of performance degradation due to numerous causes, including data and environmental changes
- Capture of user feedback and behavioral adaptations
- Support of iterative retraining, risk assessments, and governance updates

The components of a post-market monitoring plan include:

- **Performance metrics monitoring:** Continuous observation of key performance indicators (for example, accuracy, F1 score, and latency) across user segments
- **Bias and fairness audits:** Scheduled reviews using demographic parity, equal opportunity, or counterfactual fairness tests
- **Drift detection:** Monitoring input and output distributions for statistical drift using methods like Population Stability Index or Kolmogorov–Smirnov tests
- **Data flow and transfer reviews:** Periodic reviews of data transfers into and out of the AI model, other systems, and to and from the organization itself. *Data loss prevention (DLP)* tooling can help automate this monitoring. Monitoring needs to include *cross-border data transfers* to ensure the organization remains compliant with related regulations
- **User feedback channels:** Integrating feedback mechanisms into the user interface enables real-time submission of complaints or suggestions
- **Incident and issue integration:** Feeding detected anomalies into incident and issue reporting processes and tooling for tracking and resolution

Regulatory examples of post-market monitoring include:

- **EU AI Act:** Requires high-risk AI systems to implement post-market monitoring mechanisms that include logging, reporting, and human oversight
- **U.S. Algorithmic Accountability Act:** This *proposed* law calls for impact assessments and continuous evaluation of deployed systems

Governance and Accountability Structures

To be effective, accountability for business record keeping should be formally documented across all roles and functions within the AI system lifecycle. Typical roles and responsibilities include:

- **System developers:** Maintain records of system design, training data provenance, data lineage, test plans, test results, and model versioning.
- **Product managers:** Ensure that user-facing documentation is accurate and up to date.
- **Compliance officers:** Verify that record keeping aligns with legal and regulatory mandates.
- **Risk owners:** Monitor the status of open risks and risk treatment decisions and oversee mitigation plans to ensure effective risk management.
- **Workflow and process owners:** Maintain records of AI system development reviews and approvals.

TABLE 12.1 Example AI Record Keeping RACI Matrix

Task	Responsible	Accountable	Consulted	Informed
AI system design	Developers	CIO	CISO, CIO, product manager	Executive team
Document incidents	DevOps	Risk lead	Compliance team	Product manager
Maintain risk register	Risk owner	CISO	Legal counsel	Executive team
Update monitoring plan	AI/ML engineer	Head of AI	QA lead	All stakeholders

Table 12.1 shows a typical RACI for AI record keeping.

Retention and Disposal Policies

Records associated with AI system development and operations should be included in the organization's records retention schedule. Typically, records should be retained for a legally prescribed period, typically from three to seven years, depending on jurisdiction and applicable laws. Organizations should define:

- **Retention periods:** Based on system type and risk level
- **Access controls:** Ensure only authorized personnel can access records (most records should be *immutable* to prevent tampering and destruction)
- **Archival procedures:** Store in secure, tamper-evident formats for long-term preservation
- **Deletion protocols:** Wiping records at end-of-life, using methods commensurate with risk and data sensitivity

Tools and Technologies for Business Record Management

It often makes sense to utilize tooling to manage business records, particularly because modern AI systems generate vast quantities of operational data. Key logs include:

- **Prediction logs:** Model inputs and outputs for each query
- **Error logs:** Failures or exceptions during inference or data processing
- **Usage logs:** Frequency and context of system access

Emerging AI governance platforms are beginning to offer AI-specific record keeping and compliance features, including:

- Model version control and lineage

- Data provenance and lineage
- Risk register and compliance integration
- Post-deployment monitoring dashboards
- Incident response workflows

Example platforms include Fiddler AI, Arthur.ai, Credo AI, and Microsoft Responsible AI Dashboard.



Blockchain-inspired solutions, such as append-only logs or cryptographic hashing, are gaining attention for ensuring immutability, traceability, and *nonrepudiation* in AI audit trails.

Legal and Regulatory Considerations

Recordkeeping related to AI system development and operations is often driven by laws, regulations, and terms in legal agreements. Example frameworks include:

- **EU AI Act:** Requires high-risk systems to maintain detailed technical documentation, logs, and monitoring records
- **Brazil AI Act:** Requires high-risk systems to document risk classification, impact assessments, operational logs, governance proceedings, incidents, testing, and human interventions (as of the writing of this book, this law has not been enacted)
- **GDPR:** Requires organizations to provide meaningful information about automated decisions and maintain accountability, which in practice often necessitates logging
- **NIST AI RMF:** Encourages documentation at all stages for trustworthiness and auditability
- **ISO/IEC 42001:** Emphasizes logging, incident reporting, and risk documentation for AI management systems

Laws vary considerably by jurisdiction, including:

- In **Canada**, the Personal Information Protection and Electronic Documents Act (PIPEDA) requires accountability for automated decision-making outcomes, which extends to AI systems
- In **Brazil**, the Lei Geral de Proteção de Dados (LGPD) intersects with AI auditability in data processing decisions
- In **China**, AI regulations mandate algorithmic transparency and regular audits of system behavior



Organizations need to obtain or maintain a regulatory map that cross-references system types, jurisdictions, and applicable laws. This helps prioritize which records must be maintained and for how long. Because this effort is often labor-intensive, organizations usually turn to subscription services for this information.

Integration with Broader Organizational Practices

Business records related to AI systems should not be isolated, but instead, embedded into the organization's other operational, compliance, and risk management workflows and repositories.

AI record keeping should align with existing enterprise information governance policies, including document classification, enterprise content management (ECM) systems, and audit trails.

Internal and external audits should be performed to ensure that:

- AI-related records are complete and accurate, including governance proceedings, reviews, and decisions
- Role-based access controls are enforced
- Incident responses are timely and well-documented

Compliance should not be the only impetus for record keeping. Instead, record keeping helps an organization in several ways, including:

- **Incident after-action reviews and RCA reports:** These provide valuable insights for preventing future incidents.
- **Risk trend analyses:** These help identify emerging patterns and problems.
- **Issue recurrences:** These signal areas needing process improvement.

For example, an AI-powered financial fraud detection system experiences a spike in false positives. Root cause analysis reveals a data drift event, prompting the development of improved retraining protocols and enhanced monitoring to detect similar future patterns.

External Communications

External communication is a crucial aspect of AI governance, as organizations must occasionally communicate with regulators, customers, and other stakeholders. Effective and clear communications help organizations maintain trust, comply with regulatory expectations, and inform the public, users, and partners about various matters. This section describes the development and management of structured external communication plans that promote transparency, accountability, and stakeholder engagement across the AI system lifecycle.

Establishing a Communication Governance Structure

External communications deliver the tone, voice, and content of messages to various parties, including the public. This requires planning, discipline, and maturity to avoid issuing ad hoc press releases or reactive statements. A communications governance structure ensures that messages are consistent, authorized, and aligned with broader compliance and risk management policies.

Organizations should designate key personnel to oversee AI-related external communications, and roles may include:

- **Public information officer (PIO):** Official representative for the organization in public forums
- **Social media lead:** Manages social media presence and delivers communications via social media
- **AI communications lead:** Owns the AI-specific messaging strategy
- **Legal counsel:** Reviews communications for compliance, intellectual property considerations, and liability risk
- **Product manager or engineer:** Provides technical explanations

Organizations' external communications should be governed by policies that integrate with enterprise risk management, data governance, privacy, incident response protocols, and regulatory reporting obligations. These policies help define when, how, and by whom external communications should occur, especially during crisis scenarios involving the AI system. In most or all cases, communications to any external party should be formally approved by a designated leader.

For example, a healthcare startup deployed a diagnostic AI system and assigned a medical officer to explain its capabilities and limitations, ensuring that users receive medically sound guidance rather than marketing-speak.

