

1ST EDITION

AI Governance in Practice

Build responsible, audit-ready programs with ISO/IEC 42001, NIST AI RMF, and lifecycle controls

HEMANG DOSHI

AI Governance in Practice

Build responsible, audit-ready programs with ISO/IEC 42001, NIST AI RMF, and lifecycle controls

Hemang Doshi

‹packt›

AI Governance in Practice

Copyright © 2026 Packt Publishing

All rights reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, without the prior written permission of the publisher, except in the case of brief quotations embedded in critical articles or reviews.
Every effort has been made in the preparation of this book to ensure the accuracy of the information presented. However, the information contained in this book is sold without warranty, either express or implied. Neither the author, nor Packt Publishing or its dealers and distributors, will be held liable for any damages caused or alleged to have been caused directly or indirectly by this book.
Packt Publishing has endeavored to provide trademark information about all of the companies and products mentioned in this book by the appropriate use of capitals. However, Packt Publishing cannot guarantee the accuracy of this information.

Portfolio Director: Vijin Boricha
Portfolio Manager: Anindya Sil
Program Manager: Ankita Thakur
Content Engineer: Tazeen Shaikh
Technical Editor: Nithik Cheruvakodan
Copy Editor: Tazeen Shaikh
Indexer: Rekha Nair
Proofreader: Tazeen Shaikh
Production Designer: Aparna Bhagat
Growth Lead: Ankita Thakur

First published: September 2026

Production reference: 1250926

Published by Packt Publishing Ltd.
Grosvenor House
11 St Paul's Square
Birmingham
B3 1RB, UK.

ISBN 978-1-80730-081-4
www.packtpub.com

To my mother, Jyoti Doshi, and to the memory of my father, Hasmukh Doshi, for their sacrifices and for exemplifying the power of determination.

To my wife, Namrata Doshi, for being my loving partner throughout our life journey together, and to my kids, Jia and Neev, for giving me the time and space to write this book.

To my sister, Pooja Shah, my brother-in-law, Hiren Shah, and my nephew, Phenil Shah, for their love, support, and inspiration.

To my in-laws, Chandrakant Shah, Bharti Shah, and Ravish Shah, for their love and motivation.

To my mentor and guide, Dipak Mazumder, for showing me how talent and creativity evolve.

– Hemang Doshi

Contributors

About the author

Hemang Doshi has more than 20 years of experience in system auditing, IT risk and compliance, internal audit, risk management, information security auditing, third-party risk management, AI GRC, and operational risk management. He has authored several books on certifications such as CISA, CRISC, CISM, Certified in Cybersecurity, DISA, and CEH, as well as enterprise risk management. His books and lectures are available in more than 175 countries and in more than 35 languages.

About the reviewers

Mani Keerthi Nagothu is a cybersecurity professional with global work experience. She is a well-known public speaker at various cybersecurity conferences, webinars, and podcasts. Her focus areas include cybersecurity strategy, incident response, and risk management. She is passionate about teaching cybersecurity, and four of her courses, *Insider Threat Risk Management*, *Cyber Supply Chain Risk Management*, *Designing Tabletop Exercises*, and *Designing Cybersecurity Controls*, are published on LinkedIn Learning.

Javen Khoo Ai Wee, CMIIA, CIA, CISA, AAIA, CFE, ISC² CC is an internal audit and governance, risk, and compliance (GRC) professional with more than two decades of cross-functional experience across oil and gas, real estate, financial services and corporate advisory. With more than 10 years in internal audit, risk, and assurance-related roles, Javen currently serves as an independent GRC and internal audit advisor, supporting organizations in strengthening governance effectiveness, enhancing risk culture, and improving control discipline.

Javen previously held leadership roles in PETRONAS Group Internal Audit and KLCC Group, where she directed high-impact audit engagements across PETRONAS' downstream and gas value chains, corporate functions and KLCC Group's diversified businesses. Beyond leading assurance portfolios, Javen is an active contributor to the internal audit and governance profession. She is a former member of the Research and Technical Advisory Committee of The Institute of Internal Auditors Malaysia (IIA Malaysia) and is currently a technical writer for IIA Malaysia, where she has published thought-leadership articles on cybersecurity, data integrity, fraud, and the evolving value of internal auditing. She also engages with academic institutions, including University of Malaya and Tunku Abdul Rahman University of Management and Technology, as a resource person for PhD research studies on AI in internal auditing and ESG assurance, thereby bridging academic rigor with practitioner experience.

Table of Contents

Preface **xvii**

Chapter 1: Understanding Artificial Intelligence **1**

Understanding AI **2**
Differentiating normal system vis-à-vis AI-based system • 2
AI system • 4
Elements of AI **5**
Data • 6
Algorithm • 6
Model • 6
Types of algorithms **7**
Supervised learning algorithms • 7
Unsupervised learning algorithms • 7
Semi-supervised learning algorithms • 8
Reinforcement learning algorithms • 9
Machine learning **9**
Neural network • 10
Deep learning • 10
Types of AI models **11**
Decision tree model • 11
Recognition AI model • 12
Personalization model • 12
Natural language processing (NLP) model • 12
Robotic model • 12
Multimodal AI model • 12
Large language models (LLMs) • 12
AI deployment options **13**
Cloud deployment • 13
On-premises deployment • 14
Edge deployment • 14

Summary **14**

Chapter 2: Introduction to AI Governance 15

Understanding governance **15**
Elements of governance • 16
Governance versus management versus operations **17**
Governance and regulations **18**
Policies, procedures, and guidelines **18**
Differentiating AI governance and IT governance **19**
Why AI needs stronger governance than traditional systems **20**
Governing AI-based systems **21**
AI governance – roles and responsibilities **21**
AI governance – policies **22**
AI governance – KPIs **23**
Summary **25**

Chapter 3: Principles of Responsible AI 27

Who defines Responsible AI principles **28**
Responsible AI principles **29**
Accountability • 29
Accountability under EU AI Act • 30
Accountability – practical implementation • 32
Fairness and non-discrimination • 33
Fairness – practical implementation • 34
Transparency • 34
Transparency – practical implementation • 35
Explainability • 35
Explainability – practical implementation • 36
Privacy and data protection • 36
Security and resilience • 37
Security and resilience – practical implementation • 38
Safety and reliability • 39
Safety and reliability – practical implementation • 39
Human oversight • 40
Human oversight – practical implementation • 40

Risk-based approach • 41
Risk-based approach – practical implementation • 41
Ethical use • 41
Ethical use – practical implementation • 42
Continuous monitoring and improvement • 42
Continuous monitoring and improvement – practical implementation • 43
Compliance with laws and standards • 44
Compliance with laws and standards – practical implementation • 44
Summary 45

Chapter 4: Overview of Global AI Governance Frameworks – ISO/IEC 42001 and EU AI Act 47

Understanding the difference between regulations, standards, and frameworks 47
EU AI Act 49
General data protection regulation (GDPR) 51
US executive order on safe, secure, and trustworthy AI 52
ISO/IEC 42001 – artificial intelligence management systems 53
NIST AI risk management framework (AI RMF) 54
OECD AI principles 54
IEEE 7000 series 55
World economic forum AI governance playbooks 56
Common themes across the globe for AI governance 56
Summary 59

Chapter 5: AI Governance Framework – ISO/IEC 42001 61

ISO and IEC – international standards bodies 61
What is a standard? • 62
How international standards are developed • 62
Structure of AIMS under ISO/IEC 42001 63
Ten clauses (clauses 1 to 10) • 63
Four annexes (annexes A, B, C, and D) • 64
Annex A • 65
Statement of Applicability (SoA) • 66
Implementation steps for ISO/IEC 42001 66
Sample case study – implementing ISO/IEC 42001 in a financial services company • 67

1. Create an AI governance policy • 67
2. Perform gap analysis • 67
3. Prepare the SoA • 68
4. Implement and maintain controls • 68
5. Certification audit • 68
Common pitfalls when implementing ISO/IEC 42001 68
Certification audit process 69
Importance of ISO/IEC 42001 certification 72
Leveraging an existing ISO/IEC 27001 certificate to implement ISO/IEC 42001 72
1. Review the scope • 73
2. Align the risk management processes • 73
3. Align the governance process • 73
4. Align the management review process • 74
5. Align the internal audit processes • 74
Summary 75

Chapter 6: AI Governance Framework – NIST AI RMF 77

What is NIST? 77
What is NIST AI RMF? 78
Components of NIST AI RMF 78
Govern • 80
Developing policies and processes • 81
Establishing accountability and responsibility • 81
Ensuring diversity, equity, inclusion, and accessibility • 82
Establishing AI risk culture • 82
Engaging the stakeholders • 82
Managing third-party and supply chain risks • 83
Map • 83
Establishing and understanding context • 83
Categorizing the AI system • 84
Understanding AI capabilities, benefits, and costs • 84
Mapping risks and benefits across the AI system • 84
Assessing AI impacts • 84
Measure • 85
Applying AI risk metrics • 85

Evaluating AI trustworthiness • 85
Tracking AI risks over time • 86
Assessing measurement effectiveness • 86
Manage • 87
Managing AI risks • 87
Maximizing benefits and minimizing impacts • 87
Managing third-party AI risks • 88
Monitoring, improving, and responding • 88
Implementing NIST AI RMF 88
Other implementation support 89
1. AI RMF profiles • 89
2. AI RMF playbook • 91
3. AI RMF crosswalks • 92
Summary 92

Chapter 7: AI Governance Framework – OECD AI Principles 93

What is OECD? 94
Background and evolution of the OECD AI Principles 94
Scope and applicability of the OECD AI Principles 95
The five OECD AI Principles 95
Practical application of OECD AI Principles in organizations 97
Principle 1: Inclusive growth, sustainable development, and well-being • 97
Principle 2: Respect for the rule of law, human rights, and democratic values, including fairness and privacy • 98
Principle 3: Transparency and explainability • 98
Principle 4: Robustness, security, and safety • 99
Principle 5: Accountability • 99
Mapping OECD AI Principles to AI governance frameworks 100
Summary 102

Chapter 8: European Union AI Act 103

Global impact of the EU AI Act 103
Applicability of the EU AI Act • 104
AI system 104
Classification of AI systems 104

Prohibited AI • 106
High risk AI systems • 106
Limited risk AI • 108
Minimal risk AI • 108
Determining the classification of an AI system 109
Obligations by role 111
Complaint and redress mechanism 112
Penalties, fines, and enforcement 112
Practical implementation roadmap 114
Summary 116

Chapter 9: AI Impact Assessment 117

AI impact assessment requirements under different frameworks, standards and regulations 117
Benefits of AI impact assessment 118
Key areas of AI impact 118
Steps to conduct effective AI impact assessment 121
Summary 123

Chapter 10: AI Risk Assessment 125

Types of AI risks 126
AI risk assessment requirements under different frameworks, standards, and regulations ... 127
AI risk assessment steps 128
Risk identification • 129
Risk analysis • 131
Likelihood and impact • 131
Risk evaluation • 133
Risk mitigation/treatment • 133
Risk management policy 134
Risk register 135
Summary 135

Chapter 11: Comparative Analysis of AI Frameworks – ISO/IEC 42001 and EU AI Act 137

Overview of major AI governance frameworks 137

Selecting the right framework for your organization **138**
When to use the OECD AI Principles • 138
When to use the NIST AI RMF • 139
When to use ISO/IEC 42001 • 139
EU AI act • 140
The ISO advantage **140**
OECD principles mapped to ISO/IEC 42001 standard • 141
NIST AI RMF mapped to ISO/IEC 42001 standard • 143
Summary **144**

Chapter 12: AI Governance Structure **147**

Objectives of an AI governance structure **147**
Governance model **148**
RACI matrix for AI governance **150**
Implementing AI governance structures **152**
Common pitfalls in implementing AI governance **154**
Summary **155**

Chapter 13: AI Governance Documents **157**

Importance of documented AI policies and procedures **158**
Document hierarchy **159**
Important AI governance documents **161**
AI governance policy • 161
AI use case assessment and approval form • 162
AI system inventory • 162
AI impact assessment • 162
AI risk assessment • 163
AI risk register • 163
AI vendor and third-party assessment • 163
AI model and system documentation • 164
AI monitoring and review records • 164
AI incident management procedure and records • 164
AI change management records • 165
AI governance committee records • 165
Training and awareness records • 165

AI governance review and audit records • 165

Key contents of AI governance policy 166

Documentation and record keeping requirements under different regulations and frameworks. 170

EU AI Act • 170

ISO/IEC 42001 artificial intelligence management systems • 170

NIST AI Risk Management Framework (AI RMF) • 171

OECD AI Principles • 171

Stepwise procedure for creating AI documents 171

Summary 173

Chapter 14: AI Lifecycle Governance 175

AI lifecycle stages – a governance perspective 176

Ideation and use-case definition • 177

Data collection and preparation • 178

Model design and development • 178

Testing and validation • 179

Deployment and integration • 180

Monitoring and ongoing management • 180

Retirement and decommissioning • 181

Third-party and vendor AI lifecycle governance • 182

Summary 182

Chapter 15: Security of AI Systems 183

Understanding the unique security risks of AI systems 183

Protecting data, models, and infrastructure 185

Protecting AI data • 185

Protecting AI models • 187

Protecting AI infrastructure • 188

Role of the AI GRC professional in the security of AI systems 189

Summary 190

Chapter 16: AI Incident Management 191

Incident criteria 193

AI incident management framework 195

Role of the AI GRC professional in AI incident management **197**
Incident register **198**
Common pitfalls in AI incident management **200**
AI incident management checklist **201**
Summary **202**

Chapter 17: AI System Audits **203**

Responsibility of audit **204**
Role of AI GRC professional in AI system audits **205**
Key audit areas in AI systems **206**
AI governance and oversight • 206
Data quality and integrity • 207
Bias and fairness • 207
Transparency and explainability • 207
Security risks in AI systems • 207
Privacy and regulatory compliance • 207
Model performance and reliability • 207
Human oversight and accountability • 207
AI incident management • 208
Types of AI audits **208**
Summary **209**

Chapter 18: Unlock Access to the Code Bundle and the PDF Version **211**

Unlock this book's free benefits in three easy steps **212**

Other Books You May Enjoy **216**

Index **219**

Preface

Turn AI governance principles, standards, and regulatory requirements into a practical program that operates across the AI lifecycle.

This book shows you how to build responsible AI governance around ISO/IEC 42001, the NIST AI RMF, OECD AI Principles, the EU AI Act and other frameworks. You'll conduct AI risk and impact assessments, define governance roles and decision rights, establish policies and lifecycle controls, and create documentation for transparency, accountability, and AI compliance.

You'll apply governance through development, deployment, monitoring, and retirement; address AI security and safety; prepare for AI incidents; and use AI audits and monitoring findings to strengthen controls. Practical scenarios, templates, checklists, and step-by-step guidance turn frameworks into repeatable organizational practices.

Written by Hemang Doshi, who has more than 20 years of experience in system audit, IT risk and compliance, internal audit, risk management, information security audit, third-party risk management, and operational risk management, this book connects governance frameworks with practical implementation. By the end, you'll be able to build an accountable, audit-ready AI governance program and integrate AI risk management with existing compliance, security, privacy, and risk processes.

Who this book is for

AI governance, compliance, and risk management professionals responsible for overseeing AI systems, evaluating AI risks, or implementing responsible AI practices, along with data scientists, ML engineers, internal auditors, privacy leaders, and technology executives. A foundational understanding of AI concepts, basic IT risk principles, and organizational policies or governance frameworks is recommended. This book also provides a strong foundation for AI GRC certifications such as AIGP, AAIA, AAIR, AAISM, AIMS and others.

What this book covers

Chapter 1, *Understanding Artificial Intelligence*, introduces the fundamentals of AI, its key concepts, types, applications, and lifecycle. It provides a foundation for understanding AI systems, risks, and governance.

Chapter 2, *Introduction to AI Governance*, introduces the fundamentals of AI governance, its objectives, key principles, and the importance of establishing effective governance practices.

Chapter 3, *Principles of Responsible AI*, introduces the key principles of responsible AI, including fairness, transparency, accountability, privacy, safety, and human oversight.

Chapter 4, *Overview of Global AI Governance Frameworks – ISO/IEC 42001 and EU AI Act*, provides an overview of ISO/IEC 42001, NIST AI RMF, OECD AI Principles, and the EU AI Act.

Chapter 5, *AI Governance Framework – ISO/IEC 42001*, introduces ISO/IEC 42001 and explains its key requirements for establishing and maintaining an AI management system.

Chapter 6, *AI Governance Framework – NIST AI RMF*, introduces the NIST AI Risk Management Framework and its approach to managing AI-related risks and promoting trustworthy AI.

Chapter 7, *AI Governance Framework – OECD AI Principles*, introduces the OECD AI Principles and their role in promoting innovative, trustworthy, and human-centered AI.

Chapter 8, *European Union AI Act*, introduces the EU AI Act, its risk-based approach, key requirements, and obligations for AI systems.

Chapter 9, *AI Impact Assessment*, explains the purpose and process of assessing the potential impacts of AI systems on individuals, organizations, and society.

Chapter 10, *AI Risk Assessment*, explains how to identify, assess, prioritize, and manage risks associated with AI systems.

Chapter 11, *Comparative Analysis of AI Frameworks – ISO/IEC 42001 and EU AI Act*, compares ISO/IEC 42001 and the EU AI Act, highlighting their key similarities and differences.

Chapter 12, *AI Governance Structure*, explains the roles, responsibilities, accountability, and organizational structure required for effective AI governance.

Chapter 13, *AI Governance Documents*, introduces the key policies, procedures, records, and other documentation required to support AI governance.

Chapter 14, *AI Lifecycle Governance*, explains how governance practices can be integrated throughout the AI system lifecycle, from planning and development to deployment and retirement.

Chapter 15, *Security of AI Systems*, introduces key security risks and controls for protecting AI systems, data, models, infrastructure, and supporting services.

Chapter 16, *AI Incident Management*, explains how organizations can identify, respond to, manage, and learn from AI-related incidents.

Chapter 17, *AI System Audits*, introduces the principles and practices for auditing AI systems, governance processes, risks, controls, and compliance.

To get the most out of this book

You should have a basic understanding of IT, cybersecurity, risk management, and governance. No advanced AI knowledge is required.

The book includes practical examples and templates to help you understand and apply AI governance concepts. You should also refer to the latest versions of relevant standards and regulations, as they may change over time.

Download the example code files

This book includes a complete downloadable code bundle containing all the example projects and files used throughout the chapters. We recommend downloading the bundle so you can follow along smoothly and experiment with the examples.

Use the bundle as a practical starting point. Modify it, extend it, and apply what you learn by creating your own variations as you progress through the chapters.

Get the code bundle

If you bought the book directly from Packt:

1. Go to **packtpub.com**
2. Click your **profile picture** and select **Your Orders**
3. Find this book and click **Download Code**

If you bought this book from Amazon or any other channel partner:

1. Go to `packtpub.com/unlock` or scan the following QR code



2. Search for this book
3. Sign up or log in to your free Packt account
4. Upload your proof of purchase and download the code bundle locally

Usage note: You're free to use and modify this code for personal learning and non-commercial projects.

Download the color images

We also provide a PDF file that has color images of the screenshots/diagrams used in this book. You can download it here: `https://packt.link/gbp/9781807300814`.

Conventions used

There are a number of text conventions used throughout this book.

Bold: Indicates a new term, an important word, or words that you see on the screen. For instance, words in menus or dialog boxes appear in the text like this. For example: "An AI system primarily consists of three elements: **data**, **algorithm**, and **model**."

> **Note**
>
> Warnings or important notes appear like this.

> **Tip**
>
> Tips and tricks appear like this.

Get in touch

Feedback from our readers is always welcome.

General feedback: If you have questions about any aspect of this book or have any general feedback, please email us at `customercare@packt.com` and mention the book's title in the subject of your message.

Errata: Although we have taken every care to ensure the accuracy of our content, mistakes do happen. If you have found a mistake in this book, we would be grateful if you reported this to us. Please visit `http://www.packt.com/submit-errata`, click **Submit Errata**, and fill in the form.

Piracy: If you come across any illegal copies of our works in any form on the internet, we would be grateful if you would provide us with the location address or website name. Please contact us at `copyright@packt.com` with a link to the material.

If you are interested in becoming an author: If there is a topic that you have expertise in and you are interested in either writing or contributing to a book, please visit `http://authors.packt.com/`.

Share your thoughts

Once you've read *AI Governance in Practice*, we'd love to hear your thoughts! Scan the QR code below to go straight to the Amazon review page for this book and share your feedback.



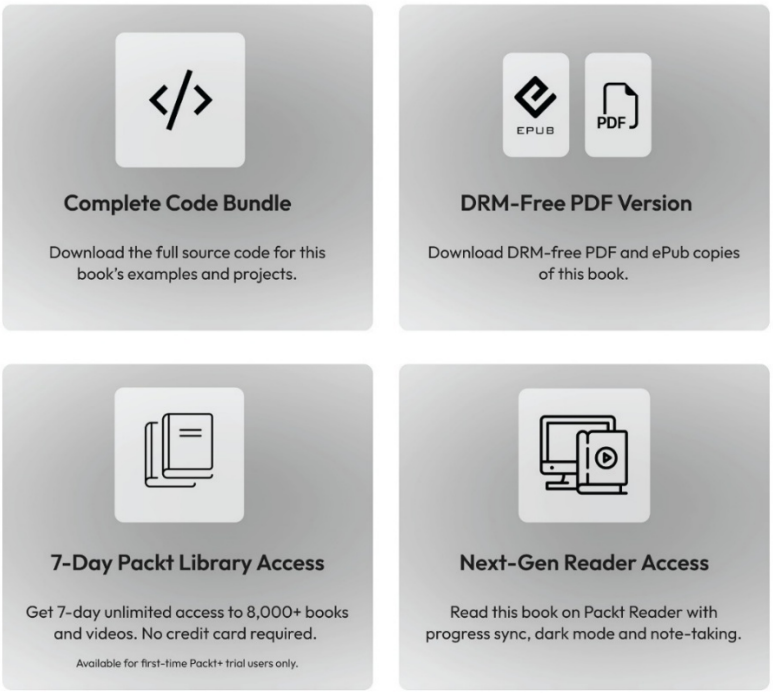
https://packt.link/r/1807300811

Your review is important to us and the tech community and will help us make sure we're delivering excellent quality content.

Free benefits with your book

This book includes free benefits designed to support your learning and help you apply what you learn effectively. Activate them now for instant access (see the *How to unlock* section for instructions).

Here's a quick overview of what you can instantly unlock with your purchase:



How to unlock

Scan the QR code (or go to `packtpub.com/unlock`). Search for this book by name, confirm the edition, and then follow the steps on the page.





Note: Have your invoice handy. Purchases made directly from the Packt website don't require an invoice.

1

Understanding Artificial Intelligence

It is Monday morning. You begin your day by checking your emails. One subject line shocks you: *"Your Citizen Credit Score Has Been Degraded."*

Your heart sinks. You immediately search online about consequences. What you find is frightening. The information suggests that once a citizen's credit score is degraded:

- You will not be eligible for state unemployment support
- Banks will not sanction any loans
- Your insurance premium will triple with immediate effect
- Except for emergency medical assistance, all other public services may be restricted

Anxiety sets in. You contact the government office mentioned in the email and speak with the concerned officer. During the conversation, you are informed that the citizen credit score is calculated using multiple parameters. However, as per regulation, the parameters and scoring logic cannot be disclosed. There is no manual override or immediate appeal mechanism.

Your survival now depends on an AI-based score you cannot see, understand, or challenge.

Please note that though it seems to be a fictional story, it can be a reality in the near future if **artificial intelligence** (**AI**) is not regulated and governed. With the speed at which AI is being adopted, there can be such negative impacts across geographies. AI can improve efficiency, but without strong governance, it can also silently control lives. This book is designed to help AI GRC professionals, business leaders, governance professionals, auditors, risk managers, compliance officers, and technology practitioners understand not only how AI works but also how it can be governed responsibly throughout its lifecycle. This opening chapter lays the foundation by introducing the fundamental concepts of artificial intelligence, enabling readers

to build the knowledge needed for the governance frameworks, risk management practices, and regulatory considerations discussed in the chapters that follow.

In this chapter, you will learn the fundamental concepts of AI that serve as the foundation for discussions on governance, risk, and compliance in this book. We begin by explaining what AI is and why it differs fundamentally from traditional software systems.

> **Note**
>
> **Download the code bundle and the PDF version of this book**
>
> Your purchase includes a DRM-free PDF copy of this book, the code bundle, and a range of exclusive benefits. To unlock everything, follow the *Free benefits with your book* section in the *Preface*.

Understanding AI

AI refers to computer systems that are designed to perform tasks that normally require human intelligence. In simple terms, AI is a technology that allows machines to:

- Learn from data
- Identify patterns
- Make decisions or predictions
- Act without being explicitly programmed for every situation

Unlike traditional software, which follows fixed rules, AI systems adapt and improve as they process more data. Let us differentiate AI systems from normal software.

Differentiating normal system vis-à-vis AI-based system

A normal (traditional) software system works on predefined rules written by developers. It does exactly what it is told to do and nothing more. If a new situation arises, the software must be manually updated to handle it.

An AI-based system, on the other hand, learns from data. Instead of relying only on fixed rules, it identifies patterns, makes predictions, and improves its behavior over time based on new data. The following table will help you to understand the differences:

Aspect	Normal System	AI-Based System
Decision logic	Based on fixed, predefined rules	Based on patterns learned from data

Aspect	Normal System	AI-Based System
Learning capability	Does not learn or improve automatically	Learns and improves with more data
Behavior over time	Always behaves the same way	Behavior can change over time
Handling new situations	Requires manual code changes	Can adapt based on training data
Transparency	Logic is usually easy to understand	Logic may be complex or opaque
Data dependency	Limited use of data	Heavily dependent on large datasets

Table 1.1: Differentiating normal system vis-à-vis AI based system

AI is when computers are designed to think, learn, or act like humans, without being explicitly programmed for every action. *'Without being explicitly programmed for every action'* is a very important differentiator between a normal system and an AI-based system.

For example, in a normal email filtering system, the rules are fixed and predefined by developers or administrators. If the email subject contains the word *Free*, it will mark it as spam. Even if the email is genuine, such emails will be moved to spam every time.

In an AI-based email filtering system, the system learns from data and user actions. The system analyzes email content and user behavior. If users move emails from spam to inbox, the system adjusts its decisions.

Another example can be of a loan approval system. In a normal system, if income is below a fixed limit, the loan will be rejected, whereas in an AI-based system, the loan may be approved after analyzing other factors such as spending habits, employment history, behavior patterns, and other data to predict the likelihood of loan repayment.

As an AI GRC professional, it is important to be very clear about these differences. AI-based systems cannot be governed by simply extending the existing control environment used for traditional systems. In a normal system, once the source code is reviewed and approved, there is usually no need for further review until a change is made to the code. The system behaves in a predictable and consistent manner.

In contrast, an AI-based system can change its behavior over time as it learns from new data. Even if the source code remains the same, the outputs may vary. Therefore, AI outputs must be reviewed and monitored on a continuous basis to detect errors, bias, drift, or unintended outcomes. We will discuss how to govern the AI systems in our subsequent lectures. Let us now understand the technical definition of AI as documented in ISO 22989.

AI system

AI terms are often confusing and used differently by vendors, regulators, and auditors. ISO/IEC 22989 solves this problem by providing clear definitions of AI-related terms. It is a highly recognized standard, as it is developed by participating countries through international consensus. ISO 22989 defines AI system as *"engineered system that generates outputs such as content, forecasts, recommendations or decisions for a given set of human-defined objectives."*

ISO 22989 further explains the definition: *"Note 1: The engineered system can use various techniques and approaches related to artificial intelligence to develop a model to represent data, knowledge, processes, and more, which can then be used to perform tasks."*

This means that an engineered system (something that is designed and built by humans) that uses artificial intelligence (machines thinking and acting like humans) to generate outputs like:

- Content (e.g., writing text, creating images)
- Forecasts (e.g., predicting weather or sales)
- Recommendations (e.g., suggesting movies, products)
- Decisions (e.g., approving a loan or detecting fraud)

These outputs are generated to meet specific goals or objectives set by humans. So, the AI doesn't decide its own goals. It works toward what people have programmed or defined for it. For example, accounting software that records transactions and generates a profit and loss statement using pre-programmed rules is not considered an AI system because the content (profit and loss statement) is not obtained as a result of intelligence (i.e., learning, adapting, or generating new insights from data).

However, if the same software provides a feature to predict future profits, then it is considered an AI-based system as it has the ability to learn from past data and generate new insights.

Note

The term AI system may have slightly different definitions across frameworks, standards, and regulations. In this book, we primarily use the ISO/IEC 22989 definition. The EU AI Act definition and its specific structure are discussed separately in *Chapter 8*.

Elements of AI

It is important to understand that every AI system is designed, trained, and operated using a structured approach. At its core, an AI system depends on a few fundamental building blocks that enable it to analyze information, learn from experience, and produce meaningful outputs. An AI system primarily consists of three elements: **data**, **algorithm**, and **model**, as can be seen in *Figure 1.1*:

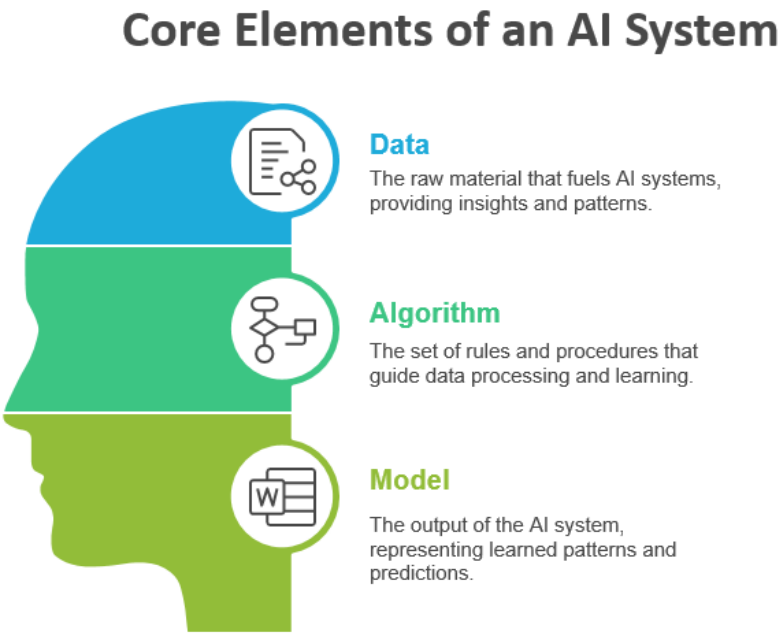


Figure 1.1: Core elements of an AI system

These elements work together to enable the AI system to learn from information, identify patterns, and make decisions or predictions. Data provides the input and knowledge base, the algorithm defines how learning takes place, and the model is the outcome of this learning process. Like any product is derived using input, process, and output, for creating an AI system, data acts as the input, providing the information on which the system learns or makes decisions. The algorithm represents the process, defining how the data is analyzed and how patterns are identified. The model is the output of this learning process and is then used to generate results or predictions for new data. Understanding these three elements is essential to understanding how an AI system functions in practice. Let us understand each element.