Communicating AI System Capabilities and Limitations

External stakeholders, including regulators, users, partners, and the public, deserve clear, concise, and accessible information about their AI systems, how they work, and what they should (and should not) be relied upon to do. Disclosures should cover the following:

- **Purpose and use cases:** What the system is designed to do
- **Limitations and known risks:** Accuracy boundaries, potential biases, data dependencies, or reliability caveats
- **Human involvement:** How and when human oversight is used
- **Data sources and model behavior:** General descriptions of training data and how decisions are made
- **Inquiries and data subject requests:** How to get assistance, appeal a decision, and correct errors

Disclosures take on many forms and can appear in:

- Product documentation and user manuals
- Legal agreements and terms of use statements
- Privacy notices and terms of use

- Web portals or model cards summarizing technical characteristics
- Public FAQs or explainers written in plain language

For example, an organization's AI-powered voice assistant product publishes a transparency page that describes the interactions that are logged, how the AI model interprets commands, and which commands trigger data retention.

The Dangers of Overclaiming AI Capabilities

Overstating an AI system's capabilities can lead to regulatory scrutiny, loss of consumer trust, or even lawsuits. Communication teams must balance enthusiasm with technical and legal accuracy.

Regulatory and Legal Communication Obligations

Some jurisdictions and regulatory frameworks require organizations to proactively or reactively disclose information about their AI systems to authorities or the public.

In some cases, such as under the EU AI Act, high-risk AI systems must be registered in a public database, which includes documentation on their intended purpose, technical specifications, risk mitigation measures, and post-market monitoring plans.

Incidents, such as biased outcomes, performance degradation, or adversarial attacks, may require external communication as required by:

- **GDPR (Article 34):** Notifying data subjects of high-risk breaches
- **EU AI Act:** Informing regulators of serious incidents and corrective actions
- **Consumer protection laws:** Alerting users to breaches, unauthorized disclosures, product defects, or misleading results

Mature organizations develop preapproved communication templates that comply with regulatory expectations while maintaining organizational tone and branding. These can streamline response during time-sensitive disclosures.

Managing Public and Media Relations

Public perception plays a significant role in shaping organizations' reputations, including how AI deployments are perceived. Even technically sound systems can lose public trust if communications are mishandled or people are confused. Effective media relations should be proactive, transparent, and informed.

In terms of their relationship with customers and the public, organizations should lean toward proactive communication, including:

- Press releases announcing new AI deployments
- Public blog posts explaining design choices

- Conference presentations or white papers
- Engaging external advisory panels or ethics boards
- Webinars and seminars to explain AI deployments

Proactive outreach helps set expectations, build confidence, and demonstrate good faith.

Even the most organized companies have occasional issues and events. When system failures or public controversies occur, a crisis communication plan should direct the PIO and others on how to communicate effectively with the right messages. Key components include:

- **Crisis response team:** Usually led by the PIO, this is a designated group authorized to speak to authorities, the press, and the public.
- **Message framing:** Centers transparency, responsibility, and action.
- **Media readiness:** Includes prewritten press releases and answers to anticipated questions.

For example, after a facial recognition system misidentified persons of color, a tech company immediately paused deployments, conducted an audit, published the audit findings, and engaged with civil rights groups in open forums to discuss response and remediation. (Note that several U.S. cities and other jurisdictions have banned or restricted the use of facial recognition.)

User-facing Communications and Transparency

The most impactful external audiences are customers and end users. Organizations must ensure that all user-facing communications promote informed, responsible use, expectations, and respectful engagement with AI tools.

The systems and sites that users and customers use should include clear labeling that indicates:

- That a system is AI-driven (for example, “AI-generated”)
- The identity of the deployer
- The nature of the system’s outputs (for example, suggestions vs. decisions)
- Any opt-out or appeal mechanisms

Examples of required labeling and disclosures include the EU AI Act’s mandated labeling of deepfakes and synthetic content, and California’s Bolstering Online Transparency (BOT) Act, which requires bots to self-identify during online interactions for commercial purposes.

Users should be able to intuitively ask questions and voice concerns about AI behavior. Users should be met with accessible help centers, contact options for escalation, and receive timely and respectful responses. Organizations should record and analyze this feedback as part of their post-market monitoring efforts.

Partner and Vendor Communications

Most organizations outsource an appreciable portion of their IT infrastructure, meaning AI governance does not stop at the organization's boundary. Vendors, developers, and third-party integrators must also be included in external communication plans.

When working with third parties, organizations should embed communication plans into:

- **Vendor agreements:** Requiring partners to report outages, incidents, and changes in system performance
- **Service level agreements (SLAs):** Defining timelines and scope for updates, disclosures, and issue resolution
- **Joint communication protocols:** Clarifying who speaks on behalf of the system in case of shared ownership

In open-source and collaborative contexts, communication with partners and vendors should support:

- Transparent publication of vulnerabilities, breaches, defects, and other issues
- Roadmaps for system and model updates
- Feedback channels from contributors

For example, an AI developer working with an academic institution posted all model updates and limitations on GitHub and notified contributors through mailing lists and community calls.

Communication Planning Across the AI Lifecycle

External communications should not be treated as one-time events but as part of an overall process mapped across all of the AI lifecycle stages, including:

- **Design and development:** Research publications and data use notices
- **Validation and testing:** Accuracy disclosures and bias audit results
- **Deployment:** Purpose statement, user guidance, and stakeholder briefings
- **Operation:** Post-market updates, incident alerts, and FAQs
- **Retirement:** Sunset notices, archival disclosures, and user offboarding info



The communications discussed in this chapter should align with documentation, monitoring, and retirement strategies discussed elsewhere in this book. Communication is the connective tissue between internal governance and public and regulator accountability.

Measuring Communication Effectiveness

Mature organizations measure their processes and public metrics to make improvements to their operations. To ensure external communication efforts are effective, organizations should evaluate:

- **Clarity:** Whether users and stakeholders understand the intended messages
- **Reach:** Whether the right audiences are informed
- **Responsiveness:** Whether inquiries and issues are handled in a timely manner
- **Trust indicators:** Whether public trust improves or erodes over time

Ideas for external communications metrics include:

- User satisfaction surveys
- Media sentiment analysis
- Volume and resolution time of inquiries
- Stakeholder feedback during consultations
- Social media engagement analytics
- Traffic metrics to web pages containing AI-specific content



Organizations should conduct periodic communication audits to review the clarity, consistency, and legal adequacy of their communications. Update FAQs, disclosures, and public statements in response to system evolution and user feedback.

AI System Retirement

Like traditional information systems and applications, AI systems are not deployed indefinitely. Like any other enterprise technology, they must eventually be upgraded, replaced, decommissioned, or withdrawn due to obsolescence, performance degradation, legal compliance issues, or strategic business shifts. A well-governed AI lifecycle must include a clear policy and a repeatable process for retiring AI systems. This section guides organizations in planning for, implementing, and managing the controlled deactivation, withdrawal, or localization of AI systems.

Retirement should be a standard part of any organization's system development lifecycle (SDLC) process, rather than an activity to be invented on the fly. Just as a responsible software product manager plans for version upgrades and end-of-life support, AI developers should embed deactivation strategies into the architectural blueprint. Planning reduces panic-driven decisions when retirement becomes necessary.

Understanding the Need for AI System Retirement

Retirement is a critical governance function that ensures AI components do not outlive their usefulness or pose ongoing and increasing risk. Several factors may trigger the need to deactivate or localize a model. Common triggers for AI system deactivation include:

- **Regulatory changes:** New legal requirements may render continued use noncompliant, or the cost of becoming compliant may be prohibitive.
- **Performance degradation:** The AI system may not perform as well as newer systems or be experiencing model drift, obsolete data, or declining accuracy.
- **Business reprioritization:** Shifts in the organization's product strategy or service offerings may render the system unneeded.
- **Security or ethical concerns:** The emergence of known vulnerabilities or harmful impacts may prompt an organization to retire the AI system.
- **Third-party model withdrawal:** An external AI vendor may have deprecated its AI model or API.
- **Localization requirements:** The AI system may be unable to comply with regional obligations (for example, *data sovereignty*).

For example, an AI trained primarily on U.S. English was found to produce biased outputs in international deployments. The organization decided to localize and retrain separate models for European markets (in part to ease GDPR compliance efforts) and retired the global version in those jurisdictions.

Building an AI System Retirement Policy

Any system's retirement should be governed by policy. A system retirement policy provides a formal structure to guide when and how AI systems are sunset. The policy should align with the broader AI governance framework and be documented in the organization's AI lifecycle management standards.

Core elements of a retirement policy include:

- **Criteria for retirement:** Define the operational, financial, strategic, ethical, and legal thresholds that trigger consideration for retirement. These criteria may include sustained KPI failure, excessive sustainment costs, changes in business strategy, confirmed regulatory conflicts, or stakeholder demands.
- **Approval and escalation process:** Identify those responsible for initiating and approving retirement actions. Escalation should occur when ethical risks or customer harm are identified.
- **Retirement planning requirements:** Include mandatory steps such as stakeholder notification, system backup, data archiving, and user transition support.

- **Documentation and audit trail:** Require detailed records of the rationale, timing, methods, stakeholders, and proceedings involved in the retirement.
- **After-action review (AAR):** Include lessons-learned analysis to inform future development and deployment practices.

AI Retirement vs. IT Decommissioning

While similar in concept, AI retirement involves more than turning off a system. The retirement of an AI system may require untraining, an ethical review, the removal of derivative models, and public communication, especially when mandated by regulations or when high-impact decisions were previously automated.

Planning for Controlled Deactivation

The controlled deactivation of an AI system should be performed as a part of a repeatable, written process, to reduce disruption, mitigate risks, and preserve vital records. The steps for a controlled deactivation include:

1. **Stakeholder mapping:** Identify stakeholders, including internal users, external customers, regulators, and partners. Determine their needs during the transition.
2. **Usage analysis:** Map dependencies to better understand where the system is embedded and which downstream tools or services rely on it (if the AI system was the subject of a business impact assessment—an activity related to disaster recovery planning—these dependencies may already be known).
3. **Fallback or transition planning:** Determine whether another model, manual process, or third-party tool will assume all or part of the functionality performed by the AI system being retired.
4. **User notification and education:** Communicate clearly with affected parties and provide guidance on changes, new procedures, the reasons for the retirement, and how they can ask additional questions.
5. **Final audit snapshot:** Capture the AI model's final state, including inputs, weights, performance metrics, and decision logs, to ensure accountability and transparency.
6. **Execution of shutdown:** Disable APIs, terminate model endpoints, and remove integration points. Update security tools such as network access controls and firewalls as needed.
7. **Post-deactivation monitoring:** Monitor system health and user behavior to ensure a successful transition.

Often, when a system is being deactivated, some or all of the automation is taken over by a replacement system. The overall activity is known as a *migration*, which often

involves migrating data from the old system to the new system. This also often requires a considerable amount of testing to ensure that data was migrated correctly and that the replacement system works correctly. Often, the system being replaced is put into a "read-only" or "reporting" mode if the organization is required to make information available for historical purposes for a specified period.

AI system retirement may introduce operational and governance challenges, including:

- **Business continuity:** Ensure that critical processes formerly powered by the AI system continue to be supported and remain operational.
- **Data integrity:** Prevent loss of audit trails or traceability.
- **Access revocation:** Deprovision API keys, credentials, and service accounts.
- **Residual risk management:** Monitor for unintended consequences of the shutdown, such as degraded user experience or manual process errors.

For example, a bank retires an AI-powered loan risk scoring model after discovering a shift in applicant demographics that the model no longer represents. Customers are notified two months in advance, and human underwriters temporarily handle applications while a replacement model is acquired, trained, and validated.

Localization and Regional Retirement Strategies

Sometimes, models are not retired globally, but instead are deactivated or *localized* for specific regions, due to differing compliance environments or cultural contexts. Sometimes it's just too difficult to get a single AI system to comply with multiple sets of regulations and customer expectations. Localization drivers may include:

- **Data sovereignty laws:** Mandates requiring training or inference to occur within national borders (for example, China or Russia).
- **Ethical and societal norms:** Regional sensitivities may necessitate adjustments in content filtering or recommendation practices.
- **Regulatory constraints:** Laws such as the EU AI Act may restrict or ban the use of specific AI system use cases deemed unacceptable or high-risk.

Approaches for performing localization include:

- **Separate model instances:** Deploy country-specific models trained on localized data.
- **Federated learning:** Use decentralized training to maintain local *data residency*.
- **Modular architecture:** Enable selective enablement of regional features or content.

For example, a global media company runs a content moderation AI across multiple continents. Due to regional speech laws, the company maintains different rule sets and retrained models in France, Germany, and the United States, retiring the global version for jurisdiction-specific deployments.

Data and Model Artifacts: Archiving and Disposal

When a system is retired, the organization must determine what information should be retained, archived, or deleted. Regulations, legal agreements, or internal policies govern these decisions.

Information that an organization may choose to archive includes:

- Final model versions and training data lineage
- Performance logs and error reports
- Incident reports and mitigation records
- Decision outputs, especially for high-impact use cases (for example, healthcare and finance)
- User and stakeholder communication records
- Design documents and other artifacts that describe the AI system
- Any other information specified by regulations as being required to retain

Information that an organization might choose to dispose of includes:

- Redundant or obsolete training data
- Temporary model artifacts (for example, intermediate checkpoint files)
- Any personally identifiable information not required for compliance

Organizations sometimes use military-grade data erasure tools to permanently discard sensitive information, ensuring that it cannot be recovered.

Organizations using an AI system hosted by a cloud service provider will need to request that the cloud service provider destroy the information, possibly including a certificate of destruction that the organization may require for audits. Similarly, an organization may need to seek an alternative hosted AI solution when its current hosted AI provider announces its own system retirement.



Organizations can conduct a Data Protection Impact Assessment (DPIA) before deletion to determine whether data or records are subject to retention requirements under the GDPR, CPRA, or other relevant frameworks.

Communication and Transparency in AI Retirement

Often, organizations need to inform external parties of their intention to retire an AI system. Concise communication is crucial to ensure that all parties are notified of decommissioning decisions and their associated impacts.

Examples of these communications include:

- **User notification:** In-product alerts, email campaigns, or support updates
- **Organization notification:** Official communications to business customers and partners may be required by legal agreements, including applicable terms and conditions, such as the time of notice

- **Public disclosures:** Press releases, social media announcements, and statements on the organization's website or transparency portal
- **Regulatory reporting:** Notification to authorities when required by law (for example, EU AI Act or consumer protection laws)
- **Internal briefings:** Notifications to legal, compliance, support, and sales teams

Not everyone will be pleased with an announcement of an upcoming system retirement. Organizations should be prepared for:

- Pushback from users who depend on the system
- Misunderstanding of the reason for retirement
- Questions regarding a migration to other service providers
- Regulatory scrutiny of timing and rationale

Early involvement of the crisis communications team and the PIO will help the organization anticipate and respond to these inquiries. Having a well-prepared crisis communication playbook can mitigate reputational risk.

For example, a tech firm retires a generative AI image tool due to concerns about misuse. The firm issues a press release, updates its responsible AI website, and offers user refunds or migration paths to alternative tools.

AI system retirement should be handled in a manner consistent with the laws and contractual obligations under which they were deployed. Applicable legal obligations include:

- **EU AI Act:** Requires ongoing monitoring and withdrawal of noncompliant high-risk systems
- **GDPR:** Requires data minimization and may necessitate the erasure of personal data
- **Consumer protection laws:** Can trigger obligations to refund or compensate users affected by the removal of AI-driven features
- **Contractual terms:** May require notice periods or availability guarantees

Litigation Risk During Retirement

Improper or abrupt retirement of an AI system may expose organizations to legal claims, particularly if customers relied on system outputs, or if discriminatory impacts are discovered after shutdown (another reason for archiving system information).

Lessons Learned and Organizational Feedback Loops

AI system retirement should serve not only as an end-of-life event but also as a source of learning and improvement for future deployments. Mature organizations seize opportunities to reflect, learn, and improve through after-action reviews (AARs).

Essential post-retirement review practices include:

- **Root cause analysis:** Why the system needed to be retired
- **Impact assessment:** The effect on users, operations, and business reputation
- **Post-retirement ethical review:** To ensure that a retired AI model is not repurposed in harmful ways by others
- **Policy evaluation:** Whether the retirement process aligned with the formal policy
- **Knowledge capture:** What information should be documented for future teams



Organizations should incorporate lessons from system retirement into training programs for data scientists, engineers, product managers, governance professionals, and others. Doing so improves institutional maturity and helps avoid repeat mistakes.