Data

Data refers to the information used by the AI system for learning and decision-making. It can be structured, such as tables and records, or unstructured, such as text, images, audio, or videos. The quality, accuracy, and relevance of data directly impact the performance of the AI system.

For example, in an email spam detection system, data consists of past emails that are labeled as *spam* or *not spam*. This historical email data is used to teach the AI system what spam emails look like.

Algorithm

An algorithm is a set of instructions or rules that tells the AI system how to learn from data. It defines the method used to analyze the data, find patterns, and adjust learning over time.

For example, in the spam detection system, the algorithm analyzes words, phrases, sender details, and patterns in emails to learn which characteristics are commonly associated with spam.

Model

A model is the result produced after an algorithm is trained on data. It represents learned patterns and knowledge and is used to make predictions or decisions. Example: A trained spam detection model decides whether a new email is spam or genuine.

As you can see, a model is created using both data and the algorithm. From a governance perspective, it is important to understand that the model is only as reliable as the data and algorithm used to build it. If the data is incomplete, biased, outdated, or incorrect, or if the algorithm is poorly designed or misconfigured, the resulting model can be misleading or even harmful.

This means that risks in AI systems often do not start at the model stage. They originate much earlier, at the data, and algorithm level. A misleading data set or a flawed algorithm can produce a misleading model, which may lead to incorrect decisions, unfair outcomes, compliance issues, or reputational damage. We will discuss in detail about the risks associated with data and algorithms in our subsequent chapters.

Data, algorithms, and models work together to enable AI systems to learn, make predictions, and support decision-making. Since the algorithm determines how the AI system learns from data, understanding the different types of AI algorithms is the next logical step in understanding how AI systems operate.

Types of algorithms

As we discussed earlier, an **algorithm** is a step-by-step set of instructions to achieve a desired result. An AI algorithm is a set of steps that allows a computer to learn from data, make decisions, or predict outcomes, instead of following only fixed rules. Algorithms used in AI systems can be broadly classified into different types based on how they learn and operate. Understanding these types is important from a governance and risk perspective, because each type carries different risks and limitations.

Supervised learning algorithms

Supervised learning is a type of machine learning where a computer learns using examples that already have correct answers. These correct answers are known as **labels**. In supervised learning, the system is given:

- Input data
- Correct output (label)

The system learns the relationship between input and output. Later, it uses this learning to predict answers for new data. For example:

- Input: Email
- Output (label): spam or not spam

The system learns from these labeled emails and then classifies new emails correctly. Prediction rate improves over time as more and more input data is fed to the algorithm. Supervised learning is primarily used for classification and prediction.

So, the technical definition of supervised learning is: *"Supervised learning is a process where an algorithm learns from labeled data to make predictions or classifications."* Common use cases of supervised learning are spam detection systems, credit approval systems, and fraud detection systems where historical data is labeled as spam or not spam, approved or not approved, fraud, or genuine.

Unsupervised learning algorithms

These algorithms work on unlabeled data. They try to find hidden patterns or groupings in data without predefined outcomes. They learn patterns from data without any labels or correct answers.

Common use cases for unsupervised learning can be customer segmentation (grouping customers based on buying behavior, age, spending habits, etc.), text and image grouping (sorting photos based on faces or scenery), behavior analysis (finding patterns in user or system behavior), and so on.

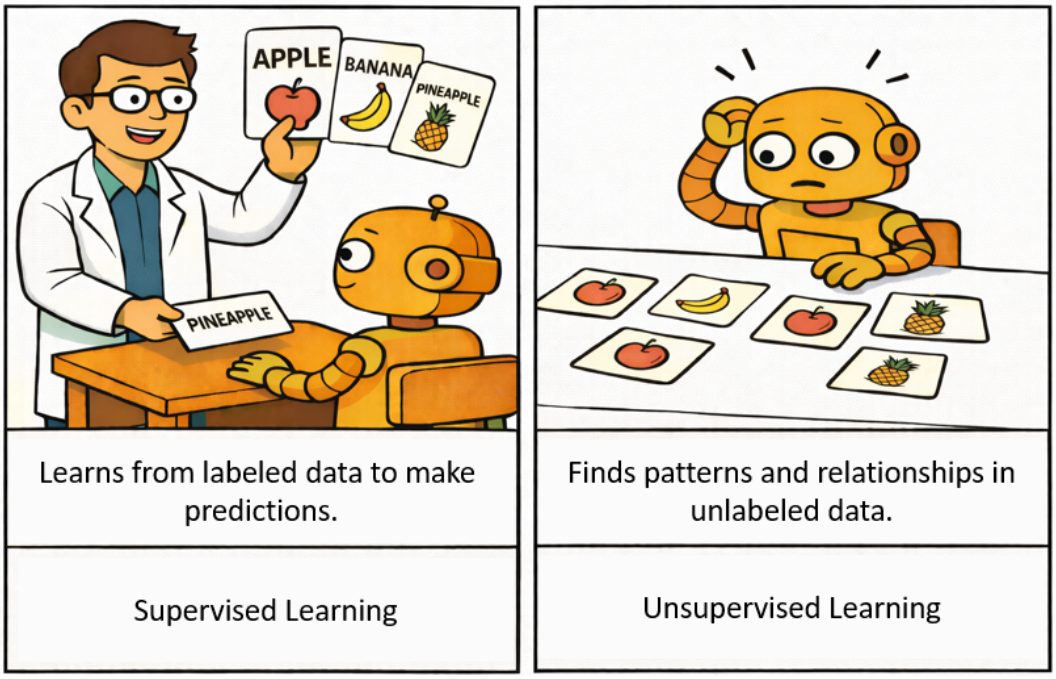


Figure 1.2: Supervised learning vis-à-vis unsupervised learning

As illustrated in the above figure, supervised and unsupervised learning differ primarily in whether labeled data is available during training. In practice, however, many real-world datasets contain only a small amount of labeled data and a large amount of unlabeled data. To address this challenge, AI systems often use semi-supervised learning, which combines the strengths of both approaches.

Semi-supervised learning algorithms

Semi-supervised learning algorithms are algorithms that learn using a small amount of labeled data and a large amount of unlabeled data. This approach is used when labeling data is costly or time-consuming, but unlabeled data is easily available. For example, a spam filter learning from a few labeled emails and many unlabeled emails.

Reinforcement learning algorithms

These algorithms learn by trial and error. They receive rewards or penalties based on actions taken and improve decisions over time. The goal is to maximize total reward over time. In reinforcement learning:

- An agent takes an action
- The environment responds
- The agent gets a reward or penalty
- The agent learns which actions give better rewards

Common use cases for reinforcement learning algorithms are teaching a robot the best path to reach a destination, self-driving cars, video game playing, and optimizing system performance.

Machine learning

ISO 22989 defines machine learning as *"process of optimizing model parameters through computational techniques, such that the model's behavior reflects the data or experience."*

In simple terms, machine learning is the ability of the machine:

- to learn from data
- behave in a way that matches what the data shows

Now let us correlate machine learning with AI system. AI is a big umbrella. It includes any technique that allows computers to act intelligently, such as solving problems, making decisions, or recognizing patterns. **Machine learning** (**ML**) is a subset of AI. It's one of the main ways we build AI systems today.

AI is the goal (to make machines smart), and ML is one way to reach that goal by letting machines learn from data instead of being told exactly what to do.

For example, an AI system that recommends movies uses machine learning to learn your likes and dislikes from your past choices.

We will now discuss models of ML.

Neural network

Neural networks are one of the models used in ML. They are primarily used when the data is complex, like images or speech. A neural network works like the human brain. It learns from data and finds patterns to make decisions or predictions.

A neural network is made up of layers of connected **neurons**, like building blocks:

- **Input layer**: Takes in raw data (like text, images, or numbers)
- **Hidden layers**: Process the data and learn patterns
- **Output layer**: Gives the result (like "cat" or "not cat")

Each connection between neurons has a weight, which changes as the system learns from data. This is how the model improves over time.

Neural networks are useful because:

- They can learn complex patterns from large amounts of data
- They work well for tasks where rules are hard to define manually (like facial recognition)
- They improve over time with more data and training

Neural networks are the core of many powerful AI systems, including image recognition, speech-to-text, and self-driving cars.

Deep learning

A neural network with one or two layers is used in basic tasks. When we add more layers, allowing the model to learn more complex patterns, it becomes a deep neural network. Using and training these deep networks is what we call **deep learning**.

For example, let's say we want to build an AI that recognizes handwritten digits. A basic neural network with one hidden layer might perform okay with simple images, while a deep neural network with 10 or more layers can detect subtle curves, edges, and styles – making it far more accurate.

Deep learning is a type of AI where the system becomes more powerful by using many layers in a neural network. These layers help the system learn more complex patterns and make better decisions. *Figure 1.3* summarizes our discussion so far:

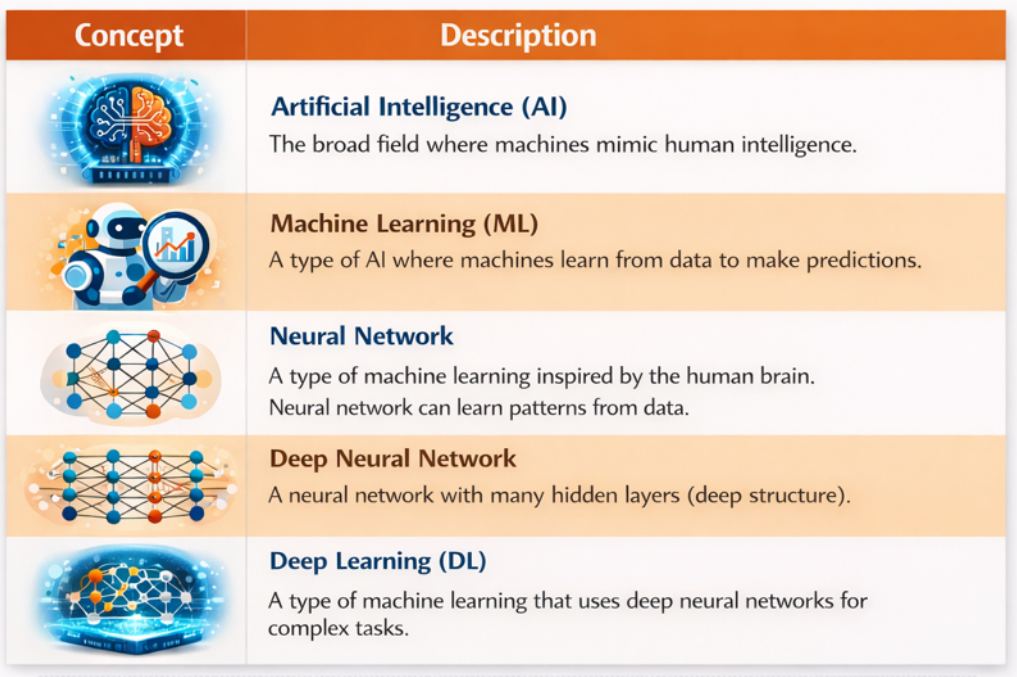


Figure 1.3: AI basics

So far, we have learned the fundamental concepts of AI and how AI systems learn from data. Let us now explore the different types of AI models and understand how they are used in real-world applications.

Types of AI models

As we discussed earlier, an AI model is a trained mathematical representation that has already learned patterns from data and can now make predictions, decisions, or classifications on new data. Following are some commonly used AI models.

Decision tree model

A **decision tree model** uses a tree-like structure of questions and answers to make decisions. It follows paths based on yes/no or true/false conditions until it reaches a final decision.

For example, consider a loan approval system with the following conditions:

- If income > ₹50,000 ⟶ Yes
- If credit score > 700 ⟶ Yes
- Else ⟶ No

Each path leads to an outcome like *Approve* or *Reject*.

Recognition AI model

These models are trained to recognize patterns, such as faces, objects, voices, or handwriting.

For example, facial recognition in a smartphone identifies unique features of a face and unlocks the phone only when it matches the stored image.

Personalization model

These models learn user behavior and preferences to deliver customized content or recommendations.

For example, an e-commerce website uses a personalization model to recommend products by analyzing your browsing and shopping behavior. It considers factors such as the products you searched for, the items you clicked on, and the preferences of users with similar interests to suggest products that you are more likely to purchase.

Natural language processing (NLP) model

NLP models help machines understand and respond to human language, both written and spoken.

For example, chatbots and virtual assistants.

Robotic model

These models are embedded in robots to help them sense, think, and act in the physical world. They may use AI for vision, movement, and decision-making. For example, a warehouse robot uses a robotic AI model to detect the correct package, plans the shortest and safest route to its destination, and picks up and delivers the package efficiently without human intervention.

Multimodal AI model

Multi-modal models can process and combine different types of data at once such as text, images, audio, and video. For example, a virtual tutor uses a multi-modal AI model to process different types of information simultaneously. It listens to your voice (audio), reads your typed questions (text), and analyzes your facial expressions (images) to better understand your needs before providing an appropriate response.

Large language models (LLMs)

The model is trained to understand and generate human language. It is called *large* because it is trained on massive amounts of text and has billions of parameters.

Common examples include, ChatGPT (Generative Pre-trained Transformer), BERT (used by Google Search for understanding queries), Claude, LLaMA, and Gemini.

Understanding different types of AI models is important from a governance perspective because each model can have different risks, uses, and control requirements.

Traditional ML models are usually developed for specific purposes, such as credit scoring or fraud detection. Governance mainly focuses on data quality, model validation, bias, explainability, performance, security, and model changes.

LLMs have broader capabilities and can be used for many purposes. They introduce additional risks such as hallucinations, prompt attacks, data leakage, inappropriate outputs, model misuse, copyright issues, and third-party dependencies.

Therefore, traditional ML governance provides a useful baseline, but additional governance controls may be needed for LLMs.

So far, we have explored different types of AI models and their applications. The next step is to understand how these trained models are deployed in real-world environments, as the choice of deployment affects performance, security, scalability, and governance.

AI deployment options

AI deployment options describe how and where a trained AI model is placed in real life. The deployment choice directly affects performance, cost, scalability, response time, security, and data privacy.

Organizations must select a deployment method based on factors such as data sensitivity, regulatory requirements, internet availability, real-time processing needs, and budget.

Common AI deployment options include cloud deployment, on-premises deployment, and edge deployment, each suited for different business and operational scenarios.

Cloud deployment

In **cloud deployment**, the AI model runs on remote servers provided by cloud service providers. This option is easy to scale, quick to implement, and works well for applications with large datasets or high user traffic. However, it requires reliable internet connectivity and may raise concerns about storing sensitive data on external servers.

For example, a retail company uses a cloud-based AI model to analyze customer data and recommend products on its website. The company can increase computing resources during high-traffic periods without purchasing additional servers.

On-premises deployment

In on-premises deployment, the AI model is hosted on the organization's own servers. This gives more control over data and system performance and is often preferred in industries with strict regulatory requirements. The downside is that it may require more technical expertise and higher setup costs.

For example, a bank hosts its AI-based fraud detection model on its own servers so that sensitive customer and transaction data remain within the organization's controlled environment.

Edge deployment

Edge deployment means running the AI model directly on local devices like mobile phones, IoT sensors, or manufacturing equipment. This method allows for real-time processing with low delay and works even in areas with limited internet access. It is ideal for tasks that require an instant response.

For example, face unlock and speech recognition on mobile devices use edge AI to process data directly on the device. Similarly, a smart camera can use edge AI to detect safety violations on a construction site.

Summary

This chapter laid out the foundation for understanding AI by explaining what it is, how it differs from traditional software systems, and why these differences matter. It introduced key concepts such as AI systems, data, algorithms, and models, and explained how machines learn through approaches like machine learning, neural networks, and deep learning. The chapter also explored common types of AI models and deployment options, highlighting how AI systems are designed, trained, and used in practice. Most importantly, it emphasized that AI systems can learn, adapt, and change their behavior over time, which introduces new risks that do not exist in traditional IT systems. This understanding is essential before discussing governance, because effective AI governance begins with a clear appreciation of how AI systems work, where risks originate, and why continuous oversight is necessary. With this foundation in place, the next chapter focuses on AI governance and how organizations can responsibly manage, control, and oversee AI systems throughout their lifecycle.

In the next chapter, we will build on these AI fundamentals by exploring the concept of governance and understanding why effective governance is essential for managing AI responsibly. We will examine the key elements of governance, the roles and responsibilities involved, and how governance helps organizations align AI initiatives with business objectives, risk management, and regulatory requirements.

2

Introduction to AI Governance

In the last chapter, we discussed how AI based systems are different from traditional systems. In this chapter, we will discuss why traditional governance may not be sufficient for AI based systems. We will first understand what governance is and how it impacts the organization's resilient environment.

In this chapter, we will build on the AI fundamentals discussed in the previous chapter and explore why effective governance is essential for the responsible use of AI. We will examine the key elements of governance, how governance differs from management and operations, and how organizations can establish structures, policies, roles, and measures to oversee AI effectively.

Understanding governance

In simple terms, governance is about direction, control, and accountability. It defines who is responsible for what, how decisions are made, and how actions are monitored. Governance ensures that an organization does not operate randomly or based only on individual decisions, but follows approved rules, structures, and objectives.

In the context of technology, and especially AI, governance becomes critical because AI systems can impact people, businesses, and society at scale. Decisions made by AI may not always be transparent. Risks such as bias, misuse, privacy violations, and legal non-compliance are real and growing.

Without proper governance, even well-designed AI systems can lead to ethical issues, regulatory penalties, reputational damage, and loss of trust. To address these risks, organizations need a structured approach to AI governance. The next section introduces the key elements of AI governance including roles and responsibilities, policies, decision-making, oversight, risk management, and accountability.

Elements of governance

Governance is built on a set of key elements that work together to ensure accountability, control, and responsible decision-making. The following are the key elements of effective governance. These elements apply to organizations in general and are especially important for technology and AI.

- **Governance committee**: This committee is a formal body responsible for overseeing and guiding key decisions related to governance. It provides direction, oversight, and accountability at an organizational level. The committee typically includes representatives from business, technology, risk management, legal, compliance, and information security. This cross-functional composition ensures that decisions consider technical feasibility, business value, risk, ethics, and compliance.
- **Documented roles and responsibilities**: Governance defines clear roles, responsibilities, and ownership. It establishes who is accountable for decisions, outcomes, and risks. For example, the AI business owner may be responsible for business outcomes, the DPO for data protection, the compliance/legal team for regulatory requirements, and the CISO for AI security risks. These responsibilities should be formally documented.
- **Documented policies and procedures**: Documented policies and procedures form the backbone of effective governance. They clearly define the rules, expectations, and processes that guide how decisions are made and how activities are performed. *Policies* describe what must be followed and why it is important, while *procedures* explain how those policies are implemented in practice. Together, they ensure consistency, accountability, and compliance across the organization.
- **Decision-making structure**: Governance defines how decisions are made, approved, escalated, or rejected, ensuring consistency and control. For example, an AI governance committee reviewing high-risk AI use cases before deployment may approve the use case if risks are within accepted limits, require additional controls if risks need to be addressed, or escalate the decision to senior management when significant risks remain.
- **Risk management**: Governance ensures risks are identified, assessed, and managed, including operational, legal, ethical, and reputational risks. For example, performing periodic risk assessments to identify and evaluate risks such as bias, privacy breaches, security threats, inaccurate outputs, and regulatory non-compliance, and taking appropriate actions to address identified risks.
- **Compliance management**: Compliance management ensures that an organization follows applicable laws, regulations, standards, and internal requirements. It provides

confidence that activities are conducted within defined legal and ethical boundaries. Effective compliance management reduces regulatory risk, avoids penalties, and strengthens trust among customers, regulators, and other stakeholders. It also supports responsible and sustainable use of AI within the organization.

- **Audit**: Audit is a key element of governance that provides independent assurance that policies, procedures, and controls are working as intended. It helps verify whether the organization is complying with defined requirements and whether risks are being effectively managed. The objective of audit is not only to check compliance, but also to identify gaps, weaknesses, and areas for improvement. Audit provides management with confidence that AI systems are being used responsibly and in line with legal, ethical, and organizational expectations. We have discussed audit in *Chapter 17*.

Now that we have discussed governance in detail, let us understand how different it is from management and operations.

Governance versus management versus operations

To understand governance properly, it is important to clearly distinguish it from management and operations. These three concepts are closely related but serve different purposes within an organization.

Governance focuses on direction, oversight, and accountability. It defines what should be done, why it should be done, and who is accountable. *Management* focuses on planning and coordination. It translates governance decisions into plans, processes, and resources. *Operations* focus on execution and day-to-day activities. A simple way to understand the difference is:

- Governance sets the rules and boundaries
- Management plans and organizes within those boundaries
- Operations carry out the actual work

For example, the governance committee sets a rule requiring human review for high-risk decisions. Management develops the implementation plan, assigns responsibilities to the relevant teams, and sets timelines for meeting these requirements. Operations then configure the AI system, implement the required controls, monitor its performance, and refer high-risk decisions to human reviewers.

Effective governance requires a clear separation and alignment between governance, management, and operations. When these roles are confused, organizations may face weak oversight, unmanaged risks, or inconsistent decision-making. When they are clearly defined and well-aligned, AI systems can be deployed responsibly, efficiently, and with confidence. Let us now understand how governance and regulations are connected.

Governance and regulations

Governance and regulations are closely connected, but they are not the same. *Regulations* are external requirements imposed by laws, authorities, or regulators, while *governance* is the internal system an organization uses to ensure that those requirements are understood, implemented, and monitored.

Governance translates legal and regulatory obligations into internal policies, procedures, roles, and controls. Through governance, organizations decide how compliance will be achieved, who is responsible, and how adherence will be monitored.

Governance also helps organizations manage regulatory change. As AI regulations evolve, governance structures such as committees, compliance functions, and review mechanisms ensure that new requirements are assessed, implemented, and communicated across the organization.

Effective governance depends on clear and well-documented requirements. Governance documents help translate organizational objectives, regulatory requirements, and management expectations into consistent practices that employees can understand and follow. The three key governance documents—policies, procedures, and guidelines—serve different purposes and together provide a structured approach to implementing governance.

Policies, procedures, and guidelines

Effective governance relies on a structured set of documents that define expectations and control how activities are performed. Policies, procedures, and guidelines each serve a distinct purpose, and understanding the difference between them is essential for practical governance.

Policies set the overall intent and direction. They define what the organization expects and why it is important. *Procedures* describe how activities are carried out in practice. They provide step-by-step instructions to implement policies consistently. *Guidelines* provide recommended practices and flexibility. They support policies and procedures by offering advice on how to apply them in different situations.

The following example shows how policies, procedures, and guidelines work together in practice:

- **Policy statement** - AI systems should be periodically reviewed and updated to ensure they remain accurate, reliable, compliant, and aligned with organizational objectives.

- **Procedures** - To implement this policy, the organization could establish the following procedures:
    - The review shall assess model performance, accuracy, bias, drift, compliance with applicable policies and regulations, and alignment with business objectives.
    - Findings from the review shall be documented, approved by the appropriate authority, and tracked to closure.
    - Any approved updates to the AI system shall be implemented in a controlled manner, followed by validation and continued monitoring.
- **Guidelines** - The following guidelines can support the implementation of these procedures:
    - Cross-functional involvement from business, risk, legal, and technical teams is recommended.
    - Mapping with NIST AI Risk Management Framework is recommended.

Together, these documents create a layered governance structure. When clearly defined and well-aligned, they help organizations manage AI responsibly, consistently, and in line with regulatory and ethical expectations.

Many organizations already have IT governance in place and may wonder why they need separate AI governance. While AI governance builds on IT governance, AI systems have additional risks that require specific oversight. Let us now understand the key differences between AI governance and IT governance.

Differentiating AI governance and IT governance

Many organizations may think that there is no need for separate AI governance because they already have IT governance in place. It must be noted that though AI governance and IT governance are related, they are not the same.

Traditional IT governance focuses on controlling information technology systems to ensure reliability, security, cost-effectiveness, and alignment with business objectives. It deals with areas such as infrastructure, applications, data centers, cybersecurity, change management, and service availability.

AI governance goes beyond traditional IT governance. While it includes many IT governance principles, it also addresses risks and challenges that are unique to AI systems. AI systems can make autonomous or semi-autonomous decisions, learn from data, and change behavior over time. This creates new governance concerns related to ethics, bias, transparency, explainability, accountability, and societal impact. We will discuss each of these AI principles in the subsequent chapter.

In IT governance, systems usually behave in a predictable and deterministic manner. Once implemented and tested, their behavior remains largely stable unless changes are made. In contrast, AI systems may evolve due to model retraining, data drift, or changes in the operating environment, requiring continuous oversight and review.

Another key difference is the impact of scope. IT governance mainly focuses on organizational impact, such as system availability or data security. AI governance must also consider broader impacts on individuals, customers, and society, including fairness, discrimination, and trust.

AI governance therefore requires additional elements such as AI ethics policies, impact assessments, human oversight mechanisms, model monitoring, and cross-functional governance committees. While IT governance provides a strong foundation, AI governance extends it to ensure that AI systems are used responsibly, transparently, and in alignment with legal, ethical, and societal expectations. Let us now understand why AI needs stronger governance as compared to traditional systems.

Why AI needs stronger governance than traditional systems

As we discussed earlier, traditional systems generally behave in a predictable manner. Once designed, tested, and deployed, their behavior remains largely stable unless explicitly changed. AI systems, on the other hand, learn from data and may change their behavior over time due to model updates, retraining, or data drift. This makes continuous oversight and review essential.

Other important reason is that AI systems often operate with limited transparency. Decisions made by AI models may be difficult to explain, even to their developers. This lack of explainability creates challenges for accountability, regulatory compliance, and trust, especially when AI decisions affect individuals' rights, access to services, or employment opportunities.

Also, the regulatory environment for AI is rapidly evolving. New laws and standards impose specific governance, documentation, and oversight requirements for AI systems. Strong AI governance helps organizations anticipate regulatory expectations, demonstrate compliance, and adapt to change.

For these reasons, AI governance must be more robust, continuous, and multidisciplinary than traditional system governance. It must combine technical controls with ethical oversight, legal compliance, and clear accountability to ensure responsible and trustworthy use of AI.

Governing AI-based systems

Governing AI-based systems means putting in place structures, rules, and oversight mechanisms to ensure that AI is designed, used, and managed responsibly. Unlike traditional systems, AI-based systems can learn from data, change behavior over time, and make decisions that directly affect people and organizations. This makes governance essential throughout the entire AI lifecycle.

Effective governance of AI-based systems starts at the idea stage. Before an AI system is developed or deployed as a service, its purpose, expected benefits, risks, and potential impact should be assessed. High-risk use cases may require additional approvals, ethical review, or regulatory assessment before proceeding.

During development and deployment, governance ensures that policies related to data quality, privacy, security, fairness, and transparency are followed. Documentation plays a critical role by providing evidence of decisions, assumptions, and controls applied.

Once deployed, AI-based systems require continuous oversight. Performance, bias, drift, and compliance must be regularly monitored, not assumed to remain stable. Governance mechanisms such as periodic reviews, audits, and reporting help detect issues early and ensure timely corrective action.

Governing AI-based systems also requires clear accountability. Roles and responsibilities must be defined across business, technical, risk, and compliance functions. Human oversight should be maintained, especially for high-impact or sensitive decisions.

AI governance — roles and responsibilities

Clear roles and responsibilities are essential for effective AI governance. AI systems involve multiple stakeholders, and without defined ownership, accountability gaps can arise, leading to unmanaged risks and weak oversight. As an AI GRC professional, you need to understand the following governance-related roles:

- **Board and senior management**: The board and senior management provide overall direction and oversight for AI governance. They approve the AI governance framework, define risk appetite, ensure alignment with organizational strategy, and hold management accountable for responsible and compliant use of AI.
- **AI governance committee**: The AI governance committee oversees AI-related decisions and risks across the organization. They generally report to the board and have responsibility to ensure that board-approved policies are implemented. It reviews and approves AI use cases, especially high-risk ones, monitors compliance and risk

reports, ensures ethical considerations are addressed, and tracks corrective actions and improvements.

An AI governance committee should be cross-functional so that decisions consider business value, technical feasibility, risk, ethics, and regulatory compliance.

- **Business and use case owners:** Business and use case owners are responsible for defining the purpose, scope, and expected outcomes of AI systems. They ensure that AI use aligns with business objectives and are accountable for the impact of AI decisions in their respective areas.
- **Data science and technical teams:** Data science and technical teams design, develop, test, deploy, and maintain AI systems in accordance with approved policies and procedures. They are responsible for data quality, model performance, documentation, monitoring, and technical issue resolution.
- **Risk management function:** The risk management function identifies, assesses, and monitors AI-related risks, including operational, ethical, legal, and reputational risks. It supports risk assessments, defines controls, and reports risk status to the governance committee and senior management.
- **Legal and compliance function:** The legal and compliance function ensures that AI systems comply with applicable laws, regulations, and contractual requirements. It advises on regulatory obligations, supports compliance assessments, and assists in responding to regulatory inquiries or changes.
- **Information security and privacy function:** The information security and privacy function ensures that AI systems meet security and data protection requirements. This includes safeguarding data, managing access controls, supporting privacy impact assessments, and addressing security incidents related to AI.
- **Internal audit:** Internal audit provides independent assurance on the effectiveness of AI governance. It evaluates whether policies, procedures, and controls are operating as intended and recommends improvements where gaps are identified.

Once roles and responsibilities are clearly defined, organizations also need documented policies to guide and control AI-related activities. These policies help put governance into practice and ensure that everyone understands what is expected of them.

AI governance — policies

As we discussed earlier, documented policies and procedures are important elements of governance. AI governance policies define the organization's intent, expectations, and boundaries for the use of artificial intelligence. They provide clear direction on how AI systems should be designed, developed, deployed, and used in a responsible and controlled manner.

They clarify what is acceptable and unacceptable use of AI, the level of risk the organization is willing to accept, and the responsibilities of different stakeholders involved in AI activities.

Policy sets the direction, while detailed implementation is handled through procedures and guidelines.

AI governance policies should not be static documents. They must be regularly reviewed and updated to reflect changes in technology, regulations, business objectives, and lessons learned from incidents or audits. Clear communication and awareness are also essential so that employees understand and follow these policies in practice.

By establishing strong AI governance policies, organizations create a foundation for responsible AI use, regulatory compliance, and long-term trust among customers, regulators, and other stakeholders.

As an AI GRC professional, one of the most important tasks is to create, maintain, implement, and monitor the AI governance policy. This policy acts as the foundation for responsible, compliant, and controlled use of AI across the organization.

Note

For your reference, a ready-to-use AI governance policy template is provided as *Template 1* and is available in the GitHub repository of this book. The template is designed for practical implementation and can be adapted to suit different organizational contexts. To access the repository link, follow the steps in the *Download the example code files* section in the *Preface*.