Summary

Ongoing AI operations require discipline and a structured approach to managing records, communicating with stakeholders, and responsibly decommissioning systems. As AI continues to mature and expand across industries, governance doesn't end at deployment; it extends throughout the system's operational life and beyond, including its retirement.

Record keeping is a critical part of AI governance. By documenting incidents, issues, risks, and post-market monitoring activities, organizations create an audit trail that supports accountability, learning, and regulatory compliance. This process involves more than reactive reporting: it includes proactive tracking of minor issues, structured risk registers, and monitoring strategies to detect model drift or unintended harms in the field. Tools such as incident logs, issue trackers, and performance dashboards enable teams to respond to real-world developments and continuously improve their AI systems.

External communication is also essential for AI governance. Organizations must maintain transparency with users, regulators, partners, and the public. This includes clear disclosures about what AI systems do, how they work, and where their limitations lie. Whether through product documentation, public model cards, or incident notifications, external communications should be accurate, accessible, and aligned with legal requirements. Regulatory expectations, including those in the EU AI Act or GDPR, often mandate these communications, while public trust depends on them. A proactive media strategy, clearly

defined roles and responsibilities, and a crisis communication plan further ensure that organizations can respond effectively and appropriately in any situation.

Eventually, every AI system reaches a point where continued operation is no longer viable. Retirement planning is a critical part of the AI lifecycle, retirement by design, enabling responsible deactivation or localization of models in response to regulatory shifts, performance degradation, and ethical concerns. Organizations must implement structured retirement policies and practices that define when and how systems should be shut down, how users and regulators should be notified, and what data must be archived or deleted. Controlled deactivation, fallback strategies, regional adjustments, and lessons-learned reviews help ensure that system retirement is orderly and low risk.

Exam Essentials

AI incidents, issues, and risks must be documented systematically to support governance, accountability, and regulatory compliance. Effective AI governance relies on maintaining accurate records of system behavior, particularly when errors, failures, or emerging risks occur. These records inform audits, help identify patterns, and demonstrate due diligence to regulators and stakeholders.

External communication plans must provide accurate, accessible, and timely information about AI system capabilities, limitations, and incidents. Communication is a core part of building public trust and meeting legal obligations. Whether disclosing an AI model's intended use or responding to a critical incident, organizations must be prepared with clear policies, messaging templates, designated spokespeople, and approved messaging channels.

AI systems should be retired through a structured, policy-driven process that minimizes disruption and preserves essential records. Decommissioning an AI system involves more than turning off the model. A comprehensive retirement process includes planning, stakeholder notification, data archiving, and risk mitigation, especially in high-risk and regulated domains.

Localization may be necessary to comply with regional regulations, cultural norms, or data residency requirements. AI systems deployed globally may need to be deactivated, modified, or retrained for use in specific jurisdictions. Localization ensures compliance with regional laws, such as the EU AI Act, and helps avoid ethical or operational problems.

Retirement planning should be integrated into the AI lifecycle from the start. Designing systems with the end in mind enables smoother transitions, reduces last-minute risks, and aligns with principles of responsible AI. Planning ahead makes retirement a routine governance activity rather than a disruptive event.

Review Questions

The following questions are designed to test your understanding of this chapter's material. For more information on how to obtain additional questions, see this book's Introduction. You can find the answers in Appendix A.

1. Which of the following is the primary purpose of documenting AI system incidents?
   - **A.** To prepare for marketing campaigns
   - **B.** To increase model accuracy during training
   - **C.** To support audits, accountability, and compliance
   - **D.** To automate retirement decisions
2. What is a key feature of a post-market monitoring plan for AI systems?
   - **A.** Ongoing evaluation of system performance and risks
   - **B.** Dynamic pricing optimization
   - **C.** Real-time stock trading integration
   - **D.** Legal review of all user interactions
3. In an external communication strategy, what should be included when describing an AI system?
   - **A.** Source code and API credentials
   - **B.** Marketing slogans and social media plans
   - **C.** User passwords and login behavior
   - **D.** System purpose, limitations, and human involvement
4. What is the primary function of an AI system retirement policy?
   - **A.** To suppress complaints from end users
   - **B.** To extend AI model use indefinitely
   - **C.** To guide the responsible deactivation of systems
   - **D.** To prevent transparency in decommissioning
5. Which of the following most likely triggers the localization of an AI system?
   - **A.** High user satisfaction in one market
   - **B.** Cultural or legal differences across regions
   - **C.** Stable training data distributions
   - **D.** Shared model weights across all use cases

6. What distinguishes an "issue" from an "incident" in AI operations?
   - **A.** Incidents only occur in training environments
   - **B.** Issues are generally lower in severity and may not require immediate response
   - **C.** Incidents are optional to report; issues are mandatory
   - **D.** Issues always involve user data loss

7. Why should deactivated AI models be archived rather than deleted immediately?
   - **A.** To support legal, audit, and accountability requirements
   - **B.** To enable unauthorized reuse in other products
   - **C.** To avoid storage costs
   - **D.** To meet SEO guidelines

8. What is a key consideration when preparing for AI system retirement?
   - **A.** Increasing the model's prediction speed
   - **B.** Releasing source code into the public domain
   - **C.** Assessing dependencies and planning fallback options
   - **D.** Disabling user feedback channels

9. Under which regulation is public registration of high-risk AI systems required?
   - **A.** GDPR
   - **B.** California Consumer Privacy Act
   - **C.** EU AI Act
   - **D.** ISO/IEC 27001

10. What is a best practice for communicating AI limitations to users?
    - **A.** Use clear, plain language with real-world examples
    - **B.** Use vague language to avoid liability
    - **C.** Provide highly technical white papers only
    - **D.** Avoid mentioning limitations to prevent user doubt



Appendix

Answers to the Review Questions

Chapter 1: AI and AI Governance

1. C. Narrow AI includes AI systems that perform highly focused tasks. Other examples are spam filters, license plate recognition systems, and Apple Face ID.
2. A. An AI agent is a system that receives input from systems and sensors and takes action to achieve specific goals.
3. D. According to the GDPR, organizations must obtain explicit consent from customers before their information can be used by the AI system for this (or any other) purpose.
4. A. Training an AI system with labeled data is known as supervised learning.
5. D. The type of harm that the bank must watch for is bias, where individuals of certain demographics may be denied loans or favorable loan terms.
6. C. The bank should implement a human in the loop to review loan applications and the AI system's recommendations to ensure that the AI system is producing correct outputs.
7. B. A biased AI algorithm in a social media platform could sway public opinion.
8. C. Because the organization uses a neural network, known for its opacity and lack of explainability, the organization will have difficulty explaining why the AI system rejected the application.
9. B. The safety of passengers and others should be the highest priority for an autonomous vehicle.
10. D. The design and operation of an AI system employing facial recognition to detect returning customers in a retail store should prioritize privacy above other considerations.

Chapter 2: Organizational Readiness

1. C. Developers build and train models, which include curating datasets and running fairness evaluations.
2. C. Startups typically employ simplified governance structures with a few cross-functional leads, due to their limited staff and resources.
3. D. Cross-functional collaboration means multiple departments sharing responsibility for governance decisions.
4. C. Training must reflect how stakeholders utilize AI to be relevant and effective.
5. C. Users are closest to the operational impact of AI and are often the first to detect unintended outcomes.

6. C. Regulated industries require strict controls, impact assessments, and compliance documentation.
7. B. “Accountable” means the person who ultimately owns the task or process and ensures it’s completed correctly.
8. C. Deployers operationalize models, ensuring they function reliably and are monitored for risk.
9. C. Collaborative reviews ensure that ethical, legal, business, and technical risks are considered together.
10. C. Real-world examples help stakeholders understand the consequences of poor governance and how to respond.