The provided template is fully aligned with the requirements of the ISO/IEC 42001 standard and supports organizations in establishing a structured and compliant AI governance framework.

Governance needs to be measured to know whether it is working effectively. Organizations cannot improve what they do not measure. Regular measurement helps an organization understand whether its governance policies, processes, and controls are being followed as expected. It also helps identify gaps, track progress, and highlight areas where corrective action or improvement is needed. This makes governance a measurable and continuously improving process rather than just a set of policies and procedures.

AI governance – KPIs

Key performance indicator (KPI) is a measurable value used to track how effectively an organization is achieving a specific objective. KPIs play an important role in measuring the

effectiveness of AI governance. KPIs help organizations understand whether AI governance policies, processes, and controls are being implemented as intended and whether they are achieving the desired outcomes.

In the context of AI governance, KPIs focus not only on performance, but also on compliance, risk management, ethics, and accountability. Well-defined KPIs provide visibility into how AI systems are governed across their lifecycle and help identify areas that require improvement.

AI GRC professionals typically work with the AI governance committee and senior management when defining KPIs. They first identify organizational objectives, key AI risks, and relevant controls, and then propose measurable KPIs. These KPIs are discussed with relevant stakeholders, such as business, risk, compliance, security, and the AI governance committee. The appropriate management or governance body approves the KPIs, thresholds, and reporting frequency. The AI GRC function then monitors the KPIs, reports results, and highlights areas requiring corrective action.

By defining and monitoring appropriate KPIs, organizations can move AI governance from a theoretical concept to a measurable and continuously improving practice. The following table provides examples of AI GRC related KPIs:

KPI Area	KPI Description	What It Measures	Threshold
AI Inventory	Percentage of AI systems registered in the AI inventory	Coverage of AI systems under governance	≥ 95%
Governance Oversight	Percentage of AI use cases reviewed and approved by the AI governance committee	Effectiveness of approval and oversight mechanisms	100%
Risk Management	Percentage of high-risk AI systems with completed risk or impact assessments	Adequacy of risk assessment for higher-risk AI	100%
Review and Monitoring	Frequency of AI system reviews as defined by policy	Compliance with review and update requirements	≥ 90% on-time reviews

KPI Area	KPI Description	What It Measures	Threshold
Incident Management	Number of AI-related incidents reported	Exposure to AI-related risks and control gaps	0 critical incidents
Issue Management	Average time taken to close AI governance issues	Efficiency of corrective action processes	≤ 30 days
Model Oversight	Percentage of AI models monitored for bias and drift	Ongoing monitoring of AI behavior and performance	≥ 95%
Awareness and Training	Percentage of employees trained in AI governance and responsible AI use	Level of organizational awareness and adoption	≥ 90%

Table 2.1: AI KPIs

Note

For your ready reference, a ready-to-use set of AI governance KPI templates is provided as *Template 2* in this book's GitHub repository. The template is designed for practical implementation and can be tailored based on the organization's size, risk appetite, and AI maturity level. To access the repository link, follow the steps in the *Download the example code files* section in the *Preface*.

Summary

In this chapter, we explored the concept of governance and its importance in the context of AI-based systems. We discussed what governance means in practice, its key elements, and how it differs from management and operations. We also examined the relationship between governance and regulations, the role of policies, procedures, and guidelines, and the distinction between AI governance and traditional IT governance.

We then explored why AI requires stronger governance than traditional systems and how AI-based systems should be governed throughout their lifecycle. We also looked at the

importance of clearly defined roles, responsibilities, policies, and KPIs in helping organizations control, monitor, and use AI responsibly.

While governance defines the structure, oversight, and accountability for AI, it must be guided by clear principles. In the next chapter, we will explore the principles of responsible AI, which provide the ethical and practical foundation for designing, deploying, and using AI systems responsibly and sustainably.

3

Principles of Responsible AI

In the previous chapter, we discussed how AI governance provides structure, oversight, and accountability for the use of AI within an organization. **Governance** defines how AI should be controlled. However, governance alone is not sufficient unless it is guided by clear, well-defined principles.

Responsible AI means developing and using artificial intelligence in ways that are ethical, trustworthy, and safe. As AI becomes more widely used in areas such as healthcare, finance, employment, and customer services, its decisions and outputs can have a significant impact on individuals and organizations. Responsible AI principles help organizations identify and address concerns such as unfairness, lack of transparency, privacy risks, security issues, and potential harm. An AI system may be technically capable or intelligent, but that does not automatically make its decisions responsible or trustworthy.



Figure 3.1: Responsible AI — intelligence is not enough

Responsible AI principles help bridge this gap by guiding organizations to consider not only what AI can do, but also how it should be designed, developed, deployed, and used. While governance establishes the structure, responsibilities, policies, and controls, Responsible AI principles provide the direction for making AI decisions in a fair, transparent, safe, and accountable manner.

In this chapter, we will explore the key principles of Responsible AI and understand what each principle means in practice. We will also learn who defines Responsible AI principles and how these principles guide the development, deployment, and use of AI systems in a responsible manner.

> **Note**
>
> **Download the code bundle and the PDF version of this book**
>
> Your purchase includes a DRM-free PDF copy of this book, the code bundle, and a range of exclusive benefits. To unlock everything, follow the *Free benefits with your book* section in the *Preface*.

Who defines Responsible AI principles

Now, the question arises: who has the authority to define what Responsible AI principles are? There is no single global authority that exclusively defines Responsible AI principles. Instead, these principles are shaped collectively by regulators, international standards bodies, governments, industry groups, and organizations themselves.

Regulatory authorities define mandatory requirements through laws and regulations (for example, EU AI Act, the US Executive Order on Safe, Secure, and Trustworthy AI, Canada's Artificial Intelligence and Data Act (AIDA), etc.). International standards bodies provide structured and widely accepted frameworks (for example, ISO/IEC 42001 for Artificial Intelligence Management Systems, the NIST AI Risk Management Framework, etc.). Industry groups and academic institutions contribute best practices and ethical guidance based on real-world use and research (OECD principles on trustworthy AI, the World Economic Forum guidelines on responsible and trustworthy AI). We have discussed all important AI governance frameworks and standards in later chapters of this book.

Before we discuss each Responsible AI principle in detail, let us have a look at the Responsible AI principles. They are derived from the above-mentioned widely accepted frameworks, standards, and regulations.

- Accountability
- Fairness and non-discrimination

- Transparency
- Explainability
- Privacy and data protection
- Security and resilience
- Safety and reliability
- Human oversight
- Risk-based approach
- Ethical and responsible use
- Continuous monitoring and improvement
- Compliance with laws and standards

At an organizational level, senior management and the board are responsible for adopting and defining Responsible AI principles that align with applicable laws, standards, and organizational values. These principles are then operationalized through governance frameworks, policies, procedures, and controls by AI GRC professionals.

Therefore, Responsible AI principles are not arbitrary or self-defined. They are derived from a combination of regulatory expectations, recognized standards, and ethical considerations, and relevant and applicable principles are implemented within organizations through strong AI governance.

Responsible AI principles

In this section, we will explore the key principles of Responsible AI and understand what each principle means in practice by examining the key best practices of each principle. These principles form the foundation on which effective AI governance is implemented and provide guidance for building trustworthy, compliant, and sustainable AI systems.

Accountability

Company A is a technology firm that develops a **large language model** (**LLM**) as a general-purpose AI system. Company B is a fintech company that uses Company A's LLM to build a loan approval solution. Company C is a bank that deploys the AI-based loan approval system in its production environment. The bank uses the system to make loan decisions that directly affect customers. Over time, several customers experience repeated loan rejections. One customer files a lawsuit alleging unfair and incorrect rejection, claiming financial harm and discrimination.

Now the question is who will be held responsible. Though a lot of other variables and input are required to decide on the above case study, it emphasizes the reason why the principle of

accountability is given utmost importance in AI governance. Responsible AI governance ensures that these accountabilities are clearly defined through contracts, policies, risk assessments, and oversight mechanisms before AI systems are deployed. AI may be complex and involve multiple parties, but accountability must always remain clear, traceable, and human-centered.

Accountability means that there must always be a clearly identified person or role responsible for an AI system and its outcomes. AI systems do not operate on their own. They are designed, trained, deployed, and monitored by people. Accountability ensures that responsibility for AI decisions, risks, and impacts cannot be shifted to the technology itself.

In practice, accountability requires clear ownership across the AI lifecycle. This includes defining who is responsible for approving AI use cases, validating models, ensuring compliance with laws and policies, monitoring performance, handling incidents, and taking corrective actions when things go wrong.

To understand accountability from a legal standpoint, the reader should refer to the consequences of misleading information provided by an AI chatbot deployed by an airline company. Let us now discuss accountability under the EU AI Act.

Accountability under EU AI Act

Though we will study the EU AI Act in detail in *Chapter 8*, at this stage it is important to understand how accountability is clearly structured in the EU AI Act. It follows a role-based accountability approach. Responsibility is assigned based on the role an organization plays in the AI value chain:

Role under the EU AI Act	Who they are	Key accountability
Provider	Entity that develops an AI system or places it on the EU market under its own name	Ensure AI system meets regulatory requirements, conduct risk management, ensure data governance, maintain technical documentation, demonstrate compliance
Deployer (user)	Entity that uses an AI system in real-world operations Example: Bank uses AI to approve or reject loan applications	Ensure lawful use, apply human oversight, monitor AI outcomes, use AI only for intended purposes, and protect individuals' rights

Role under the EU AI Act	Who they are	Key accountability
Importer	Entity bringing an AI system from outside the EU into the EU market.	Verify compliance, ensure required documentation and conformity markings are present.
Distributor	Entity making an AI system available in the EU supply chain Example: software reseller distributing AI solutions to EU customers	Ensure noncompliant systems are not distributed and take corrective action if risks are identified
General-purpose AI provider	Entity providing foundation models or general-purpose AI systems Example: company providing a large language model used by multiple vendors	Identify and mitigate systemic risks, provide technical documentation, and support downstream compliance

Table 3.1: Accountability under EU AI Act

If we apply the EU AI Act to the case study that we discussed, the probable outcome can be as follows:

Company	Role under the EU AI Act	Accountability in the lawsuit
Company A	General-purpose AI provider Developed a large language model (LLM) and provided it as a general-purpose AI	Accountable for documenting model limitations, known risks, and supporting downstream compliance Not directly responsible for loan rejection decisions unless defects or undisclosed risks in the LLM are proven

Company	Role under the EU AI Act	Accountability in the lawsuit
Company B	Provider (Used the LLM to design and market an AI-based loan approval system)	Accountable for ensuring the loan approval system meets EU AI Act requirements for high-risk AI, including risk management, data governance, testing, and documentation May share liability if the system design contributed to unfair or incorrect decisions
Company C (bank)	Deployer (user) (Deployed the AI system and used it to make real loan approval and rejection decisions)	Holds primary accountability because it uses the AI system in practice. Responsible for human oversight, lawful use, monitoring outcomes, handling customer complaints, and protecting individual rights.

Table 3.2: Accountability case study

Accountability — practical implementation

As an AI GRC professional, you need to have a basic understanding of how to implement the principle of accountability in practice. Accountability should not remain a theoretical concept; it must be embedded into governance structures, processes, and day-to-day operations. The following are some best practices for implementing accountability effectively:

- **Clearly define AI roles and ownership:** Assign clear ownership for every AI system across its lifecycle. This includes identifying a business owner, technical owner, and risk or compliance owner. Each role should have documented responsibilities and decision-making authority.
- **Maintain an AI inventory with ownership details**: Record all AI systems in an AI inventory. Each entry should include system purpose, risk classification, owners, approval status, and review frequency. This helps ensure accountability is traceable.
- **Define accountability in policies and procedures**: AI governance policies should clearly state who is accountable for AI outcomes, incidents, and compliance. Procedures should define how accountability is exercised during approvals, reviews, incidents, and corrective actions.

- **Define accountability clauses in contracts and third-party governance controls**: Accountability clauses in contracts with AI vendors and service providers should clearly specify responsibilities related to compliance, documentation, incident reporting, and support during investigations.
- **Support accountability through audits and reviews**: Conduct regular internal audits and independent reviews to verify that accountability roles are clearly defined and operating effectively. Audit findings should lead to corrective actions and continuous improvement.

By implementing these practices, AI GRC professionals can ensure that accountability remains clear, enforceable, and effective throughout the AI lifecycle.

Note

For your ready reference, a checklist for implementing all Responsible AI principles is provided as *Template 3* in this book's GitHub repository. The template is applicable for all practical implementation sections in this chapter.

To access the repository link, follow the steps in the *Download the example code files* section in the *Preface*.

Fairness and non-discrimination

A large e-commerce company proudly launched a new AI-powered hiring tool. It was designed to scan thousands of resumes and shortlist the most suitable candidates. Over time, it was noticed that the system was mostly rejecting female applicants for technical roles. The company had to pause the system, issue a public apology, and launch a full review.

The organization did not intend to discriminate, but the AI system learned patterns from historical data that reflected past hiring biases. This situation illustrates why fairness and non-discrimination are essential principles of Responsible AI.

Fairness means that AI systems should treat individuals and groups equitably and should not produce unjustified biased outcomes. *Non-discrimination* ensures that AI systems do not disadvantage people based on characteristics such as gender, age, ethnicity, religion, disability, or other attributes.

In practice, fairness requires organizations to actively identify and manage bias throughout the AI lifecycle. This includes reviewing training data for imbalance, testing models for biased outcomes, and monitoring decisions after deployment. A practical example is an AI-based loan approval system used by a bank. During testing, the bank finds that applicants from certain geographic areas are rejected more often. Further review shows that location data is indirectly

causing unfair outcomes. To address this, the bank modifies the model, applies fairness controls, and introduces human review for borderline cases. The system is then regularly monitored to ensure fair decisions across all groups.

By implementing fairness and non-discrimination, organizations reduce legal, ethical, and reputational risks while building trust with customers and regulators. Fair AI systems support equal opportunity and ensure that decisions are based on legitimate and justifiable criteria rather than unintended bias.

Fairness — practical implementation

As an AI GRC professional, you need to have a basic idea about how to implement this principle practically. The following are some of the best practices:

- **Perform bias risk assessment early:** Assess potential bias risks during the design stage of the AI system and identify sensitive attributes that may lead to unfair outcomes
- **Review and prepare training data:** Ensure training data is representative and relevant. Check for imbalance, historical bias, or missing data that could affect fairness
- **Implement human review for sensitive decisions:** Introduce human oversight for high-impact or borderline decisions, especially where individuals' rights or access to services may be affected
- **Monitor fairness continuously:** Regularly monitor AI outputs after deployment to detect emerging bias due to data drift or changes in usage patterns
- **Align with legal and ethical requirements:** Involve legal and compliance teams to ensure that fairness practices align with applicable anti-discrimination laws, regulations, and ethical expectations

By following these practices, organizations can reduce the risk of discriminatory outcomes and ensure that AI systems operate in a fair, responsible, and trustworthy manner.

Transparency

With an objective to increase the customer base, a company advertises that its customer support agents are available 24/7. Customers believe that they will be interacting with human support staff whenever they raise a query. In reality, the company has deployed AI-powered chatbots to handle most customer interactions. This is against the principle of transparency. When customers later discover that their queries were handled by AI bots rather than human agents, they feel misled. The issue is not the use of AI itself, but the lack of disclosure.

Transparency means being open and clear about the use of AI, its purpose, its scope, and its role in decision-making. It ensures that stakeholders are aware that AI is being used and understand where and how it influences decisions. Transparency does not always mean

revealing complex algorithms, but it does mean providing meaningful information about how decisions are made.

Transparency — practical implementation

Following are some of the best practices for practical implementation of the transparency principle:

- **Clearly disclose the use of AI:** Inform users and customers when AI systems are used, especially where AI interacts directly with people or influences decisions. Use stronger disclosure and transparency controls for high-risk or high-impact AI systems.
- **Be honest in communication and marketing:** Ensure that advertisements, user communications, and public statements do not misrepresent AI systems as human or conceal the use of AI.
- **Maintain an AI inventory:** Maintain a documented list of AI systems, including their purpose, scope, and usage to support internal visibility and governance.
- **Train employees on transparency obligations:** Ensure employees understand transparency requirements and communicate AI usage clearly and consistently.
- **Enable review and challenge mechanisms:** Provide mechanisms for users or customers to question, review, or appeal AI-driven decisions where applicable.

By implementing these practices, organizations can ensure that AI systems are transparent, accountable, and aligned with Responsible AI expectations.

Explainability

Imagine a customer is informed that their loan application has been rejected, but no reason is provided other than "the system decided so." The customer does not know whether the rejection was due to income, credit history, or some other factor. Even the bank staff struggle to explain the decision clearly. This lack of visibility creates confusion, frustration, and loss of trust. This is where the principle of explainability becomes important in dealing with such opaque or black box issues.

Explainability means the ability to understand and explain how an AI system arrives at a particular decision or outcome. It focuses on making AI decisions understandable to humans, especially when those decisions affect individuals' rights, opportunities, or access to services.

Explainability is different from transparency. Transparency is about disclosing the use of AI and its role, while explainability is about explaining the reasoning behind specific AI outcomes. An AI system can be transparent but still not explainable if its decisions cannot be understood or justified.

In practice, explainability requires organizations to select appropriate models or techniques that allow explanations, document decision logic, and provide meaningful reasons for AI-driven outcomes. The level of explainability required often depends on the risk and impact of the AI system.

A practical example is a bank using an AI-based credit scoring system that can explain why a loan was rejected by highlighting key factors such as credit score, income stability, or repayment history. This allows customers, regulators, and internal teams to understand and review the decision.

Explainability — practical implementation

Following are some of the best practices for practical implementation of the explainability principle. These practices help ensure that AI decisions can be understood, justified, and reviewed when required:

- **Test explanations for clarity and usefulness**: Regularly test whether explanations are clear, accurate, and useful for their intended audience
- **Enable review and challenge of decisions**: Provide mechanisms for reviewing and challenging AI-driven decisions, including human review and escalation paths
- **Train employees to explain AI decisions**: Train relevant employees so they can confidently explain AI decisions and limitations to customers and stakeholders
- **Maintain records to support audits**: Keep logs and records of decisions and explanations to support audits, investigations, and regulatory reviews

By implementing these practices, organizations can reduce reliance on black-box decisions and ensure that AI systems are explainable, accountable, and trustworthy.

Privacy and data protection

Imagine a company developing an AI system to analyze customer behavior. To improve model accuracy, the company uses large amounts of existing customer personal data without obtaining the required consent. Later, customers discover that their personal data was used without clear consent or proper safeguards. This results in legal action, reputational damage, and loss of customer trust. This situation highlights why the privacy and data protection principles are critical in Responsible AI.

Privacy and data protection mean that AI systems must handle personal and sensitive data lawfully, fairly, and securely. Organizations must ensure that data used for AI is collected for legitimate purposes, is limited to what is necessary, and is protected against unauthorized access or misuse.

Privacy is especially important because AI systems often require large amounts of data for training and operation. If personal or sensitive data is collected or used without proper

controls, organizations may face privacy violations, regulatory penalties, and loss of customer trust. Therefore, privacy should be considered throughout the AI lifecycle, from data collection and model development to deployment and monitoring.

For example, an organization developing an AI system to analyze customer behavior should first determine what personal data is required, why it is needed, and whether there is a valid legal basis for processing it. The organization should also apply appropriate safeguards to protect the data and define how long it will be retained.

This principle requires compliance with applicable data protection laws and regulations. Although Data Protection Officers are primarily responsible for this function, as an AI GRC professional, you should have a basic understanding of the following privacy principles:

- **Lawful, fair, and transparent processing:** Personal data must be collected and used based on a valid legal basis, treated fairly, and processed in a way that is open and clear to individuals
- **Consent:** Where required, personal data must be used only after obtaining clear and informed consent from individuals and they must have the ability to withdraw consent when applicable
- **Purpose limitation:** Data must be collected for specific, legitimate purposes and should not be used for unrelated activities without proper justification or consent
- **Data minimization:** Only the minimum amount of personal data required for the intended AI purpose should be collected and processed
- **Accuracy of data:** Personal data used in AI systems must be accurate and kept up to date to avoid incorrect or unfair outcomes
- **Storage limitation:** Personal data should not be stored longer than necessary and must be deleted or anonymized when no longer required
- **Integrity and confidentiality:** Personal data must be protected against unauthorized access, alteration, loss, or disclosure through appropriate security measures
- **Accountability for data processing:** Organizations must be able to demonstrate compliance with privacy principles through documentation, controls, and governance mechanisms

Security and resilience

Security means protecting AI systems, data, and models from unauthorized access, misuse, manipulation, and attacks. AI systems introduce new security risks, such as model theft, data poisoning, adversarial attacks, and unauthorized access to sensitive training data. Strong security controls are required to safeguard the confidentiality, integrity, and availability of AI systems.

Resilience means ensuring that AI systems continue to operate reliably even when errors, attacks, or unexpected conditions occur. This includes having backup mechanisms, fail-safe modes, incident response procedures, and recovery plans so that AI services can be restored quickly and safely.

Security and resilience are important because AI systems can be targeted by attackers and may also fail due to errors, unexpected conditions, or changes in the operating environment. A security incident or system failure can affect sensitive data, AI outputs, business operations, and customer trust. Therefore, organizations need appropriate security and recovery measures throughout the AI lifecycle.

Security and resilience — practical implementation

Though the **Chief Information Security Officer (CISO)** is primarily responsible for this function, as an AI GRC officer, you need to have a basic idea about the following best practices:

- **Include AI systems within information security scope:** Ensure AI applications, models, and data are covered under the organization's information security policies and controls
- **Implement strong access controls:** Restrict access to AI systems, training data, and models to authorized personnel only
- **Protect training and operational data:** Apply encryption, secure storage, and controlled data transfer mechanisms for sensitive AI data
- **Secure AI models and code:** Protect models and source code from theft, unauthorized modification, or leakage
- **Monitor for AI-specific attacks:** Implement monitoring to detect data poisoning, adversarial inputs, and unusual system behavior
- **Perform security testing on AI systems:** Conduct vulnerability assessments and security testing specific to AI applications and infrastructure
- **Integrate AI incidents into incident response plans:** Ensure incident response procedures include AI-related security events and recovery actions
- **Implement backup and recovery mechanisms:** Maintain backups of critical AI models, data, and configurations to enable quick restoration
- **Apply resilience and fail-safe mechanisms:** Design AI systems to safely fall back to manual or alternate processes when failures occur
- **Review security controls periodically:** Perform regular reviews and audits of AI security and resilience controls

Refer key frameworks and resources for AI security such as OWASP Top 10 for LLM Applications, OWASP Top 10 for Agentic Applications, Cloud Security Alliance (CSA) AI Safety/ Maturity Model, and MITRE ATLAS Framework.

Safety and reliability

Imagine a hospital deploying an AI system to support medical diagnosis. If the system produces incorrect results or behaves unpredictably, patients may receive wrong treatment, leading to serious harm. This example shows why the safety and reliability principle is critical in Responsible AI.

Safety means ensuring that AI systems do not cause unintended harm to people, organizations, or society. AI systems must be designed and tested to behave within defined limits, especially in high-impact environments such as healthcare, finance, transportation, or critical infrastructure.

Reliability means that AI systems perform consistently and as intended over time. They should deliver accurate results, maintain stable performance, and behave predictably under normal and exceptional conditions.

Safety and reliability — practical implementation

Following are some of the best practices to implement this principle:

- **Define clear performance and safety thresholds:** Establish acceptable accuracy, error rates, and performance limits before deploying AI systems
- **Conduct thorough testing before deployment:** Test AI systems under normal and extreme conditions to identify potential failures or unsafe behavior
- **Implement human oversight for high-impact decisions:** Ensure critical or high-risk AI decisions are subject to human review or intervention
- **Use controlled and phased deployment**: Deploy high-risk AI systems gradually to monitor behavior before full-scale rollout
- **Monitor performance continuously:** Track accuracy, error rates, drift, and system behavior to detect reliability issues early
- **Implement safe fallback mechanisms:** Ensure AI systems can safely switch to manual or alternative processes if failures occur
- **Log and analyze errors and failures:** Record incidents, unexpected outcomes, and near misses to improve system safety

- **Review and update models periodically:** Retrain and update AI systems to maintain accuracy and reliability over time
- **Conduct independent validation and audits:** Perform periodic validation and assurance reviews for high-impact AI systems

Human oversight

Human oversight means that AI systems should not operate without appropriate human supervision, especially when decisions affect individuals' rights, opportunities, or access to essential services. It ensures that humans remain in control of AI systems and can review, intervene, or override AI decisions when necessary.

AI systems may operate at different levels of automation. Some require humans to review every decision, while others need human intervention only in exceptional or high-risk cases. The level of oversight should be determined based on the risk and impact of the AI system.

Human oversight is important because AI systems may produce incorrect, unexpected, or unfair outcomes. Having appropriate human involvement helps organizations identify such issues, take corrective action, and prevent decisions from being made solely by an AI system when human judgment is necessary.

For example, an AI system used to screen loan applications may automatically assess applications, but a human reviewer can examine cases where the system identifies uncertainty or where the decision could significantly affect an individual. This allows the organization to benefit from AI while ensuring that important decisions remain subject to appropriate human judgment.

Human oversight — practical implementation

As an AI GRC professional, you should ensure that appropriate human oversight is defined and implemented based on the risk and impact of each AI system. The following are some practical measures:

- Define when human review is required, particularly for high-risk or high-impact decisions
- Assign clear responsibility to individuals who review, approve, or override AI decisions
- Provide appropriate training so reviewers understand the AI system, its limitations, and when intervention is necessary
- Maintain records of significant human interventions and overrides for review and audit purposes
- Periodically assess whether the level of human oversight remains appropriate as the AI system, its use, or its risks change

Risk-based approach

A risk-based approach means applying governance, controls, and oversight proportionate to the level of risk and impact of an AI system. Not all AI systems require the same level of scrutiny. High-risk AI systems need stronger controls, deeper assessments, and continuous oversight, while low-risk systems can be managed with lighter controls.

In practice, a risk-based approach requires classifying AI systems based on factors such as their impact on individuals, legal and regulatory exposure, ethical concerns, and operational criticality. Based on this classification, organizations can define appropriate requirements for risk assessments, approvals, testing, documentation, monitoring, and human oversight.

A risk-based approach is important because applying the same level of controls to every AI system may not be practical or effective. Organizations can focus greater attention and resources on AI systems that present higher risks, while applying proportionate controls to lower-risk systems.

For example, a company may use AI to filter spam emails and AI to assess job applications. The spam filter is low-risk and may require only basic testing and monitoring. In contrast, the hiring system is high-risk and requires bias assessments, explainability, human review, and governance committee approval before deployment.

Risk-based approach — practical implementation

As an AI GRC professional, you should ensure that AI systems are classified and governed according to their level of risk. The following are some practical measures:

- Classify AI systems based on their potential impact and risk
- Define different approval and assessment requirements for different risk levels
- Perform detailed risk assessments for high-risk AI systems
- Apply stronger testing, monitoring, documentation, and human oversight to high-risk systems
- Periodically review the risk classification as the AI system, its use, or its operating environment changes.

We have discussed AI risk assessment in detail in *Chapter 10*.

Ethical use

An e-commerce website deploys an AI system that offers personalized prices to customers based on their spending habits and purchase history. The system identifies customers who are likely to pay higher prices and automatically increases prices for them, while offering lower prices to others for the same product. Ethical AI use requires organizations to avoid such hidden and exploitative practices.

Ethical use of AI means ensuring that AI systems are applied in ways that align with fundamental human rights. AI should be used to create benefit and value, not to exploit, manipulate, or cause harm to individuals or groups. Ethical use goes beyond legal compliance. An AI system may be legally permissible but still ethically questionable. This principle encourages organizations to ask not only, *"Can we do this?"* but also, *"Should we do this?"*

Ethical considerations are important because AI systems can influence people, make decisions, and affect individuals or groups in ways that may not always be obvious. Organizations should therefore consider the potential ethical impact of an AI use case before deployment and continue to review it as the system and its use change.

For example, an organization considering an AI system for personalized pricing should assess whether the system could unfairly disadvantage certain customers or exploit their individual circumstances. The organization should consider these concerns before approving the use case, rather than relying only on whether the practice is legally permitted.

Ethical use — practical implementation

As an AI GRC professional, you should ensure that ethical considerations are included when AI use cases are evaluated and monitored. The following are some practical measures:

- **Involve an ethics officer**: Involve the Ethics Officer in overseeing ethical AI practices, advising on sensitive AI use cases, and supporting the governance committee in ethical decision-making
- **Define acceptable and unacceptable AI use cases**: Clearly document which AI applications are permitted, and which are prohibited, based on ethical considerations
- **Include diverse stakeholders in AI decision-making**: Involve business, legal, risk, compliance, and ethics representatives when evaluating AI use cases
- **Provide escalation mechanisms for ethical concerns**: Allow employees and users to report potentially unethical AI practices without fear of retaliation
- **Review ethical practices periodically**: Reassess AI use cases over time to ensure continued alignment with ethical expectations

Continuous monitoring and improvement

Imagine a company deploying an AI system to detect fraudulent transactions. Initially, 90% of transactions are performed through the web application, and the AI model is trained mainly on web-based customer behavior. Over time, customers gradually shift to using mobile applications. As user behavior changes, the AI system starts missing new fraud patterns specific to mobile transactions.

If the organization does not continuously monitor and update the AI system, fraud detection accuracy may decline, leading to financial losses. This situation highlights why continuous monitoring and improvement are essential in Responsible AI.

Continuous monitoring means regularly tracking the performance, behavior, and impact of AI systems after deployment. AI systems are not static and may degrade over time because of data drift, changes in the operating environment, or emerging risks. Ongoing monitoring helps organizations identify issues such as reduced accuracy, new biases, security threats, and unexpected outcomes.

Continuous monitoring is important because an AI system that performs well when initially deployed may not continue to perform in the same way as data, user behavior, or business conditions change. Regular monitoring helps organizations identify these changes early and take corrective action.

For example, if customers increasingly use a mobile application for transactions, the organization should assess whether the fraud detection model continues to identify mobile-related fraud patterns effectively. If performance declines, the model may need to be retrained using relevant data and subjected to appropriate testing before being put back into use.

Continuous monitoring and improvement — practical implementation

As an AI GRC professional, you should ensure that AI systems are regularly monitored and reviewed throughout their lifecycle. The following are some practical measures:

- **Define monitoring metrics**: Define performance, risk, and compliance metrics to track AI system behavior and effectiveness over time
- **Perform periodic AI system reviews**: Conduct scheduled reviews to assess whether AI systems remain aligned with business objectives, risk levels, and regulatory requirements
- **Monitor for data and model drift**: Monitor changes in data and AI system performance that may affect the reliability of AI outputs
- **Update governance controls based on lessons learned**: Refine policies, procedures, and controls based on monitoring results, incidents, and other findings
- **Take corrective action**: Retrain, update, or modify AI systems when monitoring identifies significant performance, risk, or compliance issues

Compliance with laws and standards

Compliance means ensuring that AI systems adhere to applicable laws, regulations, contractual obligations, and recognized standards. AI is increasingly subject to specific regulatory requirements, and organizations must ensure that their AI systems meet these obligations throughout the AI lifecycle.

Compliance is important because failure to meet legal and regulatory requirements can result in penalties, legal disputes, financial losses, and damage to the organization's reputation. AI systems may also be subject to different requirements depending on the industry, type of data used, purpose of the system, and countries in which the system operates.

For example, an organization using AI to process customer personal data should identify the applicable data protection requirements and ensure that appropriate privacy controls are implemented. If a new AI regulation becomes applicable, the organization should assess its impact and update its policies, procedures, and controls accordingly.

Compliance with laws and standards – practical implementation

As an AI GRC professional, you should ensure that applicable legal, regulatory, contractual, and standards-based requirements are identified and addressed throughout the AI lifecycle. The following are some practical measures:

- **Identify applicable AI laws and regulations**: Identify all relevant AI-related laws, data protection regulations, industry regulations, and contractual obligations that apply to AI systems
- **Map regulatory requirements to governance controls**: Translate legal and regulatory requirements into internal policies, procedures, and control measures
- **Define compliance responsibilities**: Clearly assign compliance roles across business, legal, risk, and AI governance functions
- **Monitor regulatory changes**: Track new and evolving AI laws, standards, and regulatory guidance to ensure timely updates to governance frameworks
- **Conduct periodic compliance reviews and audits**: Perform regular compliance checks and audits to verify adherence to laws and standards

Note

For your ready reference, a checklist for implementing all the above-mentioned Responsible AI principles is provided as *Template 3* in this book's GitHub repository. To access the repository link, follow the steps in the *Download the example code files* section in the *Preface*.

Summary

In this chapter, we learned about the principles of Responsible AI that guide the ethical, trustworthy, and safe use of AI. We saw how these principles form the foundation for effective AI governance and help ensure that AI systems respect human rights, avoid harm, and operate in a fair and transparent manner.

We learned how accountability establishes clear ownership for AI decisions and outcomes, while fairness helps identify and address bias. We also explored how transparency and explainability make AI systems and their decisions easier to understand and evaluate. Privacy and data protection, security and resilience, and safety and reliability help protect AI systems, their data, and the people affected by them. Human oversight, a risk-based approach, ethical use, continuous monitoring, and compliance further help organizations manage AI-related risks and ensure that AI systems continue to operate responsibly throughout their lifecycle.

Together, these principles provide the foundation for trustworthy AI governance and help organizations ensure that AI systems are developed and used in a fair, transparent, secure, accountable, and responsible manner. However, the rapid growth of AI has also raised significant legal, societal, and economic concerns worldwide.

In the next chapter, we will explore global trends in AI adoption, emerging risks, and why AI regulation has become a necessity rather than an option. We will also examine key international regulatory developments that are shaping the future of AI governance.

4

Overview of Global AI Governance Frameworks – ISO/IEC 42001 and EU AI Act

Around the world, governments, international organizations, and bodies responsible for setting technological standards have introduced frameworks and regulations to guide responsible and compliant AI use. Global AI frameworks for organizations may operate across different countries and regions, each with their own requirement. Therefore, understanding these global approaches is essential for AI GRC professionals to build governance programs that meet both ethical expectations and legal obligations. This knowledge can also be used by organizations to select and apply suitable frameworks, identify relevant controls, and adapt their AI governance programs to different regulatory environments.

In this chapter, we will explore major global AI governance frameworks and regulations, understand their purpose and structure, and learn how organizations can apply them to establish robust, compliant, and trustworthy AI governance systems.

Understanding the difference between regulations, standards, and frameworks

Organizations encounter three important types of requirements and guidance for AI use: regulations, standards, and frameworks. Although all three can support responsible AI, they serve different purposes and have different levels of enforceability.

Regulations are legally-binding requirements issued by governments or regulatory authorities. Organizations must comply with applicable regulations when developing, deploying, or using AI systems. Examples include the EU AI Act and GDPR.

Standards are documented requirements, specifications, or practices that provide a consistent way to achieve defined objectives. For example, ISO/IEC 42001 is an international standard that defines requirements for establishing, implementing, maintaining, and continually improving an **Artificial Intelligence Management System** (**AIMS**). An organization can use the standard to establish a structured AI governance program and may also seek certification against it. Standards are generally voluntary unless they become mandatory through regulations, contracts, organizational requirements, or certification commitments. For example, EU regulations for electronic identification require certain authentication-node operators to demonstrate compliance with ISO/IEC 27001.

Frameworks, on the other hand, provide structured guidance or best-practice approaches for addressing a particular area. They help organizations design and implement governance, risk management, and control processes. For example, the **NIST AI Risk Management Framework** (**AI RMF**) provides guidance for identifying, assessing, and managing AI-related risks. Frameworks are generally voluntary unless adopted through contracts, organizational requirements, or referenced by regulations.

In this chapter, we will have an overview of the following regulations, governance-related directives and frameworks:

	Category	**Issuing Body / Origin**	**Primary Focus**
EU AI Act	Regulation	European Union	Comprehensive AI law with risk-based requirements for providers and deployers
GDPR	Regulation	European Union	Data protection and privacy requirements applicable to AI systems
US Executive Order on Safe, Secure, and Trustworthy AI	Regulation	United States government	National-level AI governance and safety directives

	Category	Issuing Body / Origin	Primary Focus
ISO/IEC 42001 – Artificial Intelligence Management System	Standard	ISO/IEC	Management system for implementing and maintaining AI governance
NIST AI Risk Management Framework (AI RMF)	Framework	NIST (USA)	Risk-based approach to managing AI risks
OECD AI Principles	Framework	OECD	High-level principles for trustworthy AI
IEEE 7000 Series	Framework	IEEE	Ethical AI and system design standards
World Economic Forum AI Governance Playbooks	Framework	WEF	Board-level and industry best practices for AI governance

Table 4.1: AI governance regulations, directives and frameworks

Please note that almost all the countries have their own regulations or are in the process of adopting the same. They are primarily based on Responsible AI principles that we discussed in *Chapter 3*. In this chapter, we will discuss above mentioned regulations and framework which will give you basic idea about the overall objective of any regulation or framework for AI.

EU AI Act

The **EU AI Act** is the world's first comprehensive law specifically designed to regulate AI. It is issued by the European Union to ensure that AI systems are safe, trustworthy, and respectful of fundamental rights. The Act establishes legally binding requirements for organizations that develop, sell, or use AI systems within the EU. Following are some the highlights of the act:

- **Following a risk-based regulation strategy**: The EU AI Act classifies AI systems according to the level of risk they pose to individuals and society. The main categories are unacceptable risk, high-risk, transparency risk, and minimal or no risk.

Unacceptable-risk practices are prohibited, while high-risk systems are subject to stricter requirements such as risk management, documentation, testing, and human oversight. Transparency-risk systems are subject to specific disclosure requirements, while minimal- or no-risk systems generally have limited obligations. For instance, AI used to assess whether a person should receive a job or loan may be classified as high-risk, while an AI-powered spam filter is generally considered minimal or no risk.

- **Defining clear roles in the AI value chain:** The EU AI Act defines responsibilities for different participants in the AI value chain, including providers and deployers. Providers are generally responsible for developing or placing AI systems on the market, while deployers use AI systems in their operations. Each role has specific obligations under the Act. Consider a company that develops an AI system for employment purposes, which gives it responsibilities as a provider, while an organization that uses the system to screen job applicants may have responsibilities as a deployer. This helps establish clear accountability for AI systems and prevents organizations from avoiding responsibility for AI-related decisions.
- **Paying special attention to high-risk systems**: For high-risk AI systems, the EU AI Act requires risk assessments, data governance controls, technical documentation, human oversight, transparency measures, and continuous monitoring. Certain AI practices that pose unacceptable risks, such as harmful manipulation or social scoring, are prohibited.
- **Enforcing Responsible AI principles**: The EU AI Act is significant from an AI GRC perspective because it converts several Responsible AI principles into legally enforceable requirements. Principles such as safety, transparency, human oversight, accountability, and protection of fundamental rights are reflected in the Act's requirements. For example, high-risk AI systems are subject to requirements relating to risk management, data quality, documentation, human oversight, accuracy, robustness, and cybersecurity.
- **Identifying AI governance alignment**: Organizations operating in or serving the EU market need to identify the requirements applicable to their AI systems and align their AI governance frameworks, policies, procedures, and controls accordingly. Consider an organization deploying an AI chatbot that interacts directly with customers. In this case, they may need to identify applicable transparency requirements and establish appropriate governance controls to ensure users are informed about the use of AI.

We will discuss the entirety of the EU AI Act in *Chapter 8*. While the EU AI Act provides a broad regulatory framework for AI, the **General Data Protection Regulation** (**GDPR**) focuses specifically on the protection of personal data and privacy. Since many AI systems rely on

personal data, understanding GDPR is an important part of understanding AI governance and compliance. Let us now look at how GDPR applies to AI systems.

General data protection regulation (GDPR)

GDPR is the European Union's comprehensive law for data protection and privacy. It applies to any organization that processes personal data of individuals in the EU, including AI systems, and is legally binding with significant penalties for non-compliance. GDPR is highly relevant to AI because many AI systems rely on large volumes of personal data. It ensures that personal data used in AI systems is processed lawfully, fairly, and transparently.

GDPR requires organizations to:

- **Establish a lawful basis for AI data processing:** Identify and document the appropriate legal basis for collecting and using personal data to train, test, deploy, or operate an AI system.
- **Obtain consent where required:** Where consent is the appropriate legal basis, obtain clear and informed consent from individuals and provide the ability to withdraw consent where applicable.
- **Define the purpose of AI data use**: Clearly define why personal data is being used in an AI system and avoid using the data for purposes that are incompatible with the original purpose without an appropriate legal basis.
- **Minimize data used by AI systems**: Use only the personal data necessary for the intended AI purpose. For example, an AI model developed to recommend products should not collect unrelated sensitive information simply because it is available.
- **Maintain data accuracy**: Ensure that personal data used to train or operate AI systems is accurate and kept up to date where necessary. Inaccurate data can result in incorrect or unfair AI outputs.
- **Define data retention requirements**: Establish appropriate retention periods for personal data used in AI systems and delete or anonymize data when it is no longer required.
- **Protect personal data used by AI**: Apply appropriate technical and organizational security measures to protect personal data used for AI training, testing, and operation against unauthorized access, loss, alteration, or disclosure.
- **Conduct a Data Protection Impact Assessment (DPIA)**: Conduct a DPIA when AI-related processing is likely to result in a high risk to individuals' rights and freedoms. This can be particularly relevant for systematic and extensive profiling or large-scale processing of sensitive personal data.

- **Address automated decision-making and profiling**: Where AI systems make decisions based solely on automated processing that have legal or similarly significant effects on individuals, organizations must comply with the applicable GDPR safeguards. These may include providing information about decision-making, enabling human intervention, and allowing individuals to contest the decision.
- **Provide transparency about AI processing**: Inform individuals about how their personal data is processed, including the purposes of processing, the categories of data used, the legal basis, retention period, and, where applicable, the existence of automated decision-making and the logic involved.

From an AI GRC perspective, GDPR ensures that privacy and data protection principles are embedded into AI governance, policies, and system design, and it continues to influence global privacy and AI-related data governance practices.

While GDPR focuses primarily on data protection and privacy, AI governance also needs to address areas such as safety, security, fairness, accountability, and responsible innovation. Different countries have taken different approaches to addressing these concerns. Let's now look at the US's approach to safe, secure, and trustworthy AI.

US executive order on safe, secure, and trustworthy AI

The US Executive Order on Safe, Secure, and Trustworthy AI was issued by the United States government to guide the responsible development and use of artificial intelligence. Unlike the EU AI Act, this Executive Order is not a single comprehensive AI law. Instead, it provides strategic direction and assigns responsibilities to federal agencies to develop policies, standards, and oversight mechanisms for AI.

Under this Executive Order, US agencies such as NIST, the Department of Commerce, and sectoral regulators are tasked with issuing guidance, developing technical standards, and integrating AI risk management into existing regulatory frameworks. This approach reflects the US model of sector-based and risk-based AI governance rather than a single centralized AI law. A sector-based approach allows regulators to address industry-specific risks. For example, healthcare may focus on patient safety, while financial services may focus on fraud and consumer protection.

The Executive Order focuses on promoting AI safety, security, privacy, fairness, and accountability. It emphasizes managing risks from advanced and generative AI systems, protecting consumers and workers, and preventing misuse of AI for harmful activities. It also highlights the importance of transparency, security testing, and responsible innovation.

From an AI GRC perspective, the Executive Order signals strong regulatory direction. Organizations operating in or serving the US market are expected to align with emerging agency guidance, adopt risk management practices such as the NIST AI Risk Management Framework, and implement governance controls to ensure safe and trustworthy AI systems.

Having discussed the US approach to AI governance, let's now look at an international standard that provides a structured approach to managing AI. ISO/IEC 42001 focuses on establishing, implementing, maintaining, and continually improving an **AI Management System** (**AIMS**). This makes it useful for organizations looking to put AI governance into practice.

ISO/IEC 42001 – artificial intelligence management systems

ISO/IEC refers to standards jointly developed by the **International Organization for Standardization** (**ISO**) and the **International Electrotechnical Commission** (**IEC**). These organizations create globally recognized standards that help ensure consistency, quality, safety, and reliability across industries and technologies.

ISO/IEC 42001 is the first international management system standard specifically designed for artificial intelligence. It provides a structured framework for organizations to establish, implement, maintain, and continually improve AIMS.

The standard is built on the same management system structure used in other ISO standards such as ISO/IEC 27001. This makes it easier for organizations to integrate AI governance into existing information security, privacy, and risk management systems. ISO/IEC 42001 focuses on embedding Responsible AI principles into organizational processes rather than treating AI governance as a standalone activity.

ISO/IEC 42001 requires organizations to define AI policies, assign roles and responsibilities, perform AI risk assessments, implement controls across the AI lifecycle, monitor AI system performance, and maintain documentation and audit evidence. It also emphasizes leadership involvement, continuous improvement, and accountability for AI-related decisions.

From an AI GRC perspective, ISO/IEC 42001 provides practical guidance on how to operationalize Responsible AI principles and regulatory expectations. Organizations can use this standard to build a structured AI governance program, demonstrate due diligence, and prepare for compliance with emerging AI regulations.

ISO/IEC 42001 is unique among the AI governance frameworks discussed in this book because it provides a complete management system approach for establishing, implementing, maintaining, and continually improving AI governance. It helps organizations translate

Responsible AI principles and governance requirements into defined processes, responsibilities, controls, monitoring, and continual improvement activities. We will discuss ISO/IEC 42001 in detail in the next chapter.

NIST AI risk management framework (AI RMF)

The NIST is a US government agency that develops widely used standards and frameworks for technology and cybersecurity. The NIST AI RMF was created to help organizations manage risks associated with AI systems in a structured and practical way. Following are some highlights of the AI RMF:

- The AI RMF provides a voluntary, risk-based framework for identifying, assessing, and managing AI-related risks throughout the AI lifecycle. It is designed to support trustworthy and responsible AI development without being a regulatory requirement. However, it is widely referenced by regulators, industry, and government agencies, making it highly influential in AI governance practices.
- The framework is organized around four core functions: **Govern**, **Map**, **Measure**, and **Manage**. *Govern* focuses on establishing AI governance structures, policies, and accountability. *Map* helps organizations understand AI system context, stakeholders, and potential risks. *Measure* involves assessing and monitoring AI performance, risks, and impacts. *Manage* focuses on prioritizing risks, implementing controls, and taking corrective actions.
- From an AI GRC perspective, the NIST AI RMF provides practical guidance for building AI risk management programs, aligning AI governance with organizational objectives, and demonstrating due diligence in managing AI risks.

We will discuss NIST AI RMF in detail in *Chapter 6*.

OECD AI principles

The **Organization for Economic cooperation and Development** (**OECD**) is an international body that develops global policy guidance on economic and technological matters. Though OECD AI principles are not legally binding, they are widely adopted as a global reference point for Responsible AI and have shaped frameworks such as the EU AI Act and ISO/IEC 42001. The OECD has defined five key principles for trustworthy AI:

- AI should benefit people and the planet by driving inclusive growth, sustainable development, and well-being.
- AI systems should respect human rights, democratic values, and diversity, and should be fair and non-discriminatory.
- AI systems should be transparent and explainable so that people understand when and how AI is used.
- AI systems should be robust, secure, and safe throughout their lifecycle.
- Organizations and individuals developing or deploying AI should be accountable for its proper functioning.

We will discuss OECD principles in *Chapter 8*.

We will now explore the IEEE 7000 Series, which provides practical guidance for incorporating ethical considerations into the design and development of AI and autonomous systems.

IEEE 7000 series

The **Institute of Electrical and Electronics Engineers** (**IEEE**) is a global professional body that develops technical and ethical standards for emerging technologies. The IEEE 7000 Series is a set of standards specifically created to address ethical considerations in the design and deployment of AI and autonomous systems.

Unlike regulations, the IEEE 7000 Series provides practical guidance and engineering-focused standards to help organizations embed ethical values into AI system design from the beginning. These standards are not legally binding but are widely respected and used as best practice references in AI governance and product development.

The IEEE 7000 Series covers key areas such as ethical system design, transparency, explainability, bias reduction, data privacy, and human oversight. For example, IEEE 7000 provides a process for identifying and addressing ethical risks in system design, while IEEE 7001 focuses on transparency of autonomous systems, and IEEE 7003 addresses bias considerations in AI systems.

Moving from ethical AI design to practical governance, we will explore the **World Economic Forum AI Governance Playbooks** (**WEF AI Governance Playbooks**) that provide practical guidance for implementing AI governance, strengthening leadership oversight, and integrating AI risks into broader business processes.

World economic forum AI governance playbooks

The **World Economic Forum (WEF)** is an international organization that brings together governments, industry leaders, and experts to address global challenges, including the governance of emerging technologies. A playbook is a practical guidance document that provides step-by-step instructions, best practices, and recommended actions to help organizations implement a specific program or process.

The WEF AI Governance Playbooks are practical guidance documents designed to help organizations implement effective AI governance at both operational and board levels.

These playbooks focus on translating Responsible AI principles into real-world governance practices. They provide recommendations on establishing AI oversight structures, defining leadership accountability, integrating AI risk management into enterprise risk programs, and ensuring board-level visibility of AI-related risks and opportunities.

The WEF AI Governance Playbooks emphasize topics such as AI strategy alignment with business objectives, ethical AI decision-making, transparency and disclosure, workforce readiness, and stakeholder trust. They are designed to help senior management and board members ask the right questions and make informed decisions about AI adoption and oversight. Having reviewed the major global AI governance frameworks, regulations, standards, and initiatives, let us now examine the common themes that emerge across these approaches.

Common themes across the globe for AI governance

AI GRC becomes more effective when complex requirements are simplified into clear actions. Although global AI regulations, standards, and frameworks differ in their purpose, scope, and level of enforceability, several common themes appear across them. These common themes reflect the Responsible AI principles discussed in *Chapter 3* and demonstrate that AI governance is increasingly moving toward a consistent set of expectations for organizations. The following example highlights why simplifying common requirements across AI regulations, standards, and frameworks is important for effective AI governance:

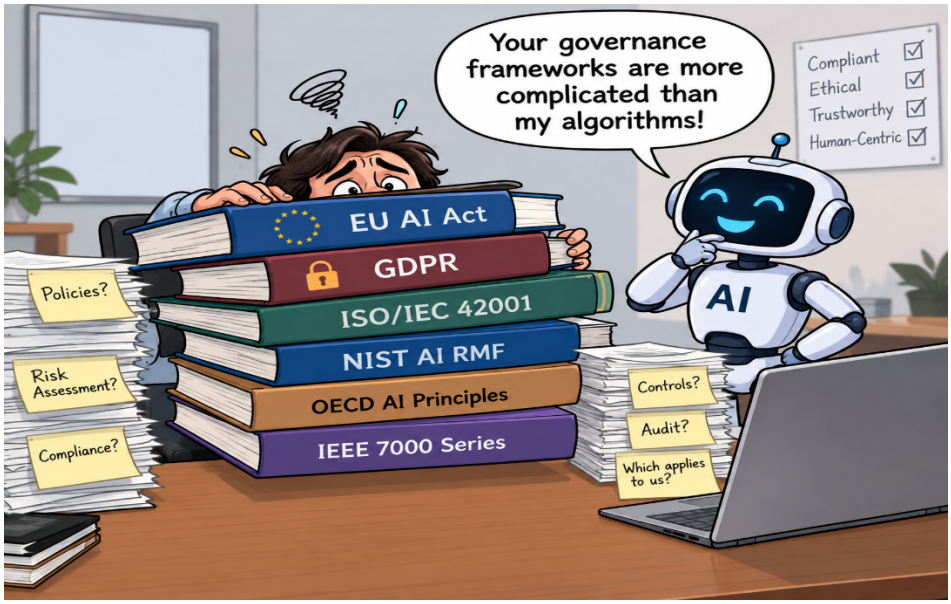


Figure 4.1: Key GRC concept - simplification

Understanding these common themes is important for AI GRC professionals because an organization operating across multiple countries may need to comply with different regulations while implementing a common governance approach. Instead of creating completely separate governance programs for each regulation or framework, organizations can identify these common requirements and establish controls that address multiple requirements simultaneously.

The following are some of the major common themes across global AI governance frameworks and regulations:

AI Principle	OECD	ISO/IEC (ISO/ IEC 42001)	NIST AI RMF	EU AI Act	World Economic Forum	IEEE
Accountability	✓	✓	✓	✓	✓	✓
Fairness and Non-Discrimination	✓	✓	✓	✓	✓	✓

AI Principle	OECD	ISO/IEC (ISO/ IEC 42001)	NIST AI RMF	EU AI Act	World Economic Forum	IEEE
Transparency	✓	✓	✓	✓	✓	✓
Explainability	✓	✓	✓	✓	✓	✓
Privacy and Data Protection	✓	✓		✓	✓	✓
Security and Resilience	✓	✓	✓	✓	✓	✓
Safety and Reliability	✓	✓	✓	✓	✓	✓
Human Oversight	✓	✓	✓	✓	✓	✓
Risk-Based Approach	✓	✓	✓	✓		
Ethical and Responsible Use	✓		✓	✓	✓	✓
Continuous Monitoring and Improvement	✓	✓	✓		✓	✓
Compliance with Laws and Standards	✓	✓	✓	✓	✓	✓

Table 4.2: Common themes across global AI governance

All the above responsible AI principles have already been discussed in *Chapter 3*.

Summary

In this chapter, we saw that regulations, standards, and frameworks serve different purposes and have different levels of enforceability. We learned that regulation is a legally binding requirement issued by a government or regulatory authority, a standard provides defined requirements or practices for achieving specific objectives, and a framework provides structured guidance or best practices for managing a particular area.

We explored major global AI governance regulations, standards, and frameworks, including the EU AI Act, GDPR, ISO/IEC 42001, NIST AI RMF, OECD AI Principles, and others, and understood that these approaches share common themes such as risk management, accountability, transparency, human oversight, security, and continuous monitoring.

We discussed that organizations need to identify the AI governance requirements applicable to their regulatory and business environment and use relevant frameworks to build and strengthen their AI governance programs. By the end of the chapter, we were able to conclude that ISO/IEC 42001 provides a complete management-system approach for implementing and continually improving AI governance.

The purpose of this chapter was to help AI GRC professionals understand the different approaches to AI governance and identify which requirements and guidance may be relevant to their organizations. This knowledge can help organizations compare global approaches, identify applicable requirements, and strengthen their AI governance programs. By understanding these global approaches, AI GRC professionals can align internal governance programs with international expectations and Responsible AI principles.

In the next chapter, we will explore its structure, key requirements, and how organizations can use it to build effective and auditable AI governance programs. Among global AI governance frameworks, ISO/IEC 42001 is unique as it provides a complete management system for implementing AI governance. In the next chapter, we will explore its structure, key requirements, and how organizations can use it to build effective and auditable AI governance programs.

5

AI Governance Framework – ISO/IEC 42001

In the previous chapter, we introduced ISO/IEC 42001 as an international standard for establishing and managing an **Artificial Intelligence Management System** (**AIMS**). We also discussed how it provides a structured approach to implementing AI governance. To build on our understanding so far, we will look at the standard in more detail and understand how organizations can implement its requirements in practice.

ISO/IEC 42001 provides a structured framework for establishing policies, defining roles and responsibilities, managing AI risks, implementing lifecycle controls, and continually improving AI governance. The standard follows the ISO management system approach, which also makes it easier to integrate AI governance with existing information security, privacy, and risk management programs.

In this chapter, we will explore the structure and requirements of ISO/IEC 42001, understand its key clauses and controls, and learn how organizations can use it to build effective, compliant, and sustainable AI governance systems.

ISO and IEC – international standards bodies

Let us first understand the issuing bodies: ISO and IEC. **ISO** stands for the **International Organization for Standardization**. It is an independent, non-governmental organization that develops and publishes international standards across a wide range of industries. ISO has member bodies from more than 160 countries. Its objective is to ensure that products, services, and systems are safe, reliable, and of good quality.

IEC stands for the **International Electrotechnical Commission**. It is an international organization that develops standards for electrical, electronic, and related technologies. IEC

has member and affiliate countries from across the world. Its objective is to ensure that electrical and electronic systems are safe, compatible, and efficient.

Together, ISO and IEC jointly publish international standards for information technology and emerging technologies, including artificial intelligence. Let us now understand what a standard is and how it is developed.

What is a standard?

We discussed in the last chapter that a **standard** is a documented set of rules, processes, or characteristics designed to ensure that a product, service, or system meets defined expectations. Standards help organizations achieve:

- Quality
- Safety
- Efficiency
- Consistency

As an AI GRC professional, it is important to understand what a standard is because standards provide a structured basis for establishing governance, demonstrating compliance, and providing assurance to stakeholders.

For example, ISO/IEC 42001 defines specific requirements for establishing an AIMS. If an organization is certified against this standard, it confirms that the organization meets all defined requirements, providing assurance to customers, regulators, and other stakeholders that AI systems are governed responsibly.

How international standards are developed

The development of international standards follows a structured and collaborative process to ensure global relevance, technical accuracy, and wide acceptance. The typical standard development process involves the following steps:

1. Formation of a Working Group under a Technical Committee.
2. Drafting of the standard by subject matter experts.
3. Circulation of the draft among member countries for review and feedback.
4. Multiple rounds of discussion and refinement based on comments received.
5. Approval of the final draft through a two-thirds majority vote of participating member countries.
6. Publication of the standard as an official international document.

This rigorous process ensures that international standards are developed through expert input, global consensus, and practical applicability across industries and regions.

Now that we understand how international standards are developed, we will move on to see how ISO/IEC 42001 is structured and how its requirements are organized.

Structure of AIMS under ISO/IEC 42001

AIMS is the set of policies, processes, roles, responsibilities, and controls that an organization uses to manage AI in a structured and controlled manner. It helps organizations manage AI-related risks, meet applicable requirements, monitor performance, and continually improve their AI governance practices. The AIMS under ISO/IEC 42001 is structured into clauses and annexes that we will look into next.

Ten clauses (clauses 1 to 10)

These clauses define the mandatory requirements for establishing, implementing, maintaining, and continually improving AIMS. They follow the standard ISO management system structure and cover areas such as leadership, planning, support, operation, performance evaluation, and improvement. The following table gives details about 10 clauses:

Clause	Title	Description
Clause 1	Scope	Defines the purpose and applicability of the standard for establishing AIMS.
Clause 2	Normative References	Lists referenced standards required for applying ISO/IEC 42001.
Clause 3	Terms and Definitions	Provides standardized definitions of key terms used in the standard.
Clause 4	Context of the Organization	Requires understanding of internal and external factors and defining the scope of AIMS.
Clause 5	Leadership	Focuses on top management responsibility, AI policy, and the assignment of roles and accountability.
Clause 6	Planning	Covers AI risk assessment, risk treatment, and AI governance objectives.
Clause 7	Support	Addresses resources, competence, awareness, communication, and documentation needs.

Clause	Title	Description
Clause 8	Operation	Defines operational controls for managing AI systems across their lifecycle.
Clause 9	Performance Evaluation	Requires monitoring, measurement, internal audits, and management review of AIMS.
Clause 10	Improvement	Focuses on corrective actions and continual improvement of AI governance practices.

Table 5.1: Ten Clauses of the ISO/IEC 42001 Standard

Clauses 1 to 3 of ISO/IEC 42001 are informational in nature and are not subject to certification. They provide background, references, and definitions to support understanding of the standard. *Clauses 4 to 10* contain the mandatory requirements of AIMS and must be implemented by organizations seeking ISO/IEC 42001 certification.

Note

For your ready reference, we have designed an AI governance policy which is completely aligned with the previously-mentioned requirements of the ISO/IEC 42001 standard. The policy is available in *Template 1* of this book's GitHub repository. To access the repository link, follow the steps in the *Download the example code files* section in the *Preface*.

Four annexes (annexes A, B, C, and D)

These annexes provide additional guidance, reference information, and suggested controls to support the implementation of AIMS. Annexes help organizations understand how to apply the requirements in practical AI governance scenarios. The following table gives details of the annexes:

Annex	Title	Description
Annex A	Reference AI Controls	Provides recommended AI governance and risk controls to support implementation.
Annex B	Implementation Guidance	Offers practical guidance on applying clauses and controls in real-world AI environments.

Annex	Title	Description
Annex C	Mapping to Other Standards	Shows alignment with related standards such as ISO/IEC 27001 and privacy frameworks.
Annex D	Sector and Use-Case Considerations	Provides additional considerations for specific industries and AI use cases.

Table 5.2: Annexes A to D of ISO/IEC 42001 Standard

Let us now understand **Annex A** in detail.

Annex A

As we discussed earlier, ISO/IEC 42001 includes four annexes that support the implementation of AIMS and provide practical guidance beyond the main clauses.

Annex A of ISO/IEC 42001 provides a set of controls that organizations can use to manage risks associated with AI systems. Annex A is particularly important because, for an organization to achieve ISO/IEC 42001 certification, it must state whether each control mentioned in Annex A is implemented and, if not implemented, needs to provide justification for exclusion.