Chapter 3: Updating Policies for AI

1. C. AI governance must be continuous and span the entire lifecycle, from use case assessment to incident management. Limiting governance to deployment or specific data types introduces risk.
2. C. Even anonymized data can be reverse-engineered or inferred from model behavior. This is a known risk in AI privacy and requires stronger safeguards.
3. B. Governance policies should demand transparency about the model and data. Documentation helps evaluate fairness, legality, and risk exposure.
4. B. Monitoring is a post-deployment function that ensures models continue to perform ethically and effectively under real-world conditions, for example, to detect emerging bias and drift issues.
5. C. Contracts should require transparency and auditability, especially for high-risk AI applications. This includes data origin, fairness testing, and monitoring results.
6. A. AI models may “remember” or leak training data. Traditional policies rarely account for the retention risks related to what a model has learned or memorized.
7. C. Risk-based tiering ensures that higher-risk vendors (for example, those processing sensitive data or supporting critical functions) are subject to more thorough review and oversight.
8. B. A use case registry enables organizations to track and enforce approved purposes for data use, preventing unauthorized repurposing.
9. B. Fairness testing against different groups helps identify and reduce disparate impact, which is essential for ethical AI development.
10. B. Model inversion is a privacy threat in which attackers use access to a model to reconstruct inputs from its outputs, posing serious risks to data confidentiality.

Chapter 4: Privacy and Data Protection Law

1. A. Article 9 of the GDPR generally prohibits the processing of special category data unless an exception applies; the most common exception is explicit consent, which must be freely given, specific, and informed.
2. C. Purpose limitation requires that personal data be used only for the purpose for which it was collected, and not repurposed unless the new use is compatible or consent is reobtained.
3. D. DPIAs are mandated under Article 35 of the GDPR when processing is likely to result in a high risk to individuals, including when AI systems process sensitive data.
4. C. A data controller is responsible for deciding why and how personal data is processed and carries the accountability for compliance with data protection laws.
5. B. Facial recognition is a form of biometric data used for identification, which qualifies as special category data and requires strict controls under laws such as the GDPR.
6. C. When AI systems infer special category attributes (for example, religion, sexual orientation), this processing is still subject to the same legal protections and may require consent or an Article 9 exception.
7. B. Privacy by design calls for integrating privacy considerations at every stage of development, from problem definition to deployment and monitoring.
8. B. Data minimization refers to collecting and using only the information necessary for a specific task. Anonymized, purpose-limited data use is a textbook example.
9. A. Controllers must have contracts with processors that define the scope, safeguards, instructions, and responsibilities as outlined in Article 28 of the GDPR.
10. C. Article 22 of the GDPR requires that individuals subject to significant automated decisions have the right to obtain human intervention, express their point of view, and contest the outcome.

Chapter 5: Sectoral and Civil Laws

1. C. Using copyrighted images without permission may infringe on reproduction rights and does not automatically qualify as fair use.
2. D. Race is explicitly protected under Title VII of the Civil Rights Act. AI tools must be audited for disparate impact across such categories.

3. C. The FTC has warned that failing to disclose AI interactions may mislead reasonable consumers and violate deceptive practice standards.
4. C. Design defects refer to flaws in the product's intended specifications, regardless of how well it was manufactured.
5. B. ECOA bars discriminatory loan and credit practices and applies to both human and algorithmic decision-making processes.
6. C. Providing health advice without disclaimers or licenses can mislead consumers and trigger enforcement under UDAP (unfair, deceptive, or abusive practices) laws.
7. A. Complexity is not a valid defense. Manufacturers are still responsible for designing reasonably safe products.
8. C. The EEOC has explicitly stated that employers are responsible for ensuring that AI hiring tools do not discriminate.
9. A. Proxy variables (like postal code) can correlate with race or income, potentially causing discriminatory outcomes.
10. B. AI performance claims must be supported by evidence and testing—just like any other advertising or product representation.

Chapter 6: The EU AI Act

1. C. The EU AI Act bans real-time biometric identification in public spaces unless specific public safety exceptions apply. This practice is considered unacceptably risky.
2. C. High-risk AI systems must include detailed technical documentation that supports transparency, traceability, and accountability.
3. D. A provider develops and markets the AI system under its own name and is responsible for ensuring compliance.
4. B. Deployers must ensure human oversight during operation to support safety and mitigate unintended consequences.
5. B. Minimal-risk AI systems, like grammar checkers and spam filters, are largely unregulated under the EU AI Act.
6. C. GPAI systems deemed to pose systemic risk, due to scale or influence, face stricter obligations such as external audits and real-time monitoring.
7. B. Importers are responsible for verifying that the AI system complies with EU conformity standards, including CE marking.

8. D. The most serious violations, such as using prohibited AI, can incur penalties of up to €35 million or 7 percent of global revenue, whichever is greater.

9. C. Distributors must act if they become aware of non-compliance; failing to do so triggers liability.

10. C. A model card is a transparency tool that describes an AI model's design, training data, capabilities, and risks to inform responsible downstream use.

Chapter 7: AI Standards and Frameworks

1. C. The OECD AI Principles offer non-binding ethical guidelines designed to foster responsible, human-centered AI across member countries. They focus on values like transparency, fairness, and accountability.

2. C. The Map function focuses on understanding the AI system's design, context of use, stakeholder impacts, and associated risks.

3. B. NIST ARIA complements the AI RMF by providing practical tools, metrics, and methodologies to test and evaluate AI systems for safety and trustworthiness.

4. C. ISO/IEC 42001 is the first AI-specific management system standard that organizations can certify against, similar to ISO/IEC 27001 for information security management.

5. C. ISO/IEC 22989 offers foundational definitions for AI systems, capabilities, and lifecycle stages, facilitating consistency across industries.

6. A. The G7 Hiroshima Process includes a voluntary code of conduct aimed at developers of AI and foundation models to promote responsible innovation.

7. B. The IEEE 7000 series supports embedding ethical considerations directly into AI system design and engineering processes.

8. B. NIST ARIA supports measurement and assurance techniques for trustworthiness traits, including interpretability, robustness, and fairness.

9. A. ISO/IEC 23894 provides detailed guidance on AI risk management processes, while ISO/IEC 42001 defines a complete management system for AI.

10. C. UNESCO's recommendation focuses on equity, sustainability, and cultural diversity, making it a practical and values-driven resource in countries without detailed AI regulations.

Chapter 8: AI System Design

1. C. Lifecycle documentation supports traceability, governance, and audit readiness, especially when evaluating compliance, fairness, or model behavior over time.
2. A. AI systems should be aligned with business objectives to ensure they serve real-world needs and deliver measurable value.
3. D. Impact assessments help surface potential harms, compliance gaps, and ethical issues early in the development process. The EU AI Act requires impact assessments for high-risk AI systems.
4. C. HITL ensures that humans can review, override, or validate AI decisions in high-stakes scenarios, increasing organizational accountability and safety.
5. A. Developing proprietary models gives organizations more control, but also increases exposure to legal, regulatory, and ethical responsibilities.
6. B. A probability/severity matrix helps quantify and prioritize risks based on their likelihood and potential impact.
7. C. Feature importance metrics and clear, contextual explanations enhance understanding for users and stakeholders.
8. D. Stakeholder mapping reveals who is impacted, what their concerns are, and how risks differ across roles and perspectives.
9. B. Shadow mode testing simulates real-world conditions without affecting users, enabling risk detection and performance validation. Shadow mode testing is similar to parallel testing in IT disaster recovery planning.
10. C. A company that develops and deploys its own AI system is legally classified as a "provider," carrying additional obligations under the EU AI Act.

Chapter 9: Data Governance and Model Training

1. C. Data governance ensures that data is lawfully collected, accurate, sufficient, and suitable for its intended AI use, forming the foundation of trustworthy system development.
2. D. Data provenance refers to the origin of the data, its collection method, and the conditions under which it was collected, which is essential for legal, ethical, and contextual evaluation.

3. D. The F1 score balances precision and recall, making it more reliable than raw accuracy when class distributions are uneven, such as in fraud or medical detection.

4. B. Neglecting subgroup validation can mask performance disparities across demographic groups, introducing or perpetuating algorithmic bias.

5. A. Data lineage tracks the lifecycle and processing of data, while provenance identifies its origin, licensing, and collection context. Both are necessary for traceability.

6. C. Adversarial examples are intentionally designed inputs that cause AI models to produce incorrect or unexpected outputs, often without obvious errors to humans.

7. C. Counterfactual fairness assesses whether a model's decision would change if a sensitive attribute (for example, gender, race, or political affiliation) were altered, while keeping all other factors constant.

8. C. Integration testing validates the end-to-end functionality of the system, ensuring that data flows, API calls, and downstream systems properly interact with the AI model.

9. C. Monitoring loss curves and training metrics during the training phase helps detect problems such as non-convergence, vanishing gradients, or early overfitting.

10. D. Model cards summarize key information about a model's training data, performance, limitations, and intended use, thereby supporting transparency and fostering stakeholder trust.

Chapter 10: Deployment and Monitoring

1. C. Model cards are standardized documentation tools that provide an explanation of an AI model's intended use, training data, evaluation results, ethical considerations, and limitations. They support transparency and accountability.

2. D. Edge deployment reduces latency and avoids transmitting sensitive data to centralized servers, making it ideal for real-time, privacy-focused applications.

3. A. Version control enables traceability and auditability by tracking model changes and ensuring consistency across development and deployment environments.

4. B. Concept drift refers to changes in the underlying patterns between input data and expected outputs, often due to real-world shifts.

5. C. RAG systems supplement model outputs by retrieving recent, relevant data from an external knowledge base, reducing hallucinations and improving accuracy.

6. B. Data governance teams are responsible for ensuring data usage aligns with policies, including tracking lineage, enforcing access controls, and maintaining compliance.

7. C. Rollback plans enable rapid recovery from harmful or faulty AI behaviors by reverting to a safe model version or configuration.
8. D. Fairness issues often appear as unequal treatment across groups. Monitoring these disparities is a key ethical and governance practice.
9. C. Shadow models run in parallel with production models to detect drift or performance degradation by comparing outputs over time.
10. D. Frequent retraining helps maintain relevance but incurs costs. High-risk, fast-changing contexts require more frequent updates to prevent harm and maintain accuracy.

Chapter 11: AI Risk Management and Assurance

1. A. Stakeholder impact mapping helps forecast downstream harm by identifying who may be affected by unintended consequences of the AI system.
2. B. Red teaming focuses on simulating attacks or misuse scenarios to uncover vulnerabilities that typical testing might miss.
3. C. Data usage rights define how vendor systems can access, use, or retain data—impacting privacy compliance and ownership.
4. D. Model drift is typically detected by monitoring statistical changes in input/output distributions over time.
5. B. In safety-critical contexts, containment is essential. Disabling the model limits exposure while the investigation is ongoing.
6. C. A shadow deployment enables the testing of a new model in production without affecting users, thereby supporting a safe evaluation.
7. A. The EU AI Act mandates post-market monitoring of high-risk AI systems, including documenting incidents and developing incident response plans.
8. C. Threat modeling is a security-focused approach, while performance audits verify system behavior against its intended outcomes.
9. B. Prompt injection manipulates the model's behavior, potentially causing it to generate outputs outside its safety constraints.
10. C. Linking incidents to the risk register ensures new or recurring risks are formally tracked and mitigated in governance processes.

Chapter 12: Ongoing AI Operations

1. C. Documenting incidents supports transparency, regulatory compliance, and audit readiness by providing a detailed record of deviations from expected behavior.
2. A. Post-market monitoring ensures that deployed AI systems are continuously evaluated for accuracy, fairness, and unintended harms in real-world use.
3. D. Effective communication includes disclosing what the system does, its limitations, and the extent of human oversight involved.
4. C. A retirement policy ensures systems are deactivated in a controlled, risk-aware manner, preserving data integrity and compliance.
5. B. Legal mandates (for example, data sovereignty) and cultural norms often require that AI systems be localized or redeployed regionally.
6. B. Issues are often minor problems or observations, whereas incidents typically involve significant failures that require prompt action.
7. A. Archiving models allows organizations to retain historical performance data and evidence of compliance with governance obligations.
8. C. Understanding how other systems rely on the AI model enables a smoother transition and prevents service disruptions.
9. C. The EU AI Act requires that high-risk AI systems be registered, accompanied by technical documentation and risk mitigation plans.
10. A. Using plain language and relatable examples helps users understand system constraints, promoting informed and responsible use.

Index

Note: Page numbers in *italics* and **bold** refers to figures and tables respectively.

A

accountability, 22, 123–125
adversarial testing, 79
AI governance, 30, **37**
  assigning responsibilities across lifecycle, 35–36, **36**
  cross-functional collaboration in, 37–43
  developers, deployers, and users in, 59–66, **61–62**
  implementation tools and techniques, 36
  roles and responsibilities for stakeholders, 30–37
  stakeholder categories and governance roles, 31–35
  strategic importance of defined roles, 31
  tailoring to organizational context, 49–59
  training and awareness program, 43–49, **45**
AI impact assessments (AIIAs), 335
AI Risk Management Framework (AI RMF), 224–235, *225*
  vs. NIST Cybersecurity Framework, 234–235, **235**
AI system bill of materials (AI-SBOMs), 186
aleatoric uncertainty, 230
Alexa, virtual assistant, 4, 172–173
algorithmic advertising and targeting, 165–166
algorithmic bias as legal risk, 161
anonymization techniques, 85
anti-discrimination in algorithmic systems, 160
Artificial General Intelligence (AGI), 5
artificial intelligence (AI), 4, 102
  agent, 5
  AI-related standards, 246–250
  AI-specific contractual clauses, 92
  AI-specific due diligence, 91
  AI-specific privacy threats, 84–85
  autonomy, 14
bias, 8, 11, 44, 62, 75, 76, 78, 92, 103, 122, 132, 161, 167
capabilities, 5
characteristics of, 12
choice and consent, 104–105
complexity, 13
data controller obligations in, 114–125
data dependency, 15–16
data minimization and privacy, 107
deception and manipulation, 173–174
drift, 33, 64, 80, 82, 122, 177, 181, 286, 323, 341
effective AI privacy notice, 103
explainability, 14, 21
fairness, 18–19, 168–169
functionalities, 5–6
global impact of, 15
hallucination, 82, 115, 341
harm, 15
intellectual property, 16
misinformation in health, finance, and safety, 175–176
notice, 102–103
opacity, 13–14
playbooks, 82
practical implications and governance, 111–114
principles of, 17–23
privacy, 16
probabilistic *vs*. deterministic outputs, 17
providers, 211–212
purpose limitation in, 105–106
risk management, 75–76
risks and harms posed by, 7–12

artificial intelligence (AI) (*Cont.*)
sensitive data requirements, 125–141
speed and scale, 14–15
standards and frameworks, 222–250
supply chain governance, 92–93
techniques, 6–7
transparency, 21, 155, 167, 171, 176, 197, 201, 202, 371–372, 395–396
types of, 4–7
unique challenges in, 103
Artificial Superintelligence (ASI), 5
Assurance of AI Systems (ARIA), 236–238
auditability, 112–113
automated decision-making, 56–57, 141
autonomous decision-making, 72–82

B

bias, 8, 11, 44, 62, 75, 92, 93, 122, 128, 130, 131, 132, 134, 135–136, 161, 164, 165, 167, 185, 205, 238, 248, 249, 263, 288, 319, 323, 326, 328, 383
bias mitigation, 19
biometrics, 128–129
black boxes, 13, 21, 103
blended learning, 46
business alignment, 58–59
business- and consumer-facing technology, 55
business context, 265–266
articulating AI use case, 267–269
assessing data availability and suitability, 268–269
definition of, 266–267
ethical and operational considerations, 269–271
human oversight and decision rights, 270
organizational readiness and governance, 266
organizational values and strategy alignment, 270–271
risk of harm and ethical tradeoffs, 269–270
stakeholder identification and engagement, 267
understanding organizational objectives, 266
workforce readiness and culture, 270
business record management, 380
broader organizational practices, integration, 386
governance and accountability structures, 383, **384**
incident documentation and tracking, 380–381
issue reporting and lifecycle management, 381–382
legal and regulatory considerations, 385
post-market monitoring plans, 382–383
retention and disposal policies, 384
risk documentation and mitigation tracking, 382
tools and technologies, 384–385