The key Annex A control areas include:

- **AI governance and accountability**: Controls related to AI governance structure, roles and responsibilities, leadership oversight, and accountability for AI outcomes
- **AI risk management**: Controls for identifying, assessing, treating, and monitoring risks associated with AI systems throughout their lifecycle
- **Data governance and management**: Controls addressing data quality, data sourcing, data protection, privacy, bias mitigation, and data lifecycle management for AI systems
- **AI lifecycle management**: Controls covering AI design, development, testing, deployment, operation, change management, and retirement
- **Transparency and explainability**: Controls to ensure appropriate disclosure of AI use, documentation of AI systems, and the ability to explain AI-supported decisions where required
- **Human oversight and intervention**: Controls to define human-in-the-loop or human-on-the-loop mechanisms, escalation paths, and override capabilities
- **Security and Resilience**: Controls focused on protecting AI systems, models, and data from security threats and ensuring resilience against failures and attacks

- **Monitoring, logging, and performance evaluation**: Controls related to ongoing monitoring of AI performance, drift detection, incident logging, and effectiveness reviews
- **Incident management and corrective action**: Controls for identifying, reporting, investigating, and responding to AI-related incidents and nonconformities
- **Third-party and external ai management**: Controls to manage risks arising from third-party AI providers, vendors, and externally developed AI systems

The selection and applicability or otherwise of these controls along with proper justification are formally documented in the **Statement of Applicability** (**SoA**) which we will dive into next.

Statement of Applicability (SoA)

The SoA is a key document under ISO/IEC 42001 that lists which AI controls are applicable to an organization and how they are implemented. The SoA identifies all relevant controls (from Annex A) and clearly states whether each control is applicable or not applicable, along with a justification. For applicable controls, the SoA describes how the control is implemented or references supporting policies, procedures, or technical measures.

From a certification and audit perspective, the SoA is critical evidence. Auditors use it to verify that AI risks have been systematically assessed and that appropriate controls have been selected, implemented, and justified.

Note

For your ready reference, a SoA template is provided as *Template 5* in this book's GitHub repository. To access the repository link, follow the steps in the *Download the example code files* section in the *Preface*.

Having understood the requirements of the ISO/IEC 42001 standard, we will now discuss how to implement it in an organization.

Implementation steps for ISO/IEC 42001

Implementation must be approached in a structured and phased manner to ensure effectiveness and audit readiness. Here are the steps that an organization should follow:

1. **Create an AI governance policy**: Define and approve an AI governance policy/AIMS Policy that sets the organization's intent, principles, and commitment toward responsible and compliant use of AI. The policy should align with the requirements of

the ISO/IEC 42001 standard. You may like to refer to *Template 1* available in this book's GitHub repository.

2. **Perform gap assessment**: Conduct a gap assessment against ISO/IEC 42001 requirements to understand the current maturity level and identify missing policies, processes, and controls. Implement the identified gaps. An implementation template, *Template 6*, is available in this book's GitHub repository. Once all the gaps are addressed, move to the next step.
3. **Prepare the SoA**: Document all controls mentioned in Annex A along with their applicability, exclusions, and implementation status in the SoA. You may like to refer to *Template 5* in the GitHub repository.
4. **Certification Audit**: Engage an accredited certification body to conduct the ISO/IEC 42001 certification audit

The steps discussed in the previous list provide a structured approach to implementing ISO/IEC 42001. To understand how these steps can be applied in practice, we will look at a case study that demonstrates the implementation journey from assessing the current state to preparing for certification.

Sample case study – implementing ISO/IEC 42001 in a financial services company

Consider a financial services company that uses AI to assess loan applications. The organization wants to establish a formal AI governance system and prepare for ISO/IEC 42001 certification. Let's see how the steps we discussed earlier are applicable here:

1. Create an AI governance policy

The organization reviews its use of AI in the loan assessment process, applicable financial, privacy, consumer protection, and anti-discrimination requirements, and Responsible AI principles. It then develops an AI Governance Policy aligned with ISO/IEC 42001, defining accountability, human oversight, transparency, data protection, risk management, and monitoring requirements for the AI-based loan assessment system.

2. Perform gap analysis

The organization compares its existing practices for the AI-based loan assessment system against ISO/IEC 42001 requirements. The assessment identifies gaps such as the absence of an AI-specific risk and impact assessment, incomplete AI system inventory, inadequate assessment of data quality and bias, limited explainability, and lack of defined human oversight and monitoring processes. The identified gaps are assigned for remediation.

3. Prepare the SoA

The organization reviews the Annex A controls and determines their applicability to the loan assessment system. Controls relating to AI impact assessment, data quality, responsible AI objectives, human oversight, documentation, and monitoring are considered based on the risks associated with AI-assisted lending decisions. The organization documents the applicability, implementation status, and justification for each control.

4. Implement and maintain controls

The organization implements the applicable controls by conducting AI risk and impact assessments, validating data quality, testing for potential bias, documenting the AI model and its limitations, establishing human review and escalation mechanisms, and monitoring model performance and outcomes. Evidence of these activities is maintained for ongoing monitoring and audit.

5. Certification audit

Once the AI management system and applicable controls are implemented and sufficient evidence is available, the organization engages an accredited certification body for the ISO/IEC 42001 certification audit. The audit examines whether the organization has effectively governed the AI-based loan assessment system and addressed relevant risks such as fairness, transparency, privacy, security, accuracy, and human oversight.

This example shows how ISO/IEC 42001 can be implemented as a practical governance program rather than simply as a documentation exercise. The organization establishes policies, identifies gaps, determines applicable controls, implements them, and maintains evidence to demonstrate that the AI governance system is operating effectively.

Organizations may face several challenges when implementing ISO/IEC 42001. Being aware of these common pitfalls can help AI GRC professionals plan implementation more effectively. Let us understand the same in our next section.

Common pitfalls when implementing ISO/IEC 42001

Let's discuss some common pitfalls organizations may face while implementing ISO/IEC 42001:

- **Treating AI governance as an IT-only responsibility:** AI governance involves business, legal, risk, compliance, security, privacy, and other functions. Limiting implementation to the IT team can result in gaps in governance and accountability.
- **Focusing only on certification:** Organizations may focus on preparing for the certification audit rather than building an effective AIMS. The objective should be to establish governance practices that continue to operate after certification.

- **Creating policies without implementation:** Documenting policies and procedures is not enough. Organizations need to implement the required controls and maintain evidence that they are operating effectively.
- **Incomplete AI system inventory:** Without a complete inventory of AI systems, organizations may not know which systems are being used, who owns them, what data they process, or what risks they present.
- **Inadequate AI risk assessment:** AI risks can change as models, data, business processes, and use cases change. Risk assessments should therefore be performed and updated periodically.
- **Treating the SoA as a formality:** The SoA should clearly document which controls are applicable, their implementation status, and the reasons for any exclusions. It should be based on the organization's actual AI risks and environment.
- **Insufficient management involvement:** Successful implementation requires active support from top management. Without clear leadership commitment, AI governance may not receive sufficient resources, authority, or attention.
- **Ignoring third-party AI systems:** Organizations often overlook risks associated with AI services provided by vendors or cloud providers. Third-party AI systems should also be assessed and governed based on their risk and intended use.
- **Lack of continual improvement:** ISO/IEC 42001 is a management system that requires ongoing monitoring and improvement. Treating implementation as a one-time project can reduce its effectiveness over time.

ISO/IEC 42001 implementation should be viewed as a continuous governance process rather than a one-time certification exercise. The focus should be on making AI governance part of the organization's normal business and risk management practices.

Let us now understand how the certification audit is conducted and what the organization should expect during the audit process.

Certification audit process

During the ISO/IEC 42001 certification audit, the auditor does not simply check whether the organization has created the required policies and documents. The auditor looks for objective evidence that the AIMS has been established, implemented, maintained, and is operating effectively. This means the organization should be able to demonstrate that its documented requirements are actually being followed in practice.

The following image illustrates why the auditor plays such a critical role in the AIMS certification process:

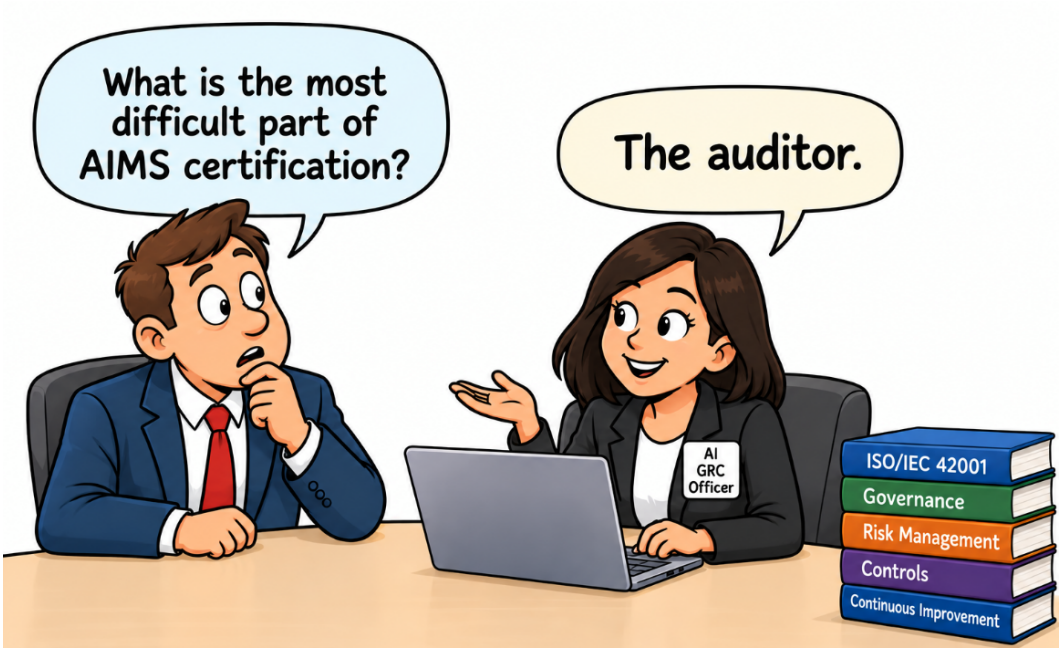


Figure 5.1: AIMS certification - preparing for the audit

For example, having an AI Risk Management Policy is not sufficient by itself. The organization should also be able to demonstrate that AI systems have been identified, risks have been assessed, appropriate controls have been implemented, and the effectiveness of those controls is being monitored.

The auditor may look for evidence in the following areas:

- **AI system inventory:** Evidence that the organization has identified and documented the AI systems within the scope of the AIMS. The inventory may include information such as the purpose of the AI system, owner, business process, data used, risk classification, and deployment status.
- **AI risk assessment:** Evidence that AI-related risks have been identified and assessed. The auditor may review whether risk assessments consider factors such as the intended use of the AI system, potential impact on individuals, data-related risks, security risks, and other relevant risks.

- **Risk treatment:** Evidence that identified risks have been appropriately addressed. The auditor may verify whether risk treatment plans have been defined, responsible persons have been assigned, and the selected controls have actually been implemented.
- **Roles and responsibilities:** Evidence that AI governance responsibilities have been clearly assigned. The auditor may interview personnel to determine whether they understand their responsibilities and whether management has provided appropriate authority and resources.
- **Implementation of controls:** Evidence that applicable controls have been implemented and are operating as intended. This may include controls related to data governance, human oversight, transparency, security, AI lifecycle management, monitoring, and third-party AI services.
- **Policies and procedures:** Evidence that relevant AI policies, procedures, and guidelines have been approved, communicated, reviewed, and followed. The auditor may compare documented requirements with actual organizational practices.
- **Monitoring and measurement:** Evidence that the organization regularly monitors the performance and effectiveness of its AIMS and relevant AI systems. This may include monitoring results, performance measures, risk indicators, incident records, and review reports.
- **Issue and incident management:** Evidence that AI-related incidents, nonconformities, and other issues are identified, documented, investigated, and addressed. The auditor may also check whether corrective actions have been implemented and whether their effectiveness has been evaluated.
- **Internal audit:** Evidence that internal audits have been conducted to assess whether the AIMS conforms to applicable requirements and is operating effectively. The auditor may review audit plans, reports, findings, and corrective actions.
- **Management review:** Evidence that top management periodically reviews the AIMS. This may include review of audit results, AI risks, performance, objectives, incidents, changes in the organization, and opportunities for improvement.
- **Training and awareness:** Evidence that personnel involved in AI-related activities have the required competence, training, and awareness. Training records and competency assessments may be reviewed where applicable.

- **Documented information**: Evidence that required records are properly maintained and controlled. Examples may include AI risk assessments, impact assessments, approvals, training records, monitoring results, audit reports, management reviews, and corrective action records.
- **Continual improvement:** Evidence that the organization uses audit findings, incidents, monitoring results, and lessons learned to improve its AIMS and AI governance practices.

The certification audit is not a document-checking exercise. The auditor will look for evidence that the AIMS requirements are implemented, consistently followed, monitored, and continually improved. An organization may have well-written policies, but if it cannot demonstrate their implementation and effectiveness, this can result in audit findings or nonconformities.

Importance of ISO/IEC 42001 certification

ISO/IEC 42001:2023 is the first international standard specifically designed for AIMS. This certification provides independent assurance that an organization has implemented a management system in line with internationally recognized standards. It demonstrates to customers, regulators, partners, and other stakeholders that the organization follows structured, consistent, and audited practices.

In the context of AI governance, ISO/IEC 42001 certification signals a strong commitment to responsible, trustworthy, and risk-based use of AI. It helps organizations build trust, improve internal governance discipline, reduce compliance risks, and enhance credibility in the global market. ISO certification also supports regulatory readiness, as many regulatory expectations align closely with ISO management system requirements.

Leveraging an existing ISO/IEC 27001 certificate to implement ISO/IEC 42001

Most GRC professionals are already familiar with ISO/IEC 27001, as it is one of the world's most widely adopted **Information Security Management System** (**ISMS**). A common question when approaching ISO/IEC 42001 is: *"Can we leverage our existing ISO/IEC 27001 certification instead of building a completely new management system?"*

Organizations that already have an established ISMS are in a strong position to implement ISO/IEC 42001. However, ISO/IEC 42001 should not be treated simply as an extension of ISO/IEC 27001. The two standards have different objectives and risk perspectives. ISO/IEC 27001 focuses primarily on information security, while ISO/IEC 42001 establishes a management system for the responsible management of artificial intelligence.

The practical advantage is that many of the management-system foundations are common. Rather than creating a completely separate governance framework, an organization can leverage its existing ISMS processes, documentation, governance structures, risk-management methodology, audit program, and continual-improvement mechanisms.

A GRC professional can approach ISO/IEC 42001 implementation using the sequence of the following sections.

1. Review the scope

Review the existing ISO/IEC 27001 scope and determine whether the same organizational units, processes, locations, products, services, and technologies should be included in the ISO/IEC 42001 scope. The scopes may be aligned, but they do not have to be identical.

2. Align the risk management processes

Organizations with ISO/IEC 27001 generally have an established risk management process. This process can be leveraged for ISO/IEC 42001 instead of creating a separate risk process.

The existing methodology for risk identification, assessment, treatment, acceptance, and monitoring can be extended to include AI-specific risks and impacts. For example, the assessment should consider risks related to fairness, transparency, explainability, safety, privacy, security, human oversight, and accountability.

The organization should therefore add AI-specific risk criteria to its existing risk assessment methodology and AI risks can be maintained within the existing risk register or GRC tool.

3. Align the governance process

Most organizations with an ISO/IEC 27001-certified ISMS already have an IS or IT Governance Committee. The members of this committee are generally similar to those expected for AI governance, including IT, information security, risk, compliance, privacy, business functions, and internal audit.

Therefore, the existing committee can usually be leveraged for AI governance rather than creating a completely separate committee. Its terms of reference and responsibilities can be expanded to include AI-related matters such as AI risks, AI inventory, responsible AI, AI incidents, and compliance.

Where necessary, AI system owners, data scientists, legal specialists, or other AI subject-matter experts can be invited to provide additional expertise. This approach helps avoid duplication and integrates AI governance into the organization's existing GRC structure.

4. Align the management review process

Organizations with ISO/IEC 27001 generally already have a defined management review process. This process can be extended to include the AIMS rather than creating a separate review process.

The existing management review can include AI-related matters such as AI risks, objectives and performance, significant AI incidents, audit findings, compliance, AI system changes, and opportunities for improvement.

This allows senior management to review information security and AI governance together and ensures that AI governance becomes part of the organization's existing management and GRC framework.

5. Align the internal audit processes

Organizations with ISO/IEC 27001 generally already have an established internal audit process. This process can be extended to cover the AIMS rather than creating a separate AI audit program.

The existing internal audit methodology, audit planning, auditor independence, reporting, finding management, and follow-up processes can be reused. The audit scope and checklist can be expanded to include AI-specific areas such as AI governance, AI risk assessments, AI inventory, responsible AI principles, human oversight, transparency, and AI lifecycle management.

Where required, auditors may need additional AI-related knowledge or subject-matter expertise. This allows the organization to integrate AI audits into its existing internal audit and GRC framework while avoiding duplication.

Summary

In this chapter, we understood that ISO/IEC 42001 provides a structured management-system approach for governing AI within an organization. We learned about the purpose and structure of the standard, including *Clauses 4–10* and the importance of Annex A controls. We understood how an organization can establish AIMS by defining its context, leadership, planning, support, operations, performance evaluation, and continual improvement processes.

We learned how to perform a gap assessment, prepare the SoA, implement applicable controls, and prepare for certification. We also understood that ISO/IEC 42001 is the first international management-system standard specifically designed for AI governance, while other ISO AI standards address specific areas such as AI terminology, AI systems, and AI risk management. We explored the benefits of adopting ISO/IEC 42001, and the common pitfalls organizations should avoid during implementation. Finally, we discussed how to leverage existing ISO 27001 standard for implementing ISO 42001.

In the next chapter, we will explore another AI governance framework, the NIST AI Risk Management Framework, in detail. We will understand its core functions, risk-based approach, and how organizations can use it to strengthen AI risk management and governance practices.

6

AI Governance Framework – NIST AI RMF

In the previous chapter, we explored **ISO/IEC 42001** and how an **Artificial Intelligence Management System** (**AIMS**) can help organizations establish a structured approach to AI governance. ISO/IEC 42001 provides a formal set of requirements for establishing and maintaining an AIMS, which is particularly useful for organizations seeking a structured management system and, where desired, certification. However, not every organization may need certification or may want to implement all the requirements of a formal management system. In such situations, a more flexible approach to AI risk management can be useful. This is where the **NIST AI Risk Management Framework** (AI RMF) can help.

Understanding the NIST AI RMF is particularly useful for AI GRC professionals because it provides a flexible approach that can be adapted to different AI systems, business environments, and risk levels.

In this chapter, we will explore the NIST AI RMF and understand how it can help organizations identify, assess, measure, and manage AI-related risks. We will look at its four functions: **Govern**, **Map**, **Measure**, and **Manage**, along with the **Categories** and **Subcategories** that provide more detailed guidance.

What is NIST?

NIST stands for the **National Institute of Standards and Technology**. It is a non-regulatory agency of the United States Department of Commerce that develops standards, frameworks, and guidelines to promote innovation, security, and trust in technology. NIST is globally recognized for issuing practical and widely adopted frameworks that help organizations manage risk and improve governance. Some well-known frameworks and standards issued by

NIST include the **Cybersecurity Framework** (**CSF**), the **NIST Privacy Framework**, **NIST SP 800-53** for security and privacy controls, **NIST SP 800-61** for incident response, and the **NIST AI Risk Management Framework** (**AI RMF**). Although the NIST AI RMF is voluntary, it is extensively used by governments, regulators, and organizations worldwide as a best practice, making NIST a highly influential authority in technology and AI governance. Now, let's discuss NIST AI RMF and how it helps organizations manage AI risks in a practical and flexible way.

What is NIST AI RMF?

The NIST AI RMF is a voluntary framework developed by the National Institute of Standards and Technology to help organizations identify, assess, manage, and mitigate risks associated with artificial intelligence systems. NIST AI RMF 1.0 identifies seven key characteristics of trustworthy AI systems: valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful bias managed. These characteristics provide a foundation for evaluating and improving the trustworthiness of AI systems and help organizations consider a broad range of AI-related risks beyond accuracy and reliability.

The NIST AI RMF is designed to be flexible and technology-neutral, allowing organizations of different sizes and industries to adapt it to their specific AI use cases. The framework is organized around four core functions: Govern, Map, Measure, and Manage. Together, these functions help organizations establish AI governance structures, understand AI system context and risks, assess and monitor those risks, and take appropriate actions to reduce or manage them.

NIST AI RMF's key components, such as functions, categories, and subcategories, provide progressively more detailed guidance for implementation.

Components of NIST AI RMF

As we read in the previous section, the NIST AI RMF is organized into a clear hierarchy of Functions, Categories and Subcategories and the framework includes four core Functions: Govern, Map, Measure, and Manage. Each Function is divided into Categories, which focus on specific areas of AI risk management. Each Category is further divided into Subcategories, which provide more specific outcomes that organizations can use to implement the framework. In simple terms, NIST AI RMF → Functions → Categories → Subcategories. The following table provides an overview of the Functions, the number of Categories and Subcategories under each, and their objectives:

Function	Categories	Subcategories	Objectives	Category Description
Govern	6	19	Establish the policies, procedures, responsibilities , and culture needed to manage AI risks.	1. Establish AI Risk Policies and Processes 2. Establish Accountability and Responsibility 3. Establish Diverse and Inclusive AI Governance 4. Establish an AI Risk Culture 5. Establish Stakeholder Engagement Processes 6. Establish Third-Party and Supply Chain Risk Processes
Map	5	18	Understand the AI system, its purpose, context, stakeholders, risks, benefits, and potential impacts.	1. Document Context 2. Categorize the AI System 3. Assess AI Capabilities, Benefits, and Costs 4. Map AI Risks and Benefits 5. Conduct AI impacts

Function	Categories	Subcategories	Objectives	Category Description
Measure	4	22	Assess and monitor AI risks, performance, and trustworthiness.	1. Apply AI Risk Metrics 2. Evaluate AI trustworthiness 3. Track AI Risks Over Time 4. Assess Measurement Effectiveness
Manage	4	13	Prioritize and treat AI risks and take appropriate corrective or risk response actions.	1. Manage AI Risks 2. Maximize AI Benefits and Minimize Impacts 3. Manage Third-Party AI Risks 4. Monitor, Improve, and Respond
Total	19	72		

Table 6.1: NIST AI RMF functions

As an AI GRC professional, you should focus on the subcategories because they provide the specific outcomes needed to implement the NIST AI RMF. One of the key advantages of the NIST AI RMF is its flexibility. You do not need to apply every Subcategory if it is not relevant to your organization. For example, an organization that only uses third-party AI tools may not need to apply some Subcategories related to developing its own AI models. The organization should assess each subcategory and focus on those relevant to its AI environment and risks. In the following subsections, we will discuss the practical aspects of implementing each function.

Govern

The primary objective of the Govern function is to establish organizational policies, processes, roles and responsibilities to ensure AI risks are identified, managed, and monitored throughout the AI lifecycle. It includes six categories that focus on defining governance structures, AI policies, roles and responsibilities, accountability mechanisms, integration with enterprise risk management, and promoting a culture of responsible AI. We will now go through expected

outcomes of the Govern function. Please note that these are summarized versions of the categories.

> **Note**
>
> You would like to use our ready-made implementation checklist available as *Template 7* in the GitHub repository. To access the repository link, follow the steps in the *Download the example code files* section in the *Preface*.

Developing policies and processes

This category requires the following outcomes:

- **Understand legal and regulatory requirements**: Identify, manage, and document laws and regulations related to AI.
- **Integrate trustworthy AI**: Include trustworthy AI principles such as fairness, privacy, security, transparency, and accountability in organizational policies and processes. For example, require an AI recruitment system to be tested for bias and fairness.
- **Set risk management activities**: Determine the level of AI risk management required based on the organization's risk tolerance. For example, an AI system used in healthcare may require a more detailed risk assessment than a simple chatbot. We have discussed AI risk assessment in *Chapter 10*.
- **Establish transparent risk management**: Define clear and transparent policies and procedures for managing AI risks.
- **Monitor and review risks**: Regularly monitor and review the risk management process, with clear roles and review schedules.
- **Maintain AI system inventory**: Maintain an inventory of AI systems and allocate resources based on their risk level.
- **Safely decommission AI**: Establish processes to retire or phase out AI systems safely. For example, when replacing an outdated AI credit-scoring system, securely remove its data and disable access.

Establishing accountability and responsibility

- **Define roles and responsibilities**: Clearly document who is responsible for identifying, assessing, and managing AI risks. For example, assign the AI system owner responsibility for monitoring model accuracy, reviewing incidents, and reporting significant issues to the AI governance committee.

- **Provide AI risk training**: Train employees and partners on AI risks, policies, and responsibilities.
- **Ensure executive responsibility**: Senior management takes responsibility for AI-related risk decisions. For example, require the Chief Risk Officer and business head to approve the use of an AI model for loan approval.

Ensuring diversity, equity, inclusion, and accessibility

- **Use diverse teams**: Include people with diverse backgrounds, skills, experience, and expertise in AI risk decisions. For example, an AI recruitment system should be reviewed by HR, IT, legal, compliance, and representatives of the employees who will use the system.
- **Define human-AI responsibilities**: Clearly define responsibilities for humans and AI systems, including human oversight. For example, an AI-based credit appraisal system can rank loan applications based on credit risk, but a credit officer must review the AI recommendation and make the final approval or rejection decision.

Establishing AI risk culture

- **Promote a safety-first culture**: Encourage critical thinking and safety when designing, developing, deploying, and using AI. For example, before deploying an AI-based medical-support system, the organization should test the system using a wide range of patient cases to identify incorrect diagnoses, unsafe recommendations, and situations where the AI may provide unreliable results. The system should have clearly defined limits, require doctors to review critical recommendations, and provide a mechanism to report and investigate harmful outputs before and after deployment.
- **Document and communicate risks**: Record AI risks and potential impacts and communicate them to relevant stakeholders.
- **Enable testing and incident reporting**: Establish practices for AI testing, incident identification, and information sharing.

Engaging the stakeholders

- **Collect external feedback**: A process should be in place to obtain and consider feedback from users, affected people, and other stakeholders about AI risks and impacts.
- **Use feedback for improvement**: A process should be in place to incorporate relevant feedback into AI system design and implementation.

Managing third-party and supply chain risks

- **Manage third-party AI risks**: Establish policies to manage risks from third-party AI systems, software, data, and intellectual property. For example, before using a vendor-provided AI credit-scoring model, assess its data sources, security controls, privacy practices, accuracy, and licensing.
- **Prepare for third-party failures**: Have contingency plans for failures or incidents involving high-risk third-party AI systems or data.

Map

The Map function requires the organization to understand the context in which an AI system is developed and used. It includes five categories that focus on defining the AI system's purpose, identifying stakeholders, understanding data and model characteristics, assessing potential impacts, and identifying AI-related risks. The following are the expected outcomes of the Map function.

Establishing and understanding context

- **Understand intended purpose and context**: Identify the AI system's purpose, users, expected benefits and harms, applicable laws, and deployment environment. For example, before deploying an AI-based loan appraisal system, identify who will use it, what decisions it will support, applicable banking regulations, and the potential impact on customers.
- **Involve diverse expertise**: Involve people with different skills, backgrounds, domain knowledge, and user experience.
- **Align with organizational mission**: Ensure the AI system supports the organization's mission and goals.
- **Define business value**: Clearly define the business purpose and value of using the AI system. For example, the objective of an AI loan appraisal system may be to reduce loan processing time from three days to a few hours while maintaining decision accuracy.
- **Define risk tolerance**: Determine and document the level of AI risk the organization is willing to accept. For example, a bank may establish a maximum acceptable error rate for an AI credit-scoring model and require human review for high-risk applications.
- **Define system requirements**: Identify requirements such as privacy, security, and fairness, while considering wider social impacts.

Categorizing the AI system

- **Define AI tasks and methods**: Identify the tasks the AI will perform and the methods used, such as classification, generation, or recommendation.
- **Document AI limitations**: Document the AI system's limitations, possible uses of its outputs, and human oversight requirements.
- **Plan testing and evaluation**: Identify testing and evaluation needs, including data quality, representativeness, reliability, and validation. For example, test the loan model using historical applications from different customer groups to check accuracy and potential bias.

Understanding AI capabilities, benefits, and costs

- **Identify benefits**: Identify and document the potential benefits of the AI system.
- **Identify costs and harms**: Identify potential financial, operational, social, and other costs resulting from AI errors or failures.
- **Define usage scope**: Clearly define where and how the AI system should be used based on its capabilities and risks.
- **Ensure user competence**: Define the skills, training, standards, and certifications needed by AI operators and users.
- **Define human oversight**: Establish how humans will monitor, review, and intervene in AI decisions.

Mapping risks and benefits across the AI system

- **Identify component risks**: Identify legal, technology, intellectual property, data, and other risks, including those from third-party components.
- **Define risk controls**: Identify and document controls to manage risks in AI system components, including third-party technologies. For example, implement access controls, encryption, vendor assessments, model validation, and regular performance monitoring for a third-party AI credit-scoring system.

Assessing AI impacts

- **Assess impact**: Identify the likelihood and severity of positive and negative impacts of the AI system. We have discussed AI Impact assessment in *Chapter 9*.
- **Collect feedback**: Establish processes to engage stakeholders, collect feedback, and address expected or unexpected impacts. For example, allow customers to appeal AI-based loan decisions and use complaints and appeal results to identify and correct problems with the AI system.

Measure

The primary objective of the Measure function is to evaluate and measure AI risks, performance, and trustworthiness using appropriate metrics and testing methods. It includes four categories that focus on deciding how AI risks should be measured, assessing whether the AI system is trustworthy, monitoring risks over time, and using feedback to improve the measurement process. Let's go through expected outcome of the Measure function:

Applying AI risk metrics

- **Select risk metrics**: Choose appropriate methods and metrics to measure identified AI risks, starting with the most significant risks. For example, measure the accuracy, false positive rate, and false negative rate of an AI-based fraud detection system.
- **Review metrics and controls**: Regularly review whether the metrics and controls are effective and update them when needed. For example, review a fraud detection model every quarter and adjust controls if false alerts increase significantly.
- **Use independent assessment**: Involve independent experts, domain experts, users, and affected communities in assessments when appropriate.

Evaluating AI trustworthiness

- **Document testing**: Record test sets, metrics, and tools used for testing and evaluation.
- **Ensure representative testing**: Ensure tests involving people meet applicable requirements and represent the relevant population. For example, test a facial-recognition system using images representing different ages, genders, and skin tones.
- **Measure performance**: Evaluate and document AI performance under conditions similar to actual use.
- **Monitor in production**: Continuously monitor the AI system and its components after deployment.
- **Validate reliability**: Confirm that the AI system is valid and reliable and document its limitations. For example, document that a credit model may be less reliable for customers with limited credit history.
- **Evaluate safety**: Regularly assess safety risks and ensure the system can fail safely within the organization's risk tolerance. For example, if an AI medical system cannot confidently assess a patient's condition, it should refer the case to a doctor rather than provide an uncertain recommendation.
- **Evaluate security and resilience**: Assess and document the system's security and ability to withstand failures or attacks.

- **Assess transparency and accountability**: Examine and document risks related to transparency and accountability. For example, clearly inform customers when they are interacting with an AI chatbot rather than a human. Also, assign a responsible person or team to review AI-related decisions, handle complaints, and take corrective action when the AI produces an incorrect or harmful outcome.
- **Explain AI models**: Explain and validate the AI model and interpret its outputs in the proper context. For example, a credit-scoring system should show the main factors that contribute to an applicant's risk score.
- **Assess privacy**: Identify and document privacy risks associated with the AI system. For example, check whether customer personal information used to train an AI model is properly protected and whether unauthorized data is being collected.
- **Evaluate fairness and bias**: Assess AI for fairness and bias and document the results. For example, compare loan approval rates across different customer groups to identify potential discriminatory outcomes.
- **Assess environmental impact**: Evaluate the environmental impact and sustainability of AI activities. For example, measure the electricity and computing resources used to train a large AI model and consider more efficient alternatives.
- **Review effectiveness**: Evaluate whether testing and evaluation methods are effective.

Tracking AI risks over time

- **Monitor emerging risks**: Regularly identify and track existing, unexpected, and new AI risks.
- **Track difficult-to-measure risks**: Develop approaches for tracking risks that cannot yet be measured effectively. For example, use expert reviews and customer feedback to monitor potential reputational or social risks that cannot be measured through a single metric.
- **Collect user feedback**: Provide ways for users and affected communities to report problems and challenge AI outcomes. For example, allow customers to appeal an AI-based loan rejection and submit supporting information for human review.

Assessing measurement effectiveness

- **Align measurement with context**: Ensure risk measurement methods are appropriate for the AI system's deployment context.

- **Validate results**: Use experts and relevant stakeholders to confirm that AI is performing as intended.
- **Track performance changes**: Identify and document improvements or declines in AI performance based on feedback and real-world data.

Manage

The Manage function requires the organization to prioritize and treat identified AI risks. It includes four categories that focus on managing AI risks, maximizing the benefits and minimizing the negative impact, managing third-party AI risks, and monitoring risk treatment, response, recovery, and communication activities. We will now look at expected outcomes of the Manage function.

Managing AI risks

- **Determine whether AI should proceed**: Confirm that the AI system meets its intended purpose and decide whether development or deployment should continue. For example, before deploying an AI-based credit appraisal system, verify that its accuracy, fairness, and reliability meet the organization's requirements.
- **Prioritize AI risks**: Address risks based on their impact, likelihood, and available resources. For example, address the risk of incorrect loan approvals before addressing a low-impact issue such as minor interface problems.
- **Develop risk responses**: Create and document responses for high-priority risks, such as mitigating, transferring, avoiding, or accepting the risk.
- **Document residual risks**: Record remaining risks that cannot be fully eliminated, especially those affecting users and downstream organizations. For example, document the possibility that an AI credit model may occasionally make an incorrect recommendation even after testing and controls are applied.

Maximizing benefits and minimizing impacts

- **Consider resources and alternatives**: Allocate resources to manage risks and consider non-AI alternatives where appropriate.
- **Maintain AI value**: Establish mechanisms to ensure the AI system continues to deliver its intended value after deployment.
- **Respond to new risks**: Follow procedures to respond to and recover from previously unknown risks.
- **Deactivate unsafe AI**: Establish processes to override, disengage, or deactivate AI systems that produce unsafe or unintended results.

Managing third-party AI risks

- **Monitor third-party risks**: Regularly monitor risks and benefits from third-party AI systems, software, and data.
- **Monitor pre-trained models**: Regularly monitor and maintain pre-trained models used in AI development.

Monitoring, improving, and responding

- **Post-deployment monitoring**: Monitor AI systems after deployment, including user feedback, appeals, overrides, incidents, recovery, decommissioning, and changes.
- **Continual improvement**: Regularly improve AI systems based on performance data and stakeholder feedback.
- **Manage incidents and errors**: Communicate AI incidents and errors to relevant stakeholders and follow documented processes for tracking, response, and recovery.

Understanding the structure of the NIST AI RMF is only the first step. The real value comes from applying its guidance to an organization's AI environment. Let us now look at how the NIST AI RMF can be implemented in practice, including practical steps, examples, and guidance for applying the relevant Subcategories.

Implementing NIST AI RMF

The NIST AI RMF provides a flexible approach to managing AI risks, but organizations may need practical guidance to translate the framework into specific actions.

> **Note**
>
> To support implementation, a ready-to-use checklist covering all the subcategories of the NIST AI RMF is available for reference in this book's GitHub repository as *Template 7*. To access the repository link, follow the steps in the *Download the example code files* section in the *Preface*.

The checklist provides practical guidance for each Subcategory and explains what an organization can do to implement it. It can be used by an AI GRC professional to assess the organization's current practices, identify gaps, define required actions, and track implementation progress.

The checklist can be used in the following manner:

1. **Review the applicable Subcategories:** Assess each Subcategory and determine whether it is relevant to the organization's AI environment.

2. **Understand the implementation outcome:** Review the practical guidance provided for each applicable Subcategory to understand what needs to be implemented.
3. **Assess the current state:** Determine whether the organization already has policies, processes, controls, or practices that address the Subcategory.
4. **Identify gaps:** Document areas where existing practices do not fully address the Subcategory.
5. **Define implementation actions:** Identify the specific actions required to address each gap, such as developing a policy, assigning responsibility, performing an assessment, implementing a control, or maintaining evidence.
6. **Assign responsibility:** Allocate each action to an appropriate owner or function and establish a target completion date.
7. **Maintain evidence:** Keep appropriate documentation and records to demonstrate that the Subcategory has been implemented.
8. **Review periodically:** Reassess the checklist when the organization introduces new AI systems, changes existing systems, or when AI-related risks and requirements change.

The checklist should not be treated as a simple compliance checklist. It should be used as an implementation tool to help the organization understand what each Subcategory means in practice, determine what is applicable, identify gaps, and establish evidence-based AI governance.

Other implementation support

NIST has developed additional resources to help organizations move from understanding the AI RMF to implementing it in practice. These resources provide practical guidance, examples, assessment approaches, and ways to connect the AI RMF with other standards and frameworks.

1. AI RMF profiles

AI RMF Profiles provide a structured way for an organization to describe its current AI risk management practices and define the desired future state.

A *Current Profile* describes the AI risk management outcomes that an organization currently achieves. A *Target Profile* describes the outcomes the organization wants to achieve. Following illustrative table will help you to understand the profile:

Category	Subcategory	Implementation Outcome	Current Profile	Target Profile
GOVERN	GOVERN 1.1	Legal and regulatory requirements involving AI are understood, managed, and documented.	Applicable AI laws and regulations are identified, but documentation and periodic review are not formally established.	All applicable AI laws and regulations are to be documented, assigned to responsible owners, and reviewed periodically.
GOVERN	GOVERN 1.2	Characteristics of trustworthy AI are integrated into organizational policies, processes, procedures, and practices.	Trustworthy AI principles are considered, but they are not consistently incorporated into policies and procedures.	Trustworthy AI principles to be integrated into AI policies, processes, procedures, and practices.
GOVERN	GOVERN 1.3	Processes are established to determine the level of risk management activities required based on organizational risk tolerance.	AI systems are assessed individually, but there is no consistent process for determining the level of risk management required.	AI risk management to be integrated with ERM, with AI-related risks documented and regularly updated in the organizational risk register.

Table 6.2: Example of Current and Target Profile

Organizations can compare the *Current* and *Target Profiles* to identify gaps in AI risk management, determine areas requiring improvement, prioritize actions based on risk and

available resources, develop an implementation plan, and track progress over time. Profiles are particularly useful for gap assessment and prioritizing AI risk management improvements. Our implementation checklist (*Template 7*) has provision to capture current as well as target profile for effective implementation.

2. AI RMF playbook

The AI RMF Playbook provides practical suggestions for implementing the AI RMF. It organizes implementation guidance around the four AI RMF Functions: Govern, Map, Measure, and Manage.

The Playbook helps organizations translate the high-level outcomes described in the AI RMF into practical activities. It provides suggestions that organizations can consider when establishing policies, identifying risks, evaluating AI systems, and managing identified risks. For example, for sub-category Govern 1.1 (which requires legal and regulatory requirements involving AI are understood, managed, and documented), playbook provides following guidance:

- Maintain awareness of the applicable legal and regulatory considerations and requirements specific to industry, sector, and business purpose, as well as the application context of the deployed AI system.
- Align risk management efforts with applicable legal standards.
- Maintain policies for training (and re-training) organizational staff about necessary legal or regulatory considerations that may impact AI-related design, development and deployment activities.

Organizations can document the following.

1. To what extent has the entity defined and documented the regulatory environment—including minimum requirements in laws and regulations?
2. Has the system been reviewed for its compliance with applicable laws, regulations, standards, and guidance?
3. To what extent has the entity defined and documented the regulatory environment—including applicable requirements in laws and regulations?
4. Has the system been reviewed for its compliance with relevant applicable laws, regulations, standards, and guidance?

We have mapped all the playbook guidance in our implementation checklist (*Template 7*) for easy reference.

3. AI RMF crosswalks

AI RMF crosswalks help organizations understand the relationship between the NIST AI RMF and other relevant standards, frameworks, guidelines, and regulations.

A crosswalk maps concepts or requirements from one framework to corresponding concepts in another framework. This can help organizations avoid duplicating work when they are already using other risk management or governance frameworks.

For example, an organization may already have an information security management system based on ISO/IEC 27001. Using an AI RMF crosswalk, the organization can identify areas where existing security controls support AI RMF requirements and determine where additional AI-specific controls may be needed. You can access the document from: `https://airc.nist.gov/airmf-resources/crosswalks/`

Summary

In this chapter, we discussed in detail how the NIST AI RMF provides a flexible approach for managing AI-related risks and can be adapted to different organizations and AI use cases. We saw that the framework is organized around four functions: Govern, Map, Measure, and Manage, with Categories and Subcategories providing more detailed guidance. We were able to conclude that organizations should focus on the Subcategories that are relevant to their AI environment and risk profile rather than applying every Subcategory in the same way. Effective implementation of this framework requires clear accountability, an understanding of AI system risks and impacts, appropriate measurement and monitoring, and timely risk treatment.

We saw various examples that illustrated how NIST AI RMF can be integrated with existing governance and risk management processes rather than creating a completely separate structure. We also discussed in the Manage function that AI risk management should be treated as an ongoing activity because AI systems, data, business requirements, and risks can change over time.

While frameworks like the NIST AI RMF focus on managing AI risks at an organizational level, global principles play an equally important role in shaping how AI should be developed and used across societies. One of the most influential sets of global principles for Responsible AI has been developed by the **Organization for Economic cooperation and Development** (**OECD**).

In the next chapter, we will explore the OECD AI principles, understand their core values, and examine how they influence AI regulations, governance frameworks, and organizational AI practices worldwide.

7

AI Governance Framework – OECD AI Principles

In the previous chapters, we explored key AI governance concepts, responsible AI principles, and major global regulations and frameworks. We now turn our attention to the OECD AI Principles, which provide an internationally recognized foundation for the responsible development and use of AI.

While regulations and frameworks provide structure and enforcement mechanisms, principles help define the underlying values that AI systems should respect. The AI principles defined by the **Organization for Economic Co-operation and Development** (**OECD**) play a critical role in this context by offering a shared, international understanding of what trustworthy and responsible AI should look like.

The OECD AI Principles were among the first globally agreed principles on artificial intelligence, endorsed by governments across the world. They focus on ensuring that AI benefits people and society, respects human rights and democratic values, operates transparently, remains robust and secure, and includes clear accountability. These principles have influenced many national AI strategies, regulations, and governance frameworks, including those discussed in earlier chapters.

In this chapter, we will examine the OECD AI Principles in detail. We will understand their intent, explore each principle individually, and discuss how organizations can apply them in practice. This chapter will help AI GRC professionals and practitioners align their AI governance programs with globally accepted ethical and societal expectations. Let us first understand what OECD is.

What is OECD?

The OECD is an international organization established in 1961 that brings together governments from around the world to promote economic growth, stability, and social well-being. It provides a platform for countries to collaborate, share best practices, and develop evidence-based policies on economic, technological, and social issues. It has **38 member countries**.

The OECD is widely respected for developing globally accepted guidelines, recommendations, and principles in areas such as digital economy, data governance, innovation, and artificial intelligence. While OECD instruments are not legally binding, they carry strong international influence and often shape national policies, regulations, and global standards. Understanding the OECD and its role provides useful context. Let us now look at how the OECD AI Principles were developed and how they have evolved over time.

Background and evolution of the OECD AI Principles

In 2018, OECD initiated a multi-stakeholder process involving governments, industry experts, academia, civil society, and technical specialists. The objective was to create principles that were practical, future-ready, and aligned with democratic values and human rights, while still supporting innovation and economic growth.

In 2019, the OECD formally adopted the OECD AI Principles, making them the first intergovernmental standard on artificial intelligence. These principles were endorsed by OECD member countries and several non-member countries, giving them global legitimacy and reach. At the same time, the OECD also introduced recommendations for governments on how to support trustworthy AI through policy, research, and international cooperation.

The OECD AI Principles were significantly updated in May 2024 to reflect developments in AI, including generative AI, and emerging concerns related to safety, information integrity, privacy, intellectual property, and environmental sustainability.

Over time, the OECD AI Principles have become a foundational reference for AI governance worldwide. They have directly influenced national AI strategies, international standards, and regulatory initiatives such as the EU AI Act, ISO/IEC AI standards, and various risk-based AI governance frameworks. Rather than being static, the principles are designed to remain relevant as AI technologies evolve, emphasizing adaptability, accountability, and societal benefit. The development of the principles explains their background, but it is equally important to understand where and to whom they apply. Let us now look at their scope and applicability.

Scope and applicability of the OECD AI Principles

The OECD AI Principles are designed to be broad, flexible, and globally applicable. They apply to a wide range of stakeholders involved in the AI ecosystem, rather than being limited to a specific industry, technology, or country. The principles apply to governments, public sector organizations, private organizations, AI developers, AI deployers, and organizations that procure or use AI systems. They are relevant across the entire AI lifecycle, including design, development, testing, deployment, operation, and retirement of AI systems.

The OECD AI Principles are technology-neutral, meaning they are not tied to a specific type of AI such as machine learning, generative AI, or autonomous systems. This allows them to remain relevant as AI technologies evolve over time. From a regulatory perspective, the OECD AI Principles are not legally binding. However, they strongly influence national AI strategies, laws, regulations, and international standards. Many governments use these principles as a foundation when drafting AI regulations or policy frameworks. Knowing the OECD AI Principles gives AI GRC practitioners a common international vocabulary and structure for developing internal AI governance policies. Let us now examine the five principles that form the foundation of the OECD approach to trustworthy AI.

The five OECD AI Principles

The OECD has defined five core principles to promote the responsible and trustworthy development and use of AI. Together, these principles provide a value-based foundation to ensure that AI systems benefit society and respect human rights. They also promote transparency, safety, security, and clear accountability. Before discussing each principle in detail, it is important to understand that these principles are interrelated and should be applied collectively rather than in isolation.

- **Principle 1, Inclusive growth, sustainable development, and well-being**: This principle emphasizes that AI should contribute positively to economic growth, social progress, and environmental sustainability. AI systems should be designed and used in ways that benefit people and society as a whole, rather than concentrating benefits among a few. For businesses, this means using AI to improve productivity, innovation, and customer experience while also considering broader societal and environmental impacts. For society, it means ensuring AI supports long-term well-being and does not deepen inequalities.
  AI can provide significant benefits for environmental sustainability, but developing and operating AI systems can also consume substantial amounts of energy, water, and computing resources. The 2024 update to the OECD AI Principles explicitly strengthened the focus on environmental sustainability. Therefore, organizations

should consider the environmental impact of AI throughout its lifecycle and, where appropriate, look for ways to improve energy efficiency and reduce resource consumption.

- **Principle 2, Respect for the rule of law, human rights and democratic values, including fairness and privacy:** This principle focuses on respecting human rights, democratic values, and diversity. AI systems should be fair, non-discriminatory, and inclusive, and they should not undermine human dignity or freedoms. Organizations are expected to identify and mitigate bias, avoid unfair treatment of individuals or groups, and consider the societal impact of AI-driven decisions. Human-centered AI places people at the core of system design and use.
- **Principle 3, Transparency and explainability:** Transparency and explainability ensure that people understand when AI is being used and how it influences decisions. This principle requires organizations to disclose the use of AI where appropriate and provide meaningful explanations of AI-driven outcomes, especially in high-impact scenarios. Transparency builds trust, supports accountability, and enables affected individuals to question or challenge AI decisions.
- **Principle 4, Robustness, security, and safety:** This principle requires AI systems to be reliable, secure, and safe throughout their lifecycle:
    - **Reliable** means that the AI system performs consistently and as intended under normal and reasonably foreseeable conditions
    - **Secure** means that the system and its data are protected against unauthorized access, attacks, manipulation, and other security threats
    - **Safe** means that the system does not cause unreasonable harm to people, property, or the environment when used as intended or in reasonably foreseeable ways

  For example, a loan approval AI that produces consistent results is *reliable*; protecting it from hacking and unauthorized modification makes it *secure*; and ensuring that its operation does not cause unreasonable harm to customers makes it *safe*.

  Organizations should address risks such as system failures, adversarial attacks, misuse, and unintended behavior, and ensure that AI systems can be monitored, tested, controlled, and, where necessary, safely overridden, repaired, or decommissioned.
- **Principle 5, Accountability:** Accountability ensures that there are clear roles and responsibilities for AI systems and their outcomes. Organizations and individuals involved in developing, deploying, or operating AI must be accountable for proper functioning, compliance, and overall impact. This includes having governance

structures, oversight mechanisms, and processes to address issues caused by AI systems.

Together, these five OECD AI Principles provide a global ethical and governance baseline that organizations can use to shape responsible AI policies, governance frameworks, and operational practices. The five principles provide a high-level foundation, but principles alone are not enough to govern AI effectively. Organizations need to translate them into practical policies, processes, and controls. Let us now look at how these principles can be applied in practice.

Practical application of OECD AI Principles in organizations

The OECD AI Principles are intentionally high-level and value-driven. To be effective, organizations must translate these principles into concrete policies, processes, and controls that can be implemented, monitored, and improved over time. Let us understand how each principle can be implemented in practice.

Principle 1: Inclusive growth, sustainable development, and well-being

This principle is important because AI should create benefits for society while avoiding increased inequality and unnecessary environmental impact.

- **Assess social and economic impact**: Evaluate the potential social and economic impacts of AI use cases before approval. For example, before deploying an AI hiring system, the organization can assess whether it may disadvantage certain groups or create other negative social impacts.
- **Promote inclusion and accessibility**: Ensure that AI systems can be used fairly and effectively by different groups, including people with disabilities and underserved communities. For example, an AI-based customer service system can provide accessible interfaces for users with visual or hearing impairments.
- **Reduce environmental impact**: Consider the energy and resource consumption of AI models and computing infrastructure and identify opportunities to improve efficiency. For example, an organization can select an energy-efficient AI model when it provides comparable results to a more resource-intensive model.
- **Monitor positive outcomes**: Define and monitor appropriate KPIs to ensure that AI continues to deliver positive social, economic, and environmental outcomes. For example, KPIs can include productivity, customer satisfaction, accessibility, and energy consumption.

Principle 2: Respect for the rule of law, human rights, and democratic values, including fairness and privacy

Organizations should translate this principle into practical safeguards that protect human rights, promote fairness, and ensure appropriate human oversight throughout the AI lifecycle. Following are some of the methods:

- **Design for people**: Consider human rights, dignity, and societal impact when designing AI systems. For example, before deploying an AI system for employee evaluation, the organization can assess how it may affect employees and their opportunities.
- **Identify and mitigate bias**: Assess training data, models, and outputs for potential bias and discriminatory outcomes. For example, an AI hiring system can be tested across different demographic groups to identify whether it unfairly rejects candidates from a particular group.
- **Ensure human oversight**: Maintain appropriate human involvement in high-impact or sensitive AI decisions. For example, a human reviewer can review an AI-generated loan rejection before the final decision is communicated to the customer.
- **Provide review and appeal mechanisms**: Allow individuals to question, appeal, or seek human review of AI-driven decisions. For example, a customer can request a review if an AI system incorrectly denies a service.

These steps help organizations ensure that AI systems are fair, inclusive, and aligned with human-centered values in real-world use.

Principle 3: Transparency and explainability

Organizations should put this principle into practice through appropriate disclosure, documentation, and explainability measures. Following are some of the practical aspects:

- **Disclose AI use**: Clearly inform users when they are interacting with an AI system, especially in customer, employee, or public interactions. For example, a chatbot can inform customers that they are communicating with an AI assistant.
- **Document AI systems**: Document the purpose, data sources, key model information, and limitations of the AI system. For example, an organization can maintain an AI system record describing what the system does, what data it uses, and its limitations.

- **Explain AI decisions**: Provide meaningful explanations for important AI-driven decisions. For example, a loan applicant should be given understandable reasons when an AI system contributes to a loan rejection.
- **Maintain records**: Keep appropriate records and logs to support audits, investigations, and accountability. For example, an organization can maintain logs showing the inputs, outputs, and significant decisions made by an AI system.

These actions help build trust, support accountability, and enable responsible use of AI across the organization.

Principle 4: Robustness, security, and safety

Organizations should ensure that AI systems perform as intended, are protected from attacks and misuse, and do not cause unreasonable harm. This principle can be put into practice through the following measures:

- **Ensure reliability**: Design AI systems to perform as intended under normal and unexpected conditions. For example, an AI customer service system should handle unexpected inputs without producing harmful or misleading responses.
- **Test and validate**: Regularly test AI systems to ensure they remain accurate, stable, and fit for their intended purpose. For example, an AI medical diagnosis system can be regularly tested against validated data to identify performance issues.
- **Protect AI systems**: Implement security controls to protect AI models, data, and infrastructure from unauthorized access, manipulation, and attacks. For example, access to an AI model and its training data can be restricted to authorized users.
- **Monitor and respond**: Monitor AI systems for performance issues, data drift, abnormal behavior, and security incidents, and maintain fail-safe and incident-response procedures. For example, an organization can automatically stop or route an AI system to human review when it detects unexpected or potentially harmful behavior.

These measures help ensure that AI systems remain safe, secure, and resilient throughout their lifecycle.

Principle 5: Accountability

In practice, clear ownership, roles, responsibilities, and oversight should be established throughout the AI lifecycle. This principle can be applied through the following measures:

- **Define AI ownership and accountability**: Clearly assign ownership for each AI system, including responsibility for performance, risk management, and compliance.

For example, an AI system owner can be responsible for ensuring that the system operates as intended and complies with applicable requirements.

- **Establish AI governance structures**: Establish appropriate governance structures, such as an AI governance committee, to oversee AI use, risks, and key decisions. For example, the committee can review and approve high-risk AI use cases before deployment.
- **Assign roles and responsibilities**: Define roles and responsibilities across the AI lifecycle, including development, deployment, monitoring, and retirement. For example, responsibility for monitoring an AI model can be assigned to a specific team with defined reporting and escalation responsibilities.
- **Maintain documentation and records**: Maintain documentation of AI decisions, risk assessments, approvals, changes, and other relevant activities to support traceability and audits. For example, an organization can maintain an AI system record containing its risk assessment, approval history, and significant changes.
- **Establish escalation and incident response**: Implement mechanisms to identify, escalate, and respond promptly to AI-related issues, errors, or harms. For example, significant AI incidents can be reported to the AI governance committee for investigation and corrective action.
- **Provide complaint and remediation mechanisms**: Provide affected individuals with channels to raise complaints, appeal AI-driven decisions, and seek appropriate remediation. For example, a customer can request human review when an AI system makes an incorrect decision that affects access to a service.

These actions ensure that AI systems are not treated as autonomous entities, and that clear human and organizational accountability exists for their outcomes. The OECD principles provide high-level guidance, while frameworks such as ISO/IEC 42001 and the NIST AI RMF provide more structured approaches for putting AI governance into practice. Let us now see how the OECD principles align with these frameworks.

Mapping OECD AI Principles to AI governance frameworks

The OECD AI Principles provide high-level, value-driven guidance on Responsible AI. To implement these principles in practice, organizations typically align them with structured governance frameworks such as **ISO/IEC 42001** and the **NIST AI Risk Management Framework** (**AI RMF**). These frameworks translate ethical principles into concrete policies, processes, controls, and risk management activities.

The following table shows a high-level alignment between the OECD AI Principles and the two widely adopted AI governance frameworks:

OECD AI Principle	**Alignment with ISO/IEC 42001**	**Alignment with NIST AI RMF**
Inclusive Growth, Sustainable Development, and Well-being	Addressed through AI policy, objectives, risk assessment, and continual improvement requirements (*Clauses 4–10*), including consideration of societal and environmental impacts	Reflected across Govern and Map functions by requiring consideration of impacts, benefits, costs, and societal outcomes
Human-Centered Values and Fairness	Implemented through data governance, risk management, human oversight, and bias mitigation controls in *Annex A*	Addressed through Map (impact identification), Measure (fairness and bias evaluation), and Manage (risk treatment)
Transparency and Explainability	Covered through documentation, communication, explainability, and disclosure controls in *Annex A* and *Clause 7* (Support)	Addressed through Map (system understanding), Measure (explainability evaluation), and Manage (communication and response)
Robustness, Security, and Safety	Implemented through lifecycle controls, monitoring, incident management, and security-related controls in Annex A	Addressed through Measure (safety, security, resilience evaluation) and Manage (risk response and recovery)
Accountability	Established through leadership commitment, defined roles and responsibilities, governance structures, audits, and management reviews	Central to the Govern function, with clear accountability, oversight, and decision-making responsibilities

Table 7.1: Mapping OECD AI Principles with ISO/IEC 42001 Standard and NIST AI RMF

In short, if your organization has already implemented the ISO/IEC 42001 standard or NIST AI RMF, all OECD AI Principles have been addressed appropriately.

Summary

In this chapter, we explored how the OECD AI Principles provide a global foundation for Responsible AI, promoting fairness, transparency, safety, and accountability.

We discussed that these five principles are voluntary and technology-neutral but have significantly influenced AI governance worldwide. They guide organizations in developing and using AI responsibly.

We also saw how frameworks such as ISO/IEC 42001 and NIST AI RMF help organizations translate these principles into practical governance, risk management, and controls, and discussed how organizations can put these principles into practice.

While the OECD AI Principles provide voluntary, value-driven guidance, governments are increasingly translating such principles into binding legal requirements. One of the most significant and comprehensive examples of this is the **European Union Artificial Intelligence Act** (**EU AI Act**).

In the next chapter, we will explore the EU AI Act in detail. We will understand its objectives, risk-based classification of AI systems, key obligations for organizations, and its impact on global AI governance and compliance expectations.

8

European Union AI Act

The **European Union Artificial Intelligence Act** (**EU AI Act**) is the world's first comprehensive legal framework created specifically to regulate **artificial intelligence** (**AI**). Its main objective is to ensure that AI systems used in the European Union are safe, trustworthy, transparent, and respect fundamental rights.

Unlike voluntary guidelines or ethical principles, the EU AI Act is legally binding. Organizations that develop, sell, deploy, or use AI systems that affect people in the EU must comply with its requirements, regardless of where the organization is located.

This chapter explains the EU AI Act from a **practical AI governance perspective**. Instead of legal complexity, it focuses on what organizations need to understand, implement, and demonstrate in real-world environments.

Global impact of the EU AI Act

The EU AI Act has an impact beyond the European Union. It is the first comprehensive law on AI and also applies to organizations outside the EU. For example, a company located in India, the US, or another country may be subject to the Act if it places an AI system or general-purpose AI model on the EU market, or if the output of an AI system operated outside the EU is used in the EU.

Because it is the first major comprehensive AI regulation, other countries are closely monitoring the EU AI Act while developing their own AI laws and governance frameworks. It may therefore serve as an important reference for countries developing their approaches to AI regulation.

As an AI GRC officer, you should closely monitor the act and adopt best practices even if your organization is not currently subject to any law. You can visit the official website of the EU to

get updates about the Act at `https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai`

Applicability of the EU AI Act

The EU AI Act applies to a wide range of organizations and AI-related activities based on use and impact, not just location. It covers AI systems that are placed on the EU market, put into service within the EU, or whose outputs affect individuals in the EU. This means that both EU-based and non-EU organizations may be required to comply if their AI systems are used in contexts such as employment, financial services, healthcare, education, or access to essential services involving EU residents. The Act applies to AI providers, deployers (users), importers, and distributors, each with defined responsibilities across the AI lifecycle. As a result, any organization that develops, sells, or uses AI systems influencing people in the EU must assess its role and ensure compliance with the applicable requirements of the EU AI Act.

Before examining the requirements of the EU AI Act, it is important to understand what the Act considers an AI system and how this definition affects its applicability.

AI system

Under the EU AI Act, an AI system is defined as a machine-based system that operates with a certain degree of autonomy and is designed to achieve specific objectives by processing data using techniques such as learning, reasoning, or inference. The system generates output in the form of predictions, recommendations, or decisions, which influence physical or virtual environments.

The definition focuses on how the system functions (i.e. learning, reasoning, or inference) and the impact of its outputs (i.e. predictions, recommendations, or decisions), rather than on the complexity of the technology used. As a result, both advanced machine learning models and simpler data-driven systems may qualify as AI systems if they influence decision-making. Traditional software that relies solely on fixed, hard-coded rules without any inference or learning capability is generally not considered an AI system under the EU AI Act. Now, let us understand how AI systems are classified under different categories.

Classification of AI systems

Under the EU AI Act, AI systems are classified using a risk-based approach that determines the level of regulatory control based on the potential harm an AI system may cause to individuals, society, or fundamental rights. The Act categorizes AI systems into **unacceptable risk**, **high risk**, **limited risk**, and **minimal risk**.

AI systems posing *unacceptable risk* are prohibited due to their inherent threat to fundamental rights. *High risk* AI systems are permitted but subject to strict compliance requirements such as

risk management, data governance, human oversight, and documentation. *Limited risk* AI systems are mainly subject to transparency obligations, while *minimal risk* AI systems face little or no regulatory burden.

This classification framework ensures that regulatory obligations are proportionate to the risks associated with each type of AI system and forms the foundation of compliance under the EU AI Act. Let us understand each classification with examples:

Risk Category	Description	Examples
Unacceptable risk (Prohibited) AI	AI systems that are a clear threat to people's fundamental rights and safety. These systems are not allowed.	• Social scoring systems that rank people based on behavior or personal traits. • AI that uses subliminal techniques to manipulate people • Real-time biometric surveillance in public places (with limited exceptions) • Emotion recognition in the workplace or schools
High risk AI	AI systems that can affect health, safety, or fundamental rights. Allowed only if strict conditions are met.	• AI used in job recruitment or employee evaluation. • AI used in credit scoring and loan approval • AI used in medical devices or treatment planning • AI used in critical infrastructure like water and power systems
Limited risk AI	AI systems with moderate risks that require some transparency.	• Chatbots • Emotion recognition tools • AI generating realistic images or video content

Risk Category	Description	Examples
Minimal risk AI	Low risk systems that do not require special legal rules.	• Email spam filters • AI used in video games • Recommendation engines for songs or movies

Table 8.1: Risk categories

Now we will discuss each of these risk categories in detail.