C

ChatGPT, 5, 7, 16, 175, 248
Chief AI Ethics Officer (CAIEO), role of, 33
Claude, 16
company size, 50–52
consent, 104–105
vs. contract, 133
containerization tools, 336
continuous integration (CI), 316
continuous integration/continuous deployment (CI/CD), 363
continuous monitoring, 343
detecting drift and performance degradation, 344–345
fairness and ethical monitoring, 345
foundations of AI system monitoring, 343–344
handling emerging risks, 346–347
human oversight, 347
maintenance schedules, 346
technical frameworks, 344
third parties, 347–348

Copilot, 16
copyrighting AI-generated works, 153–154
copyright law, 149
cost-reduction organizations, 59
credit scoring models, bias in, 164
cross-border data transfers, 118–120
cross-border IP conflicts and jurisdictional complexities, 155–157, **156**
cross-functional product safety governance, 187
cross-functional review teams, building, 88
cyber insurance policy, 334
cybersecurity and risk management teams, 34

D

dark patterns, 105, 173
data acquisition and use, 77
data controllers, 114–125
data provenance logs, 152
data flow diagrams (DFDs), 302
data governance, 302–303
  benefits in AI development, 309
  common pitfalls and failure modes, 308–309
  data classification, 302
  data quality and integrity, 305
  data stewardship and roles, 306–307
  data sufficient and representative data, 305–306
  defining for AI, 303
  ethical considerations, 307
  fit-for-purpose evaluation, 306
  legal authority and lawful use of data, 303–304
  organizational maturity and implementation, 308
  privacy risk assessments, 307
  security controls, 307
data labelling, 93
data lineage and provenance, 309–310, *309*
  benefits of strong traceability, 314
  building framework, 313–314
  capturing data provenance, 311–312
  common failure modes, 312–313
  definition of, 310
  role in AI, 310–311
  traceability challenges and tradeoffs, 314–315
  tracking data lineage, 312
  use cases and industry examples, 314
data loss prevention (DLP), 303
data minimization, 20, 107–108, 135
Data Processing Agreement (DPA), 117
Data Protection Impact Assessment (DPIA), 40, 114–116
  and risk management, 137–139
data subject requests (DSRs), 139
data subject rights, 86–87, 120–122
  and sensitive data, 139–141
decision-rights frameworks, 39
decision support AI, 56
deepfakes, 174–175
deep learning, 6
deep neural networks, 83
defect in AI-driven products, 181
deployers, 207, 212
deployment considerations, 338
  cross-functional governance alignment, 342–343, **343**
  data governance, 339–340
  enforcing deployment policies and procedures, 338–339
  issue management and incident response, 341
  risk management, 340
  user training and enablement, 342
design defects, in AI systems, 182
design risks, 281–282
  framing AI, 282
  internal *vs.* external risk sources, 282
  risk governance integration, 286–287
  risk matrix, 283, **283**
  risk mitigation hierarchy, 283–284
  stakeholder mapping and risk perception, 284

design risks (*Cont.*)
systematic risk identification techniques, 283–284
testing, piloting and validation, 285–286
use case evaluation and scenario planning, 284–285
distributors, 207, 213
documentation, 81, 112–113
drift, 33, 64, 80, 82, 110, 118, 122, 177, 181, 282, 286, 287, 305, 323, 340, 344–345, 358
detection, 340–341
external communications, 386
AI lifecycle planning, 390
AI system capabilities and limitations, 387–388
governance structure, 386–387
managing public and media relations, 388–389
measuring effectiveness, 391
partner and vendor, 390
regulatory and legal obligations, 388
user-facing and transparency, 389
extraterritoriality, 211

E

effective cross-functional governance, principles of, 38–39
efficiency-oriented organizations, 59
employee use of generative AI tools, 94
employment, AI in, 161–162
epistemic uncertainty, 230
Equal Credit Opportunity Act (ECOA), 162–163
Equal Employment Opportunity Commission (EEOC), 161–164
Ethically Aligned Design (EAD), 247
ethics, 17–18, 40, 76–77
EU AI Act, 75, 194, 288–289
classification requirements, 199–204
enforcement and penalties, 208–211
implementation, 214–215, **215**
organizational context, 211–214
requirements for general-purpose AI, 204–208
risk classification, 194–198, **198**
European Union (EU), 158, 185, 246–247
Everalbum app, 172
expert systems, 7
explainability, 14, 21
explainable AI (XAI), 21
explicit consent, 104, 132–133
Exploratory Data Analysis (EDA), 324

F

Fair Credit Reporting Act (FCRA), 163
Fair Housing Act (FHA), 164–165
fairness, 18–19, 168–169
field testing, in ARIA, 237
full-stack security testing, 364

G

G7 Hiroshima Process and Code of Conduct, 247
General Data Protection Regulation (GDPR), 102, 122–123
Article 9 and prohibition baseline, 126
general-purpose AI (GPAI), 204–208
Google Translate, 6
Govern function, in NIST AI RMF, 226–228
Grok, 16

H

hallucinations, 94, 115, 177, 336, 341, 346, 347
harm, 15
HeyGen, 4

High-level Expert Group (HLEG), 246–247
high-risk AI, 195–196, 199–202
high-risk categories, of sensitive data, 128–132
human-machine interface and user misunderstanding, 183

I

impact assessments, 40, 271
  conduction of, 273–275
  development and deployment integration, 276
  preparation for, 272–273
  reviewing and operationalizing, 275–277
  understanding AI, 271–272
importer, 213
incident management, 81–82, 122–123, 367
  categorizing and escalating incidents, 369
  documenting incidents, issues and risks, 369–370
  identifying and managing AI incidents, 368
  organizational preparedness and readiness, 372–373
  stakeholders collaboration, understand root causes, 370–371
  transparency and public disclosures, 371–372
incident response (IR), 81, 232
industry sector-specific governance imperatives, 54–56
inference of sensitive data, 136
inferential privacy violations, 136
innovation-driven organizations, 58
Institute of Electrical and Electronics Engineers (IEEE), 247
  Ethically Aligned Design (EAD) framework, 247–248
insurance, 165–166
intellectual property (IP), 148–151
internal and external reporting, 81
internal use AI, 56
International Electrotechnical Commission (IEC), 238
International Organization for Standardization (ISO), 238
  for AI governance, 244–245
  AI standards in global governance landscape, 245–246
  ISO/IEC 5338, 240
  ISO/IEC 22989, 238–240, **239**
  ISO/IEC 23053, 240
  ISO/IEC 23894, 244
  ISO/IEC 38507, 244
  ISO/IEC 42001, 240–244
  ISO/IEC 42005, 245
  ISO/IEC JTC 1/SC 42 roadmap and future standards, 249–250
  ISO/IEC TR 24028, 245
international treaties, on intellectual property, 157
interpretability, 14, 21
IT governance, 72
  combining AI governance and, 75

J

joint authorship, and AI as tool, 154
jurisdictional clarity, for GPAI providers, 211

K

key performance indicators (KPIs), 41, 48

L

large language models (LLMs), 4, 5, 16, 83
legal basis and consent, for sensitive data, 132–133
licensing models for training data, 152
lifecycle-based data retention controls, 86
lifecycle documentation, 259–260

lifecycle documentation (*Cont.*)
designing and building process, 261–262
maintaining traceability and audit readiness, 264–265
requirements for, 260
testing and validation documentation, 264
training documentation, 263
lifecycle policy model, 73–74
limited-risk AI, 196–197, 202–203
localization, 394

M

machine learning (ML), 6, 60, 106
Manage function, in NIST AI RMF, 231–233
mandatory *vs.* elective content, 47
manufacturing defects, in AI-integrated products, 182–183
Map function, in NIST AI RMF, 228–229
maturity models, 51–54
Measure function, in NIST AI RMF, 230–231
metadata, 311
minimal-risk AI, 197–198, 203–204
model card, 39, 40, 65, 78, 110, 113, 123-125, 158, 231, 262, 314, 320, 321, 334, 367
model development, 77–78
model registry, 339
model testing, in ARIA, 236
modern AI systems, 89
multi-party liability, and AI supply chain, 185
mutual assistance framework, 210

N

National Institute of Standards and Technology (NIST), 224
AI Risk Management Framework (AI RMF), 224–235
Assurance of AI Systems (ARIA), 236–238
Cybersecurity Framework (CSF), 234–235
natural language processing (NLP), 7, 262
notice of privacy practices (NOPP), 16, 85

O

open-source AI, risk policies, 93
operationalizing
AI ethics, 76
privacy principles in AI, 111
opt-out rights and limitations in AI contexts, 140–141
Organisation for Economic Co-operation and Development (OECD), 222–224, 248
organizational context
and EU AI Act, 211–214
tailoring AI governance to, 49–59
organizational maturity, and AI governance, 51
overclaiming, 388

P

patent law and AI-generated inventions, 149–150
personally identifiable information (PII), 8
policies
autonomous decision-making, 72–82
existing data privacy and security policies, 83–88
third-party risk, managing, 88–95
policy revision, practical steps for, 87–88
post-deployment monitoring, 80
pre-deployment controls, 80
pre-release readiness, 333
AI model deployment considerations, 335–336, **335**
AI model optimization for production fit, 336–337, **336**
assessing AI system readiness for production, 334
final gate reviews and stakeholder signoff, 337–338

human-in-the-loop and release protocols, 337
operational readiness and risk evaluation, 334–335
privacy
impact assessments, 114–116
policy, 16, 84–85
purpose limitation, 84, 85, 105–106, 111, 269, 339
regulations concerning sensitive data, **127**
of responsible AI, 20
roles and responsibilities of AI, 112
tradeoffs in digital therapeutics, 130
violations, 8
privacy by design (PbD), 108–111
proactive approach and consumer protection, 178–179
proactive risk mitigation strategies, 185
product liability, modernizing, 187–188
proprietary model considerations, 287–288
data rights, quality, and governance, 289–290
intellectual property and competitive risk, 292–293
model explainability, testing, and monitoring, 290–292
ownership and accountability, 288–289
resourcing and organizational readiness, 293–295
pseudonymization techniques, 85

Q

quality management system (QMS), 202

R

real-time/high-stakes AI, 57
recordkeeping and accountability, 123–125
red teaming
and adversarial testing, 79
in ARIA, 236–237
regional variability, database rights and, 151
regulations
consumer protection laws, 169–179
intellectual property laws, 148–159
non-discrimination laws, 159–169
product liability laws, 179–188
reinforcement learning, 7
reporting, 81
reputation-focused organizations, 59
Responsible, Accountable, Consulted, and Informed (RACI) matrix, 36, **37**, 39
responsible AI, 17–23
retirement risk, 396
Retrieval Augmented Generation (RAG), 8
risk assessment forms, 40
risk assessments, 335
risk management, 114–116, 355
documentation and assurance, 360
DPIAs and, 137–139
forecasting secondary and downstream harms, 358
governance structures, 359
proactively identifying and categorizing, 356
regulatory considerations, 360
risk forecasting techniques, 358–359
supply chains, 356–357
system, 199–200
tailored for AI, 75–76
risk mitigation and compliance strategies
consumer protection laws and, 176–178
intellectual property laws and AI, 154–155
non-discrimination laws and AI, 167–168
risk rating frameworks, 75
risk register, 34, 113, 125, 138–139, 243, 287, 343, 359, 360, 364, 382, 385
risk tolerance of organization, 57–58
role-based governance controls, 65–66

S

scaling governance, 113–114
sectoral and civil laws
  consumer protection laws, 169–179
  intellectual property laws, 148–159
  non-discrimination laws, 159–169
  product liability laws, 179–188
security policy updates, 87
semi-supervised learning, 7
shadow AI, 313, 323, 363
Singapore Model AI governance framework, 249
Siri, virtual assistant, 4
software bill of materials (SBOM), 87, 186
stakeholder categories and governance roles, 31–35
supervised learning, 7
synthetic data, 77, 108, 152, 324
system audit and testing, 361
  audit documentation and evidence retention, 366–367
  audit roles and responsibilities, 366
  integrating testing into lifecycle, 365–366
  performance audits and model validation, 361
  real-time testing and continuous monitoring, 365
  red teaming for adversarial robustness, 362
  regulatory and compliance audits, 364–365
  security testing for, confidentiality and integrity, 363–364
  threat modeling, AI-specific attack surfaces, 362–363
  tooling and automation, 367
system design considerations, 277
  architecture and model selection, 278–279
  human oversight and feedback loops, 279–280
  metric selection, 280
  operational controls, 281
  purpose and framing the problem, 277–278
  threshold setting and adjustment, 281
system development lifecycle (SDLC), 391
system retirement, 391
  communication and transparency, 395–396
  data and model artifacts, 395
  learnings and organizational feedback loops, 397
  localization and regional retirement strategies, 394
  planning for controlled deactivation, 393–394
  policy, 392–393
  understanding the need, 392
system testing, 315
  bias and fairness testing, 319
  challenges and tradeoffs, 321
  discipline of, 315
  documentation and traceability, 320–321
  human-in-the-loop and scenario testing, 320
  integration testing, 316–317
  interpretability and explainability testing, 319
  organizational practices, 321
  performance testing, 317–318
  security testing, 318
  understanding objectives, 315–316
  unit and component testing, 316
  validation testing, 317

T

technical and engineering teams, 33–34
technical design, building ethics into, 76–77
tenant screening and credit-based bias, 165
testing risks, 321–322
  autonomous vehicle simulation, 327
  case studies and examples, 326–327
  common testing pitfalls, 326
  financial services outlier, 327
  hiring algorithm bias, 326
  identification techniques, 324
  risk mitigation strategies, 324
  testing tools and infrastructure, 326

training lifecycle, 325
types of, 322–323
unique nature of AI testing, 322
third-party processors, in AI projects, 116–118
third-party risk, 88–95
third-party risk management (TPRM), 357
trade secrets, and AI models, 150–151
traditional data governance, AI disrupts, 83–84
training and awareness program
on AI terminology, strategy, and governance, 43
avoiding training pitfalls, 48, **49**
building awareness beyond formal training, 47–48
curriculum structure, 44–45
delivering training, 45–47
foundation of, 43–44
learning styles and strengths, **46**
maintaining training records, 47
mapping audiences to training needs and topics, **45**, **46**
scope of, 44
sustaining program over time, 49
tracking and measuring success, 48
training data dilemma, 151–152
Transfer Impact Assessments (TIAs), 119–120
transparency, 14, 21, 62, 66, 81, 106, 155, 167, 171–172, 176, 197, 201–203, 205, 206, 223, 274, 276, 314, 371–372, 389, 395–396

U

UNESCO recommendation, on ethics of AI, 248–249
unsupervised learning, 7
U.S. Copyright Act, 149
use case assessment, 74–75
U.S. Federal Trade Commission (FTC), 170–173
U.S. Legislation of AI regulations, 157–158

V

vendor risk tiers, 91

W

World Economic Forum (WEF), AI governance toolkits, 248

Z

zero-shot learning (ZSL), 7

Online Test Bank

To help you study for your IAPP AIGP Artificial Intelligence Governance Professional Study Guide, register to gain one year of FREE access after activation to the online interactive test bank—included with your purchase of this book!

To access our learning environment, simply visit `www.wiley.com/go/sybextestprep`, follow the instructions to register your book, and instantly gain one year of FREE access after activation to:

- Practice test questions, so you can practice in a timed and graded setting
- Flashcards
- A searchable glossary



Keep Learning about AI Governance!

Video Course:

Artificial Intelligence Governance Professional (AIGP) Certification: Your Guide to Building a Management and Governance System to Ensure Responsible and Ethical AI



PRESENTED BY

Peter H. Gregory

CISSP, CISM, CISA, CRISC, CDPSE, CIPM, CCSK

Peter is a career cybersecurity leader and practitioner with 30 years of real-world security experience. He has written over 50 books on cybersecurity and privacy certifications, as well as training courses on O'Reilly. He also has 10 years of experience teaching cybersecurity at the University of Washington.

O'REILLY®

https://learning.oreilly.com/course/artificial-intelligence-governance/0642572035624/

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley’s ebook EULA.