Prohibited AI

Under the EU AI Act, prohibited AI practices are AI uses that are considered to pose an unacceptable risk to fundamental rights, human dignity, and democratic values. The intent behind these prohibitions is to prevent AI systems that manipulate human behavior, enable indiscriminate surveillance, or cause significant harm that cannot be mitigated through controls or safeguards. The Act bans practices such as social scoring of individuals by public authorities, AI systems that exploit vulnerabilities of specific groups, certain forms of biometric identification in public spaces, and AI used for manipulative or deceptive purposes.

Organizations should be particularly cautious with AI applications involving behavioral profiling, emotion recognition, mass surveillance, or automated decision-making that significantly impact individuals without safeguards.

To ensure that prohibited uses do not slip into operations, organizations should maintain a clear AI inventory, conduct risk and use-case assessments before deployment, implement governance approval processes, and regularly review AI systems to confirm continued compliance with the EU AI Act.

High risk AI systems

Under the EU AI Act, high risk AI systems are those that can significantly affect individuals' rights, safety, or access to essential services. A common example is an AI-based recruitment screening system used to shortlist job applicants, where the output directly influences employment opportunities. Another example is a credit scoring system used by a bank to decide loan approvals. Because such systems can cause serious harm if they are inaccurate, biased, or poorly controlled, the EU AI Act imposes strict core requirements. The key requirements for high-risk AI systems include:

- **Risk management system (Article 9):** Organizations must establish and maintain a risk management system to identify, analyze, and mitigate risks associated with the AI system throughout its entire lifecycle, including development, deployment, and use.

- **Data and data governance (Article 10):** Training, validation, and testing data must be relevant, representative, and appropriate for the intended purpose of the AI system, with measures in place to detect and reduce bias and data-related risks.
- **Technical documentation (Article 11):** Detailed technical documentation must be prepared and kept up to date to demonstrate how the AI system is designed, how it operates, and how it complies with the requirements of the EU AI Act.
- **Record-keeping (Article 12):** High risk AI systems must enable automatic logging of events and decisions to ensure traceability, support audits, and allow investigation of incidents or malfunctions.
- **Transparency and provision of information to deployers (Article 13):** Clear instructions and information must be provided to users so they can understand the system's capabilities, limitations, and how to appropriately interpret and use its outputs.
- **Human oversight (Article 14):** Appropriate human oversight measures must be in place to monitor the AI system, intervene when necessary, and prevent or minimize risks to individuals and fundamental rights.
- **Accuracy, robustness, and cybersecurity (Article 15):** High risk AI systems must achieve suitable levels of accuracy, be resilient to errors or misuse, and include safeguards to protect against cybersecurity threats.
- **Conformity assessment and CE marking (Articles 43 and 48)**: Before a high risk AI system is placed on the EU market or put into service, the provider must conduct a conformity assessment to verify compliance with all applicable requirements. Once conformity is established, the system must carry the CE marking, indicating that it meets EU regulatory standards. Providers must ensure ongoing compliance through post-market monitoring and reassess conformity if the system is significantly modified or its intended purpose changes.
- **Post-market monitoring and incident management (Articles 61 and 62)**: Providers must implement a post-market monitoring system to continuously evaluate the AI system's real-world performance and identify emerging risks. Serious incidents or malfunctions that result in harm or rights violations must be reported to the relevant authorities without undue delay, along with appropriate corrective actions.

These requirements ensure that high risk AI systems are governed in a structured, transparent, and accountable manner, reducing harm while enabling lawful and responsible AI use.

Limited risk AI

Under the EU AI Act, limited risk AI systems are those that pose lower risks to individuals but still require specific safeguards to ensure transparency and informed use. A common example is a customer support chatbot used on a company website to answer frequently asked questions. Another example is an AI tool that generates images, videos, or audio content for marketing or social media purposes. While these systems do not usually make high-impact decisions about individuals, they can still mislead users if the AI involvement is not clearly disclosed. The key requirements for limited risk AI systems are included in **Article 50** which states:

- **Transparency obligations:** Organizations must clearly inform users when they are interacting with an AI system, such as a chatbot or virtual assistant, unless this is obvious from the context.
- **Disclosure of AI-generated content:** AI-generated or AI-manipulated content, such as deepfake images or videos, must be clearly labeled, so users understand that the content was produced by AI.
- **User awareness and informed consent:** Users should not be misled into believing they are interacting with a human or viewing authentic content when this is not the case.

These requirements ensure that users remain informed and aware of AI involvement, while allowing organizations to use limited risk AI systems with minimal compliance burden.

Minimal risk AI

Under the EU AI Act, minimal risk AI systems are those that pose little or no risk to individuals' rights, safety, or freedoms. These systems are largely exempt from regulatory obligations because they do not significantly affect decision-making or cause harm. Examples include AI used in video games, photo enhancement tools, spam filters, recommendation systems for movies or music, and AI features that improve user experience without influencing critical outcomes.

For minimal risk AI, the EU AI Act does not impose mandatory compliance requirements, allowing organizations to use and innovate freely. However, organizations are encouraged to follow voluntary codes of conduct and responsible AI practices to maintain trust and good governance, even where formal regulation does not apply. Understanding the different risk categories is useful, but organizations also need a practical way to determine where a particular AI system belongs. Let us now walk through the classification process.

Determining the classification of an AI system

To determine which classification a particular AI system falls under the EU AI Act, organizations should follow a structured, use-case-driven assessment rather than focusing only on the technology itself. The classification depends on the purpose of the AI system, the context in which it is used, and the potential impact of its outputs on individuals and society.

The first step is to clearly define the intended purpose of the AI system and understand how its outputs are used in decision-making. The organization should then verify whether the system falls under any prohibited AI practices, such as manipulative or social scoring use cases, in which case the system cannot be used. If the system is not prohibited, the next step is to assess whether it is used in a high-risk area listed in the Act, such as recruitment, creditworthiness, healthcare, education, biometric identification, or access to essential services. Systems used in these areas are classified as high risk and must meet strict compliance requirements.

If the system is not high risk, the organization should assess whether transparency obligations apply, for example in the case of chatbots or AI-generated content, which would classify the system as limited risk. AI systems that do not fall into any of these categories and do not significantly impact individuals' rights, safety, or freedoms are classified as minimal risk. Organizations should document the reasoning behind the classification and periodically review it to ensure continued compliance as the system or its use evolves.

The following table will help you determine the likely classification for these commonly used AI systems:

Common AI System	Typical Use Case	EU AI Act Classification
Recruitment and resume screening tools	Shortlisting, ranking, or rejecting job applicants	High risk AI
Credit scoring and loan approval systems	Assessing creditworthiness and loan eligibility	High risk AI
Medical diagnosis and clinical decision support systems	Supporting healthcare and treatment decisions	High risk AI
Biometric identification systems (non-public or controlled use)	Identity verification or access control in controlled environments	High risk AI

Common AI System	Typical Use Case	EU AI Act Classification
Biometric identification systems in public spaces (remote, real-time)	Mass surveillance or real-time identification of individuals	Prohibited AI
AI-based student assessment or exam proctoring tools	Evaluating students or monitoring examinations	High risk AI
Customer service chatbots	Responding to customer queries and support requests	Limited risk AI
AI-generated content tools (text, image, audio, video)	Creating content for marketing, media, or communication	Limited risk AI
Fraud detection and transaction monitoring systems	Detecting suspicious financial transactions	High risk AI
Recommendation systems (movies, music, products)	Suggesting content or products to users	Minimal risk AI
Spam filters and malware detection tools	Filtering unwanted emails or detecting threats	Minimal risk AI
AI in video games and entertainment	Enhancing gameplay or user experience	Minimal risk AI
Emotion recognition systems in workplaces or educational institutions	Inferring emotions or mental states of employees or students	Prohibited AI

Table 8.2: Classification for common AI systems

The classification depends on the intended purpose and real-world impact of the AI system. The same technology may fall into different categories based on how and where it is used. For example, consider an AI system that uses facial recognition technology. If it is used to unlock an employee's office door, it may be subject to relatively limited regulatory requirements. However, if the same technology is used by law enforcement for remote biometric identification in public spaces, it is prohibited under the EU AI Act. Thus, the technology itself

does not determine the classification; its intended purpose and the context in which it is used are important factors.

Also, it is advisable to seek legal opinion to determine the correct classification. Classification determines many of the applicable requirements, but compliance also depends on the organization's role in the AI lifecycle. In the next section, we will examine the responsibilities of providers, deployers, importers, and distributors.

Obligations by role

Under the EU AI Act, compliance obligations vary depending on the role an organization plays in the AI lifecycle. Clearly identifying whether an organization acts as an **AI provider**, **deployer**, **importer**, or **distributor** is essential for effective AI governance, as each role carries distinct responsibilities.

AI providers are organizations that develop AI systems or place them on the EU market or put them into service under their own name or trademark. Their obligations depend on the type and risk level of the AI system. In general, providers are responsible for ensuring that their AI systems meet the applicable requirements of the EU AI Act, which may include requirements related to transparency, risk management, data governance, technical documentation, testing, monitoring, and human oversight. Additional requirements apply to providers of high-risk AI systems, including conformity assessment, post-market monitoring, and reporting of serious incidents or malfunctions.

AI deployers **(Users)** are organizations that use AI systems in their operations. Their responsibilities include using AI systems in accordance with the provider's instructions, ensuring appropriate human oversight, monitoring system performance, and taking corrective actions if risks or issues are identified. Deployers of high-risk AI systems must also ensure that the system's use does not result in unlawful or discriminatory outcomes and must cooperate with authorities when required.

Importers are organizations that place AI systems from non-EU providers onto the EU market. They must verify that the AI system complies with the EU AI Act before making it available, ensure that conformity assessments have been completed, and confirm that technical documentation and required information are available. Importers also have a duty to act if they believe an AI system is non-compliant.

Distributors are organizations that make AI systems available on the EU market without modifying them. Their obligations include checking that the system carries the required conformity markings, ensuring that instructions and documentation are provided, and refraining from distributing AI systems they know or should reasonably know are non-compliant.

While providers, importers, and distributors have specific responsibilities under the EU AI Act, the affected person also needs mechanisms to raise concerns when AI systems may not comply with the Act. The Act therefore provides complaint and redress mechanisms to support accountability and protect individuals' rights.

Complaint and redress mechanism

The EU AI Act gives individuals and organizations clear rights to *complain* if they believe an AI system is not compliant. Key provisions include:

- **Right to lodge a complaint:** Any person has the right to file a complaint with their national **market surveillance authority** if they think an AI system violates the Act.
- **Right to an effective remedy:** Complainants must receive a proper response and have the right to challenge decisions in court if authorities fail to act.
- **Transparency from authorities:** National authorities must provide information on how to submit complaints and handle them in a transparent and timely manner.
- **Cross-border complaints:** Since the Act applies across the EU, complaints can be raised in any member state where the AI system is marketed or used.
- **Protection of fundamental rights:** The complaint process is designed to uphold individuals' fundamental rights (privacy, non-discrimination, etc.) when AI systems are misused or non-compliant.

These obligations are supported by enforcement mechanisms to dissuade noncompliance through various penalties and fines.

Penalties, fines, and enforcement

Under the EU AI Act, penalties and enforcement mechanisms are designed to ensure strong compliance and accountability across the AI lifecycle. The Act introduces a tiered penalty structure, where the severity of fines depends on the type and seriousness of the violation. The following image highlights how the EU AI Act's significant penalties can bring greater attention to AI compliance:

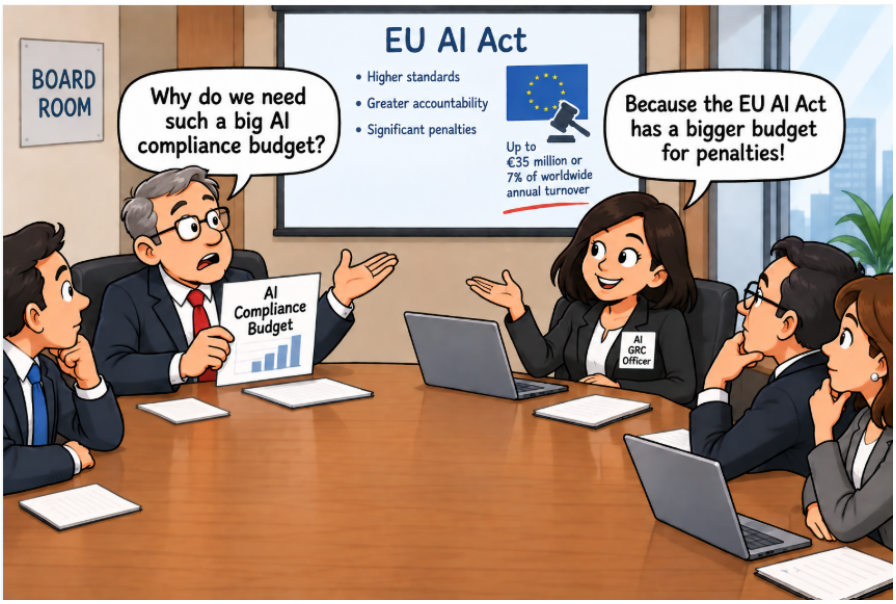


Image 8.1: EU AI Act – penalties

The most serious breaches, such as the use of prohibited AI practices, can attract fines of up to €35 million or up to 7% of the organization's total worldwide annual turnover, whichever is higher. Non-compliance with obligations related to high-risk AI systems, including risk management, data governance, or conformity assessment, can result in significant fines, though at a lower threshold. Supplying incorrect, incomplete, or misleading information to authorities is also subject to penalties. As an AI GRC professional, you should keep yourself updated about penalties levied over different organizations under the EU AI Act. You can track the same here: `https://www.gage.academy/tools/eu-ai-act-enforcement-tracker`

Enforcement is carried out by national competent authorities within EU member states, supported by EU-level coordination bodies. These authorities have the power to investigate AI systems, request documentation, conduct audits, order corrective actions, restrict or withdraw non-compliant AI systems from the market, and impose fines. Importantly, enforcement follows a risk-based and proportional approach, taking into account factors such as the nature of the infringement, intent or negligence, duration, and the organization's cooperation with authorities. From a governance perspective, these enforcement provisions make it clear that AI compliance is not optional and that organizations must treat AI risks with the same seriousness as other regulated business risks. The following table summarizes the penalty provisions:

Violation Type	Maximum Penalty
Use of prohibited AI systems (e.g., social scoring, manipulative AI)	€35 million or 7% of total worldwide annual turnover, whichever is higher.
Non-compliance with high-risk AI obligations	€15 million or 3% of total worldwide annual turnover, whichever is higher.
Supplying incorrect, incomplete, or misleading information to regulators	€7.5 million or 1.5% of turnover, whichever is higher.

Table 8.3: Penalty provisions

An AI GRC professional should not assume that an AI system remains in the same risk category throughout its lifecycle. If a new feature, change in intended purpose, or change in use could bring the system within the high-risk categories under **Article 6** and **Annex III**, the organization should reassess its classification. **For example,** a bank may initially use an AI system only to provide general information to customers. Later, it adds a feature that evaluates customers for creditworthiness and influences loan decisions. The bank should reassess the system because its new intended purpose may bring it within the high-risk classification rules.

AI GRC professionals should first assess the system against the **EU AI Act's classification rules, particularly Article 6 and Annex III**, and document the reasoning. If there is uncertainty, the organization can seek guidance from the **relevant national competent authority**, particularly the designated market surveillance authority in the Member State concerned.

Understanding the legal requirements is important, but AI GRC professionals ultimately need to translate them into practical governance activities, so we will now go through a step-by-step approach for implementing the EU AI Act.

Practical implementation roadmap

Implementing the EU AI Act in practice requires a structured, step-by-step approach that integrates legal, technical, and governance activities.

1. The first operational step is to create and maintain a complete inventory of all AI systems in use or under development, including third-party and embedded AI. Each system should then be assessed to determine whether it qualifies as an AI system under the Act and classified according to its risk level by the organization.

2. Once classification is completed, organizations should identify applicable obligations based on their role (provider, deployer, importer, or distributor) and the risk category of each AI system
3. Organizations should take appropriate action based on the classification of each AI system. For example, prohibited AI practices must not be used, or high-risk AI systems must meet the applicable requirements.
4. Organizations should then develop or update policies, procedures, and controls to embed these requirements into the AI lifecycle, from design and procurement to deployment and operation

The roadmap should also include training and awareness programs to ensure that relevant staff understand their responsibilities, as well as mechanisms for continuous monitoring, incident reporting, and regulatory engagement.

Most importantly, organizations should regularly review and update their AI governance framework to reflect changes in AI use, regulatory guidance, or business objectives. This practical, lifecycle-based roadmap helps organizations move from regulatory understanding to sustained compliance and responsible AI deployment.

AI GRC practitioners play an important coordinating role because they sit at the intersection of legal, product, data, engineering, risk, security, privacy, and compliance teams. They help translate regulatory requirements into practical governance requirements, coordinate assessments and evidence gathering, and ensure that the appropriate teams understand and fulfill their responsibilities.

Note

For your ready reference, we have created a template for managing requirements under the EU AI Act. Please refer to *Template 8* in this book's GitHub repository. To access the repository link, follow the steps in the *Download the example code files* section in the *Preface*.

AI GRC practitioners should implement the EU AI Act in consultation with Legal and relevant stakeholders, as its applicability and interpretation may vary based on the organization's circumstances and AI use cases. This chapter provides practical governance guidance and is not a substitute for legal advice.

Summary

In this chapter, we learnt that the EU AI Act is a legally binding, risk-based framework that regulates AI based on its potential impact on individuals and fundamental rights. We went through the different AI risk classifications and understood that the level of risk determines the level of regulatory requirements.

We also learnt that organizations must identify their role as a provider, deployer, importer, or distributor and apply the relevant requirements. For high-risk AI systems, we discussed the importance of risk management, data governance, human oversight, documentation, security, conformity assessment, and continuous monitoring. We discussed that it is important that AI systems should be reassessed when their purpose or functionality changes and that compliance is an ongoing governance responsibility.

One of the fundamental aspects in this chapter was determining the classification of the Systems. This is where **AI Impact Assessment** becomes critical. An AI Impact Assessment helps organizations systematically evaluate how an AI system may affect individuals, society, and fundamental rights, and whether adequate controls are in place to mitigate those risks. In our next chapter, we will discuss how to conduct an AI Impact Assessment.

9

AI Impact Assessment

An **AI Impact Assessment** (**AIIA**) is a structured process used to identify and evaluate the potential impacts and risks of an AI system before and after deployment. It helps organizations assess areas such as individual rights, compliance, safety, security, and reputation. Unlike traditional IT risk assessments, AIIA focuses not only on system failures but also on how AI decisions may affect people, fairness, transparency, and trust.

This chapter explains how AIIA can be practically implemented within an organization's AI governance framework. It focuses on what needs to be assessed, who should be involved, how impacts are identified and mitigated, and how decisions are documented and approved. The chapter also demonstrates how AIIA is not a one-time compliance activity, but an ongoing process that evolves throughout the AI system lifecycle. Before looking at how to conduct AIIA, it is important to understand why different frameworks, standards, and regulations expect organizations to perform such assessments and what they have in common.

AI impact assessment requirements under different frameworks, standards and regulations

Regulators and standards bodies increasingly expect organizations to perform AIIA as part of responsible AI governance. The EU AI Act, for example, requires high-risk AI systems to undergo systematic risk and impact analysis, while frameworks such as the **NIST AI Risk Management Framework** and **OECD AI Principles** emphasize continuous identification and mitigation of AI-related harms. The ISO/IEC 42001 standard requires the organization to establish a defined process for assessing how an AI system may affect individuals, groups of individuals, and society. This assessment must examine the potential consequences arising from the AI system's deployment, its intended use, and any reasonably foreseeable misuse. Under the **General Data Protection Regulation** (**GDPR**), organizations are required to

conduct a **Data Protection Impact Assessment** when processing activities are likely to result in high risk to individuals' rights and freedoms.

Despite differences in terminology and legal force, most frameworks share common expectations:

- Early identification of AI-related harms
- Focus on affected individuals and stakeholders
- Proportional assessment based on risk level
- Documentation and traceability
- Continuous monitoring and reassessment

Understanding the requirements provides the foundation, but the next question is what exactly should be assessed. Let us now look at the key areas that AIIA should consider.

Note

The definition of an AI system may vary across different laws, regulations, and standards. When conducting an AIIA, always consider the applicable legal or regulatory framework and the geographical context in which the AI system is used.

Benefits of AI impact assessment

AIIA helps organizations identify and address potential risks before and after an AI system is deployed. It also helps organizations understand and meet applicable legal and regulatory requirements. By identifying potential risks early, organizations can reduce financial, operational, legal, and reputational risks. It supports better AI governance by providing a structured approach to assess AI systems, define responsibilities, and implement appropriate controls.

It helps protect individuals and their rights by identifying concerns related to fairness, privacy, discrimination, and safety. It also improves transparency and builds trust among customers, employees, regulators, and other stakeholders.

It provides a basis for continuous monitoring and review as the AI system, its data, or its intended use changes, which is important, as discussed in the previous chapter.

Key areas of AI impact

An AIIA should evaluate the potential consequences of an AI system across multiple impact areas. These areas ensure that the assessment goes beyond technical performance and considers effects on different stakeholders. The depth of assessment should be proportionate

to the risk and context of the AI system. The following image highlights the importance of assessing an AI system's broader impact on customers, employees, and other stakeholders:

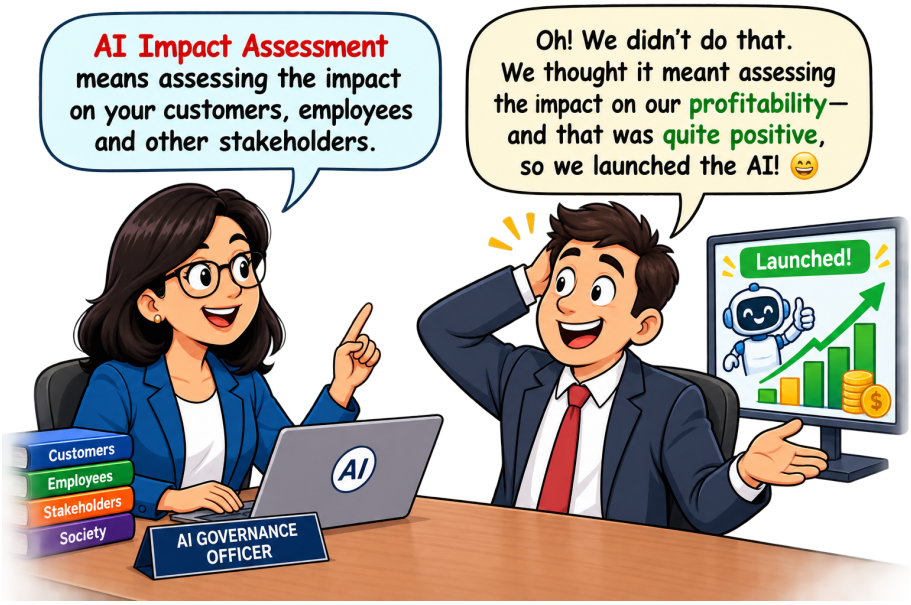


Figure 9.1: Areas of impact assessment

To build a complete picture of an AI system's potential impact, an AIIA should consider the following key areas:

- **Impact on individuals and fundamental rights:** This area assesses how the AI system may affect individuals or groups of individuals. It includes the risk of unfair treatment, discrimination, or bias, particularly for vulnerable groups. The assessment should also consider impacts on privacy, autonomy, dignity, and the ability of individuals to understand, challenge, or seek redress for AI-driven decisions. For example, an AI-based loan approval system may unfairly reject applications from a particular demographic group because the training data contains historical bias. This could result in discrimination and unfair access to financial services. An AIIA should identify such risks and recommend measures such as bias testing, human review, and an appeal mechanism.
- **Ethical and societal impact**: This impact area focuses on broader societal consequences of the AI system. It examines whether the AI system aligns with ethical principles such as fairness, transparency, accountability, and human oversight. It also considers social trust and public perception of the AI system. The assessment should

also consider long-term societal effects, such as exclusion, over-reliance on automation, or erosion of human judgment. For example, an e-commerce platform may use AI to offer discounts based on customer behavior, such as how frequently they shop, what products they view, how often they abandon their shopping cart, or how much they typically spend. If the AI provides different discounts to customers based on these behaviors, it could raise concerns about fairness, transparency, and customer trust.

- **Legal and regulatory impact:** This area assesses whether the AI system complies with applicable laws, regulations, and contractual requirements. It also considers data protection, sector-specific rules, cross-border requirements, and potential legal or regulatory risks. Say a bank uses AI to assess customers for loan approval. In this case, this bank must ensure that the system complies with applicable data protection laws, banking regulations, and requirements related to fair lending. Failure to comply could result in regulatory action, penalties, or legal claims.
- **Technical and operational impact:** This impact area addresses the technical behavior of the AI system and its operational reliability. It includes assessing whether the AI model produces accurate and consistent results, performs reliably under different conditions, and can provide understandable explanations for its outputs. The assessment should also identify the model's limitations, such as situations where it may produce incorrect results or may not be suitable for its intended purpose. For example, a bank's AI loan approval system may be highly accurate for most customers but produce incorrect results for customers with limited credit history. AIIA should assess the model's accuracy, test how reliably it performs in different situations, understand why it produces certain decisions, and identify limitations where human review may be required.
- **Safety and security impact:** This area focuses on risks to physical, digital, and information security. It evaluates whether the AI system could cause physical harm, unsafe outcomes, or security breaches. This includes risks from adversarial attacks, model manipulation, unauthorized access, and dependency on third-party components. For example, a bank's AI loan approval system could be manipulated by an attacker who changes input data to influence loan decisions. Unauthorized access to the AI model or customer data could also result in a security breach. AIIA should identify these risks and ensure appropriate security controls, access restrictions, testing, and monitoring are in place.

- **Organizational and business impact:** This impact area examines how the AI system may affect the organization itself. It includes potential reputational damage, financial loss, operational disruption, and impact on decision-making processes. It also considers workforce impacts, such as changes to roles, skills, and accountability. For example, if a bank relies heavily on an AI system for loan decisions and the system produces incorrect results, it could lead to financial losses, customer complaints, reputational damage, and regulatory action. It may also change the roles and responsibilities of loan officers, requiring them to develop new skills to review and oversee AI decisions.
- **Environmental impact : F**or certain AI systems, especially those with high computational requirements, environmental impacts should be considered. This includes energy consumption, resource usage, and sustainability implications of developing, training, and operating the AI system.
- **Impact on stakeholders and third parties:** This area assesses how the AI system may affect external stakeholders such as customers, suppliers, partners, and service providers. It includes dependencies on third-party AI models or data sources, shared responsibilities, and potential downstream impacts. For example, a company may use a third-party cloud-based AI service to analyze customer feedback. If the service provider experiences a data breach or the AI produces inaccurate results, the company's customers and business partners may be affected. AIIA should therefore assess third-party risks, data sharing, service dependencies, and clearly define the responsibilities of each party.

Not all impact areas will apply equally to every AI system. Organizations should assess impacts based on the AI system's purpose, risk level, and deployment context. A structured and documented evaluation of these impact areas supports compliance with standards and regulations while enabling responsible and trustworthy use of AI.

Steps to conduct effective AI impact assessment

An AIIA should be a structured process that helps an organization understand the potential effects of an AI system, identify risks, and determine appropriate safeguards. The following steps can be used in practice:

1. **Understand the AI system and its purpose**: Start by understanding what the AI system does, why it is being used, and who may be affected by its outputs.
   For example, if a bank plans to use an AI system to assess customers' creditworthiness and support loan approval decisions, it should document the system's purpose, users, data used, and customers affected.

2. **Identify potential impacts and risks**: Identify how the AI system could affect individuals, society, the organization, and other stakeholders. Consider areas such as fairness, privacy, legal compliance, safety, security, and reputation. Building on the example from the previous point, the bank may identify risks such as biased loan decisions, misuse of customer data, inaccurate predictions, or cybersecurity attacks
3. **Assess the severity and likelihood**: Assess how serious each identified risk could be and how likely it is to occur. This helps the organization prioritize risks that require greater attention. If the AI incorrectly rejects eligible customers, the impact could be significant because customers may lose access to financial services. If testing shows that this could happen frequently, the risk should receive a high priority.
4. **Determine and implement controls:** Determine measures to prevent, reduce, or manage the identified risks. These may include human oversight, bias testing, data protection, security controls, and monitoring. The bank, from our earlier example, could require a loan officer to review high-impact AI decisions, regularly test the model for bias, and monitor the system for inaccurate results.
5. **Document the assessment and decisions**: Record the identified impacts, risks, controls, responsibilities, and decisions made during the assessment. The bank should record the results of bias testing, the controls implemented, and the decision to approve the AI system for use.
6. **Obtain appropriate approval:** The assessment should be reviewed by relevant stakeholders such as risk, compliance, legal, information security, or the AI governance committee. The bank's AI governance committee should review the assessment and confirm that the identified risks have been appropriately addressed before approving the system for production use.
7. **Monitor and review**: The assessment should be regularly reviewed and updated when there are significant changes to the AI system, its purpose, data, regulations, or risk profile. For instance, if the bank later changes the AI model or starts using it for a different type of loan, it should reassess the potential impacts and determine whether additional controls are required.

The following image reinforces why an AIIA should be treated as an ongoing activity rather than a one-time assessment:



Figure 9.2: AIIA – it's not a one-time activity

Note

For your ready reference, a template on AIIA is available in this book's GitHub repository as *Template 9*. To access the repository link, follow the steps in the *Download the example code files* section in the *Preface*.

Summary

In this chapter, we discussed how AIIA helps organizations identify and evaluate the potential impacts and risks of an AI system. We learned that an assessment should consider areas such as individual rights, fairness, legal compliance, safety, security, technical performance, business impact, environment, and stakeholders. We also learned that an effective assessment involves identifying risks, evaluating their severity, implementing controls, documenting decisions, obtaining approval, and continuously monitoring the AI system.

By the end of this chapter, we understood that AIIA is not a one-time activity. It should be reviewed and updated when the AI system, its purpose, data, or risks change.

While AIIA focuses on understanding how an AI system may affect people, society, and the organization, impact identification alone is not sufficient. Organizations must also evaluate these impacts in terms of likelihood, severity, and acceptability, and decide how they should be treated.

The next chapter builds on the outcomes of the AIIA and explains how identified impacts are translated into measurable risks.

10

AI Risk Assessment

As organizations increasingly adopt **artificial intelligence** (**AI**) in business processes and decision-making, managing AI-related risks becomes essential. So far, we've seen that AI systems can influence hiring decisions, credit approvals, medical recommendations, security monitoring, and many other areas that directly affect individuals and organizations. If these systems are not properly assessed and controlled, they can cause unintended harm, legal non-compliance, and loss of trust. You can refer to varied consequences of unmonitored AI at `https://incidentdatabase.ai/`.

AI risk assessment provides a structured way to understand what can go wrong in an AI system, how serious the consequences could be, and what measures are required to reduce those risks. It helps organizations move from high-level AI governance principles to practical, operational decisions that support responsible and compliant AI use.

This chapter explains how to identify, analyze, evaluate, and treat risks across the entire AI lifecycle, from data collection and model development to deployment and continuous monitoring. It also highlights how AI risk assessment differs from traditional IT risk assessment and why continuous evaluation is necessary for AI systems that learn and evolve over time. Before assessing AI risks, it is important to understand the different types of risks that can arise from AI systems.

Types of AI risks

Understanding different types of AI risks helps organizations assess them systematically and apply appropriate controls. The following common AI risks were identified by reviewing different AI standards, frameworks, regulations, and risk taxonomies.

- **Technical risks:** These risks relate to the performance and reliability of the AI system. They arise when the AI model does not behave as expected or degrades over time. Common technical risks include model errors, inaccurate predictions, data drift, and performance degradation. For example, an AI model trained on historical data may perform poorly when real-world conditions change, leading to incorrect outputs. If technical risks are not monitored, the AI system may continue producing unreliable results without being detected.
- **Ethical risks:** These risks arise when AI systems produce outcomes that are unfair, biased, or discriminatory. They are often linked to biased training data, inappropriate feature selection, or a lack of fairness testing.
  Consider an AI-based recruitment system that disadvantages certain groups because the historical hiring data it was trained on reflects past bias. Ethical risks can harm individuals, violate organizational values, and undermine responsible AI principles.
- **Legal and regulatory risks**: These risks occur when AI systems fail to comply with applicable laws, regulations, or contractual obligations. They may result in penalties, legal actions, or enforcement measures.
  Examples include violations of data protection laws, failure to meet transparency or explainability requirements, or non-compliance with AI-specific regulations such as the EU AI Act. Organizations must ensure that AI systems meet legal obligations throughout their lifecycle.
- **Security risks:** These are threats that target the AI system, its data, or its models. AI systems are vulnerable to attacks such as model poisoning, adversarial attacks, and unauthorized access to training or inference data.
  For example, attackers may manipulate training data to influence AI outcomes or exploit model weaknesses to extract sensitive information. Poorly managed security risks can lead to data breaches and compromised decision-making.

- **Operational risks:** These risks arise from how an organization uses and manages AI systems. They include over-reliance on AI outputs, insufficient human oversight, unclear accountability, and inadequate operational procedures.
  For instance, if staff blindly trust AI decisions without review, errors may go unnoticed. Clear roles, responsibilities, and escalation mechanisms are necessary to manage operational risks effectively.
- **Reputational Risks:** These risks result from public perception and stakeholder trust in the organization's use of AI. Even technically correct AI systems can cause reputational harm if they are perceived as unfair, intrusive, or unethical.
  Negative media coverage, public backlash, or loss of customer confidence can occur when AI failures become visible. Reputational risks often amplify other risk types and can have long-term business impacts.

Understanding the different types of risks provides the foundation for assessment. Let us now see how major frameworks, standards, and regulations address AI risk assessment.

AI risk assessment requirements under different frameworks, standards, and regulations

AI risk assessment has become a core requirement across multiple AI governance frameworks, standards, and regulations. We have already discussed **ISO/IEC 42001**, **NIST AI RMF**, and the **EU AI Act** in our earlier chapters. While each framework, standard and regulation uses different terminology and levels of prescriptiveness, they all emphasize the need to identify AI-related risks, assess their potential impact, implement mitigation measures, and continuously monitor residual risks throughout the AI lifecycle.

ISO/IEC 42001 requires organizations to establish a structured and repeatable approach to identifying and managing AI risks as part of an **Artificial Intelligence Management System (AIMS)**. The standard expects organizations to assess risks related to data quality, model behavior, human oversight, and the impact of AI systems on individuals and society. Risk assessment under ISO/IEC 42001 is not a one-time activity and must be documented, reviewed periodically, and aligned with organizational objectives and stakeholder expectations.

The NIST AI RMF adopts a voluntary, risk-based approach that integrates AI risk assessment across governance, design, development, and deployment activities. It requires organizations to map AI risks and affected stakeholders, measure the likelihood and severity of potential harms, and manage risks through prioritization and controls. The framework places strong emphasis on continuous evaluation, transparency, and accountability rather than checklist-based compliance.

The EU AI Act establishes legally binding requirements for AI risk assessment, particularly for high-risk AI systems. It mandates the implementation of a continuous risk management system that identifies known and foreseeable risks to health, safety, and fundamental rights. Organizations must assess these risks throughout the AI lifecycle, apply mitigation measures, and maintain documented evidence to demonstrate compliance, including post-market monitoring and incident reporting.

Data protection laws such as the **General Data Protection Regulation** (**GDPR**) impose AI-related risk assessment obligations when AI systems process personal data or involve automated decision-making. These laws require organizations to evaluate risks to individuals' rights and freedoms, assess the impact of profiling and automation, and implement safeguards such as transparency, explainability, and human intervention. In practice, AI risk assessments are often closely aligned with **Data Protection Impact Assessments** for such systems.

The OECD AI Principles provide high-level guidance that influences many national regulations and organizational AI governance programs. Although not legally binding, the principles emphasize identifying and mitigating risks related to fairness, transparency, safety, and societal impact. They encourage organizations to adopt proactive and responsible risk assessment practices throughout the AI lifecycle.

Across all these frameworks, standards, and regulations, there is a clear and consistent expectation that AI risk assessment should be continuous, multidisciplinary, and focused on real-world impact. Technical reliability alone is not sufficient; organizations must also assess ethical, legal, operational, and societal risks. A well-designed AI risk assessment process can therefore support compliance with multiple frameworks simultaneously while strengthening overall AI governance. Different frameworks may use different terminology, but the basic approach to AI risk assessment remains similar. Let us now look at the key steps for identifying, analyzing, evaluating, and treating AI risks.

AI risk assessment steps

AI risk assessment follows a structured and logical sequence of steps that helps organizations understand AI-related risks and decide how to manage them. While the terminology may vary across frameworks, the core steps remain consistent: risk identification, risk analysis, risk evaluation, and risk mitigation, as shown in the following figure:

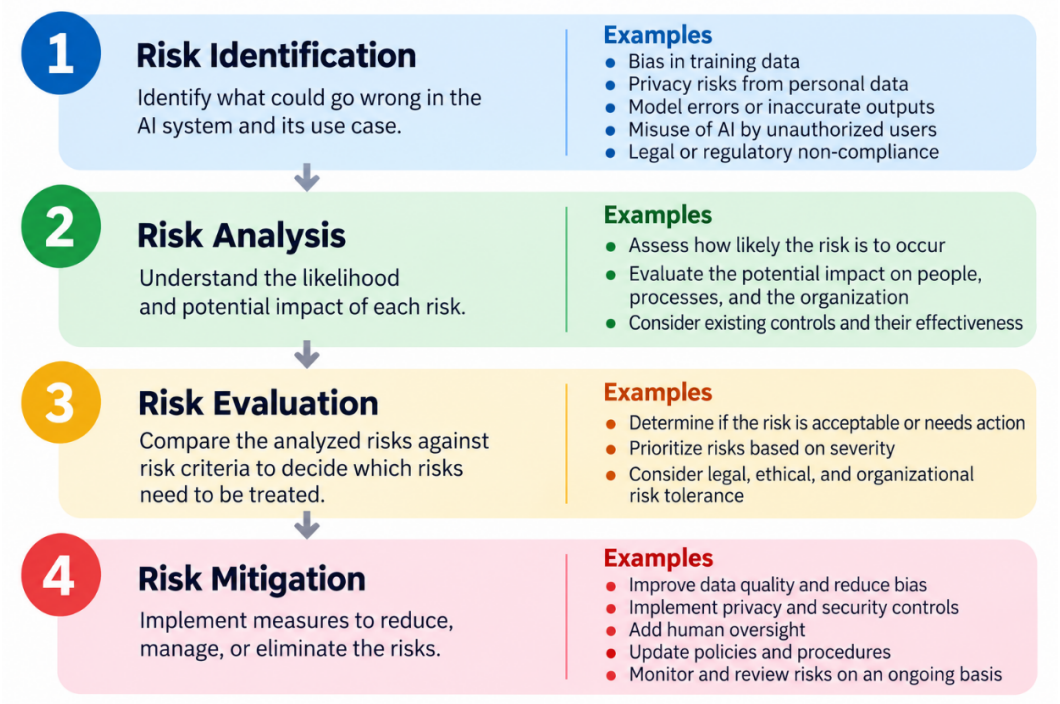


Figure 10.1: Risk assessment steps

Let us understand each step in detail.

Risk identification

Risk identification involves identifying what could go wrong in an AI system and why. At this stage, organizations list potential risks related to data, models, human oversight, security, legal compliance, and ethical impact. The focus is on identifying foreseeable and realistic risks across the AI lifecycle rather than assessing their severity. The following are some of the widely used risk identification techniques:

- **Brainstorming and expert workshops:** AI risk identification often begins with structured brainstorming sessions and expert workshops. Cross-functional teams, including data scientists, business owners, legal, compliance, and risk professionals, collaboratively review the AI use case to identify potential risk scenarios. This technique is especially effective for uncovering ethical, operational, and governance-related risks that may not be apparent through technical analysis alone.

- **Use of AI risk taxonomies and checklists**: AI risk taxonomies and checklists provide a structured way to identify common and known AI risks. These tools categorize risks such as bias, explainability, security, privacy, and safety, helping organizations ensure that no major risk category is overlooked. This technique is particularly useful for standardizing risk identification across multiple AI systems. Readers can find reference for AI risk taxonomies at `https://airisk.mit.edu/risks`.
- **Review of historical incidents and AI failures:** Reviewing past AI-related incidents and failures is a practical way to identify realistic and foreseeable risks. This includes internal incidents, industry case studies, regulatory enforcement actions, and publicly reported AI failures. Learning from past events helps organizations anticipate similar risks in their own AI systems.
- **Stakeholder and user impact analysis:** Impact analysis focuses on identifying risks based on who may be affected by the AI system. It includes employees, customers, third parties, and the wider public. Risks are identified by examining how AI decisions or errors could impact rights, fairness, safety, and trust.
- **Technical review**: A technical review helps identify risks by examining how the AI system is built and operates. It can reveal risks such as poor data quality, inaccurate model outputs, weak security, lack of robustness, system dependencies, or performance degradation
  For example, testing a loan approval model may show that it produces less accurate results for customers with limited credit history. This identifies a potential accuracy and fairness risk.
- **Model documentation review**: It helps identify risks by checking whether the documented information about the model is complete, accurate, and consistent with its actual use. Missing information about the model's limitations, data sources, assumptions, or intended use can itself indicate a governance or compliance risk.
  For example, if the documentation does not clearly state that the loan model has not been tested for certain customer groups, this gap can indicate a potential fairness and model performance risk.
- **Lifecycle-based risk identification:** This ensures that risks are identified at every stage of the AI system's lifecycle. Different risks may arise during data collection, model development, testing, deployment, operation, and retirement. This approach helps organizations manage AI risks continuously rather than only at a single point in time.

Risk analysis

Risk analysis focuses on understanding the nature of identified risks and assessing their potential impact and likelihood. This step examines how severe the harm could be if the risk materializes and how likely it is to occur. Both technical consequences and impact on individuals, business operations, and compliance are considered.

Risk can be analyzed using several approaches depending on the organization's maturity, the criticality of the AI system, and regulatory expectations. Let's discuss these techniques of risk analysis:

- **Qualitative risk analysis** uses descriptive scales such as *High*, *Medium*, and *Low* to assess likelihood and impact. This approach relies on expert judgment and predefined criteria rather than numerical data. It is easy to apply, widely used in early-stage AI programs, and suitable when precise data is not available.
- **Semi-quantitative risk analysis** combines qualitative judgment with numerical scoring. Likelihood and impact are assigned numerical values, such as *1* to *5*, which are then combined to calculate a risk score. This approach provides better prioritization while remaining relatively simple and understandable.
- **Quantitative risk analysis** uses detailed numerical data, statistical models, or probabilities to estimate risk. It may involve metrics such as financial loss, probability distributions, or error rates. Quantitative analysis is more complex and is typically used for high-impact or regulated AI systems where sufficient data is available.

Likelihood and impact

Risk analysis usually considers two key factors: likelihood and impact. **Likelihood** refers to how probable it is that a risk will occur, while **impact** refers to the severity of harm if the risk materializes. Each of these factors can be rated as *High*, *Medium*, or *Low* based on predefined criteria.

For example, a risk may be rated High likelihood if it is expected to occur frequently or has already occurred in similar AI systems. Impact may be rated High if the risk could cause significant harm to individuals, lead to regulatory penalties, or seriously damage organizational reputation.

Once likelihood and impact are assessed, the overall risk level is determined. A risk with High likelihood and High impact is typically classified as High risk. A risk with Medium likelihood and Low impact may be classified as Low risk, while a risk with Low likelihood and Medium impact is generally classified as Low risk. Organizations need to have their own risk analysis chart.

See the following chart as an example:

Impact	**Likelihood**		
	High (3)	Medium (2)	Low (1)
High (3)	9	6	3
Medium (2)	6	4	2
Low (1)	3	2	1

Table 10.1: Impact likelihood table

Risk Score (Impact * likelihood)	**Risk Level**
0 - 3	Low
4 - 6	Medium
Above 6	High

Table 10.2: Classification table

The first part of the table shows how the risk score is calculated by multiplying impact and likelihood values. For example, a risk with High impact (3) and High likelihood (3) results in a risk score of 9. A risk with High impact (3) and Medium likelihood (2) results in a score of 6, while a risk with Low impact (1) and Low likelihood (1) results in a score of 1. This simple multiplication approach provides a consistent way to compare different risks.

The second part of the table explains how risk scores are classified into risk levels. Risk scores from 0 to 3 are categorized as Low risk. These risks usually have limited impact and can be managed through standard controls and routine monitoring. Risk scores from 4 to 6 are categorized as Medium risk. These risks require attention, defined mitigation measures, and periodic review. Risk scores above 6 are categorized as High risk and demand immediate action, strong mitigation controls, and senior management oversight.

In practice, this risk matrix helps organizations prioritize AI risks in a structured and transparent manner. High-risk AI issues, such as biased decision-making in critical systems, are addressed first, while lower-risk issues are monitored and managed appropriately. The approach is simple, easy to communicate, and suitable for qualitative and semi-quantitative AI risk assessments.

> **Note**
>
> For your ready reference, a template of a structured AI risk assessment register is available as *Template 10* in this book's GitHub repository. To access the repository link, follow the steps in the *Download the example code files* section in the *Preface*.

Risk evaluation

Risk evaluation involves comparing analyzed risks against predefined risk acceptance criteria to determine whether the risk is acceptable, tolerable with controls, or unacceptable. This step helps prioritize risks and supports management decision-making.

In the recruitment system, for example, the organization may determine that high and medium risks are unacceptable and need to be mitigated.

Risk mitigation/treatment

Risk treatment is the process of deciding how identified and evaluated AI risks will be addressed and implementing appropriate actions to manage them. After risks are analyzed and evaluated, organizations must determine whether each risk should be avoided, reduced, transferred, or accepted. The objective of risk treatment is to ensure that residual AI risks remain within the organization's defined risk appetite.

Risk treatment decisions should be proportionate to the severity and impact of the AI risk. High risks, especially those affecting individuals' rights, safety, or legal compliance, require strong and well-documented treatment measures. Lower risks may be managed through standard controls and ongoing monitoring. A risk can be treated in one of the following ways:

- **Avoidance**: Risk avoidance involves deciding not to proceed with an AI system or a specific AI functionality when the associated risks are unacceptable and cannot be adequately controlled. This may include canceling an AI project, limiting its scope, or removing automated decision-making features in sensitive use cases.
- **Reduction/mitigation:** Risk reduction focuses on implementing technical, organizational, and procedural controls to lower the likelihood or impact of AI risks. Examples include improving training data quality, conducting bias and fairness testing, enhancing explainability, introducing human-in-the-loop controls, strengthening security safeguards, and implementing continuous performance monitoring.
- **Transfer:** Risk transfer involves shifting part of the AI-related risk to third parties through contractual arrangements or insurance. This may include vendor warranties, indemnity clauses, service-level agreements, or reliance on certified third-party service

providers. While risk transfer reduces exposure, it does not remove the organization's accountability for AI outcomes.

- **Acceptance:** Risk acceptance applies when residual AI risks are assessed as low or within acceptable limits. In such cases, the organization formally accepts the risk and continues to monitor it. Risk acceptance decisions should be documented, approved by appropriate management, and reviewed periodically, especially when the AI system or its use changes.

Effective risk assessment requires a consistent approach across the organization. Let us now look at how a risk management policy can establish the overall approach for identifying, assessing, treating, and monitoring AI risks.

Risk management policy

The purpose of the **risk management policy** is to establish a structured and consistent approach for identifying, assessing, treating, and monitoring risks. The policy aims to ensure that risks are managed responsibly, in alignment with organizational objectives, ethical principles, and applicable legal and regulatory requirements.

Ideally, organizations should maintain a single, enterprise-level risk management policy that covers all types of risks, including strategic, operational, financial, cybersecurity, and AI risks. AI risk management should be integrated into the enterprise risk management framework rather than managed through a completely separate policy. This approach ensures consistency in risk treatment, avoids duplication of controls, and aligns AI-related risk decisions with the organization's overall risk appetite and governance structure.

The enterprise risk management policy should clearly define the scope of risk management, risk management objectives, and the organization's risk appetite and tolerance levels. It should describe the risk assessment methodology to be used, including risk identification, analysis, evaluation, and treatment processes. The policy should also define risk rating criteria, such as impact and likelihood scales, and specify how risks are classified, prioritized, and escalated.

Senior management is responsible for approving this policy and defining the organization's risk appetite, including AI-related risks. Risk management, legal, and compliance functions support oversight, guidance, and monitoring of AI risks across the organization.

> **Note**
>
> For your ready reference, a template of a structured risk management policy is available as *Template 9* in this book's GitHub repository.

A policy establishes how risks should be managed, but organizations also need a practical way to record and track individual risks. This is where the risk register becomes important.

Risk register

A **risk register** is a central document used to record, track, and manage risks identified within an organization or a specific system. In the context of AI, a risk register captures all identified AI-related risks, including technical, ethical, legal, security, operational, and reputational risks. It provides a structured way to document each risk, assess its impact and likelihood, assign ownership, define mitigation measures, and monitor residual risk over time.

Ideally, organizations should maintain a common, enterprise-wide risk register that consolidates all categories of risks, including strategic, operational, cybersecurity, and AI-related risks. AI risks should be integrated into this enterprise risk register rather than maintained in isolated or standalone documents. This approach ensures consistency in risk evaluation, avoids duplication, and enables senior management to view AI risks in the context of overall organizational risk exposure.

The risk register serves as a living document that supports continuous risk management throughout the AI lifecycle. As AI systems evolve due to data changes, model updates, or expanded use cases, new risks may emerge and existing risks may change in severity. The risk register ensures that these changes are systematically recorded, reviewed, and escalated when required.

An effective risk register typically includes details such as a unique risk identifier, risk description, source or cause of the risk, impact and likelihood ratings, overall risk level, selected risk treatment strategy, mitigation controls, risk owner, and review status. Maintaining this information in a single, structured register improves transparency, accountability, and traceability.

> **Note**
>
> For your ready reference, a template of a structured AI risk register is available as *Template 10* in this book's GitHub repository.

Summary

In this chapter, we learned that AI risk assessment helps organizations identify, analyze, evaluate, and manage risks throughout the AI lifecycle. AI risks can be technical, ethical, legal, security-related, operational, or reputational. We discussed how AI risks can be identified using techniques such as expert workshops, AI risk taxonomies and checklists, review of past AI incidents, stakeholder impact analysis, technical review, model documentation review, and

lifecycle-based risk identification. These techniques help identify risks from different perspectives and across different stages of the AI lifecycle.

We went through the four main steps of AI risk assessment: identification, analysis, evaluation, and treatment. Risks are assessed based on their likelihood and impact and then prioritized for appropriate action. We examined how risks can be treated by avoiding, reducing, transferring, or accepting them, depending on the level of risk and the organization's risk appetite. We also discussed that AI risks should be integrated into the organization's enterprise risk management framework and maintained in a common risk register. The risk register should be continuously updated as AI systems, data, and use cases change.

Having understood different AI governance frameworks and regulations, let us now understand how to align an organization's AI governance approach with multiple frameworks and choose the most appropriate framework for your organizational context in our next chapter.

11

Comparative Analysis of AI Frameworks – ISO/IEC 42001 and EU AI Act

In previous chapters, we examined key AI governance structures such as the **OECD AI Principles**, **ISO/IEC 42001**, the **NIST AI Risk Management Framework** (**NIST AI RMF**), and the **EU AI Act**. We will build on our understanding through a comparative analysis of these structures.

In this chapter, we will cover the following: an overview of the major AI governance frameworks we've discussed so far, guidance on selecting the right framework for your organization's specific needs, and why ISO standards offer the most advantage in this space.

Overview of major AI governance frameworks

In the previous chapters, we examined the major AI governance frameworks, standards, and principles in detail. Before comparing them, let us briefly revisit the key instruments discussed earlier:

- ISO/IEC 42001 is an international standard for an **Artificial Intelligence Management System** (**AIMS**). It provides a structured management-system approach covering organizational context, leadership, planning, support, operation, performance evaluation, and continual improvement, supported by the controls in Annex A. See *Chapter 5*.
- NIST AI RMF provides a flexible approach for managing AI risks and promoting trustworthy AI. It has four functions: **Govern**, **Map**, **Measure**, and **Manage**. See *Chapter 6*.

- OECD AI Principles provide high-level principles for promoting trustworthy and responsible AI. They emphasize human-centered values and fairness, transparency and explainability, robustness, security and safety, and accountability. See *Chapter 7*.
- EU AI Act is a major regulatory framework governing the development, placing on the market, and use of AI systems in the European Union. Unlike voluntary frameworks and standards, the EU AI Act establishes legally binding requirements and follows a risk-based approach, with different obligations depending on the classification of an AI system. See *Chapter 8*.

At a broad level, all four promote responsible and trustworthy AI. They recognize the importance of managing AI-related risks, considering impacts on individuals and society, establishing accountability, maintaining appropriate oversight, and ensuring that AI systems are developed and used responsibly.

However, their approaches are different. The OECD AI Principles provide high-level direction, NIST AI RMF provides a risk management methodology, ISO/IEC 42001 provides an organizational management-system structure, and the EU AI Act establishes legally binding requirements for AI systems within its scope. This distinction is important because these instruments should generally be viewed as complementary rather than competing approaches.

Selecting the right framework for your organization

After understanding the major AI governance instruments, an important question arises: *Which framework should an organization choose?* The appropriate choice depends on the organization's objectives, regulatory obligations, industry, geographic presence, existing governance structure, and level of AI maturity.

When to use the OECD AI Principles

The OECD AI Principles are particularly appropriate for organizations that are at an early stage of AI adoption and want to establish a responsible AI direction without immediately introducing a detailed governance framework.

Consider an organization that has recently started adopting AI systems. The organization is currently not subject to any specific AI regulation and does not want to begin with extensive policies, procedures, documentation requirements, or formal governance processes. Instead, it wants to gradually develop its AI governance practices based on its experience, lessons learned, and changing AI use cases.

For such an organization, adopting the OECD AI Principles can be an appropriate starting point. The organization can use the principles to establish broad expectations around areas such as fairness, transparency, accountability, safety, security, and human-centered values.

These principles can guide employees and AI system owners without requiring the organization to immediately establish a comprehensive management system.

As the organization's use of AI grows and its governance maturity increases, it can gradually introduce more structured AI risk management policies and procedures.

When to use the NIST AI RMF

The NIST AI RMF is particularly appropriate for organizations that are ready to move from broad responsible AI principles toward more detailed AI governance policies and procedures but want to retain flexibility in how those requirements are implemented.

Consider an organization that has been using AI for some time and has gained sufficient experience to understand the types of AI risks it faces. The organization now wants to formalize its approach by developing AI policies, procedures, risk assessment processes, roles and responsibilities, monitoring activities, and other governance practices. However, it does not want to adopt a highly prescriptive approach or implement a formal management system at this stage.

For such an organization, the NIST AI RMF can provide an appropriate foundation. It provides a structured approach through its four functions, i.e., Govern, Map, Measure, and Manage, while allowing the organization to determine how these activities should be implemented based on its own business context, AI use cases, risk appetite, and maturity.

When to use ISO/IEC 42001

ISO/IEC 42001 is particularly appropriate for organizations that want to establish a formal, structured, and organization-wide approach to managing AI. It can be a suitable choice when an organization:

- Wants to establish a formal AI management system rather than managing AI governance through individual policies and procedures
- Has significant or increasing reliance on AI systems across multiple business functions
- Wants clearly defined governance roles, responsibilities, accountability, and decision-making authority
- Needs consistent AI governance processes across different AI systems, business units, or locations
- Wants AI risk management to be integrated with existing governance, risk, compliance, information security, privacy, or quality management processes
- Requires stronger documentation, evidence, monitoring, internal audit, and continual improvement mechanisms

- Wants to demonstrate its AI governance practices to customers, regulators, business partners, or other stakeholders
- Is considering certification against an AI management system standard

ISO/IEC 42001 can provide a structured approach for AI governance. It can bring together AI policies, risk management, roles and responsibilities, operational processes, monitoring, performance evaluation, internal audit, corrective actions, and continual improvement within a single management-system structure. However, please note that to obtain ISO/IEC 42001 certification, an organization must meet the applicable requirements of the standard as a whole. It cannot select only specific requirements and still obtain certification.

EU AI act

The EU AI Act should be viewed differently from the other instruments because it is a regulation rather than a voluntary framework or standard.

The first question is therefore not whether an organization wants to adopt the EU AI Act, but whether the regulation applies to the organization, its AI system, or its activities. Please note that we are discussing the EU AI Act, as currently it is the only regulation in the world impacting the use of AI. In the near future, most countries will have their own regulations on the use of AI. As an AI GRC officer, you need to be aware of AI-related regulations impacting your geography, business activities, or AI use cases.

Determining applicable regulations is one of the most important roles of an AI GRC professional because AI compliance always starts with knowing which laws apply in the first place. AI GRC professionals must first map where the organization operates, where the AI system is developed, where it is deployed, and who is affected by it. Based on this, they must identify all relevant AI-related laws, sectoral regulations, data protection requirements, and supervisory expectations. That being said, most AI GRC professionals would prefer ISO certification. ISO has a clear advantage over NIST and OECD for practical AI governance and regulatory management.

The ISO advantage

We've discussed repeatedly that the selection of a framework or standard depends upon an organization's specific requirements, but on the ground, ISO has a clear advantage over NIST and OECD for practical AI governance and regulatory management. Let's discuss why.

- **ISO standards are designed as management systems**: This means AI requirements can be embedded into existing enterprise governance structures, such as risk management, information security, privacy, quality, and compliance. For organizations already aligned with ISO-based systems, AI governance can be integrated without creating parallel or duplicate processes.

- **ISO standards are implementation-oriented rather than purely principle- or guidance-based**: While OECD focuses on high-level principles and NIST provides flexible risk management guidance, ISO translates governance expectations into defined policies, roles, processes, controls, documentation, and continual improvement mechanisms. This makes ISO more suitable for day-to-day operationalization of legal and regulatory requirements.
- **ISO supports auditability and assurance**: ISO-based approaches are designed to generate evidence, maintain records, and demonstrate consistent application of controls. This is critical when responding to regulators, customers, or partners who expect proof of compliance rather than intent or maturity statements. An ISO/IEC 42001 certification demonstrates that an organization has implemented all the requirements under the standard. An organization cannot be certified on partial compliance.
- **ISO enables enterprise-wide consistency**: AI risks are rarely isolated to a single team or function. ISO management system structures ensure consistent governance across business units, geographies, and third parties, which is essential for organizations deploying AI at scale.
- **ISO aligns well with a lifecycle-based approach to AI governance**: From design and development to deployment, monitoring, change management, and retirement, ISO-style structures support continuous oversight rather than one-time risk assessments. As we've observed, ISO offers a practical advantage over other frameworks because it enables AI legal requirements to be systematically embedded into existing GRC structures, managed consistently across the enterprise, and demonstrated through auditable governance mechanisms.

Next, we'll map OECD and NIST AI RMF principles against ISO standards to understand their correlation better.

OECD principles mapped to ISO/IEC 42001 standard

When you implement an ISO/IEC standard, you ultimately take care of all the OECD AI Principles as well, as can be seen from the mapping table below. This is because OECD principles define what responsible and trustworthy AI should achieve at a high level, while ISO/IEC standards define how those principles are operationalized within an organization:

OECD AI Principle (OECD)	How it is addressed through the ISO/IEC 42001 standard
Human-centered values and fairness	ISO/IEC standards require defined policies, ethical considerations, risk assessments, and controls to identify and mitigate bias, discrimination, and adverse impact on individuals. Human oversight, accountability, and decision review mechanisms operationalize this principle. (clause/ Annex A reference: 6.1.2, 6.1.4, 8.4, A.5.2–A.5.5, A.9.2–A.9.4)
Transparency and explainability	ISO/IEC mandates documentation, traceability, role clarity, and record keeping across the AI lifecycle. These requirements support explainability, transparency of decision-making, and availability of evidence for internal and external review (clause / Annex A reference: 7.5, 8.1, A.6.2.3, A.6.2.7, A.6.2.8, A. 7.5, A.8.2, A.8.5).
Robustness, security, and safety	Risk management, secure design, change management, monitoring, and incident handling controls within ISO/IEC standards ensure that AI systems are reliable, resilient, secure, and safe throughout their lifecycle. (clause / Annex A reference: 6.1.2, 6.1.3, 8.1, A.6.2.4, A.6.2.6, A.6.2.8, A.8.4)
Accountability	ISO/IEC standards explicitly define roles, responsibilities, ownership, escalation mechanisms, and governance structures, ensuring clear accountability for AI decisions, outcomes, and compliance obligations (clause / Annex A reference: 5.1, 5.2, 5.3, 6.1, A.3.2, A.10.2).
Inclusive growth, sustainable development, and well-being	ISO/IEC governance frameworks require consideration of organizational context, stakeholder impact, legal and regulatory obligations, and continual improvement, supporting responsible deployment aligned with societal and sustainability objectives (clause / Annex A reference: 4.1, 4.2, 6.1.4, A.5.2–A.5.5, A.9.3, 10.1).

Table 11.1: Mapping OECD AI Principles with AIMS

This is why, in practice, organizations implementing ISO/IEC standards naturally align with and operationalize OECD AI Principles through structured, auditable, and repeatable processes.

NIST AI RMF mapped to ISO/IEC 42001 standard

Similarly, requirements of NIST AI RMF can be operationalized through the ISO/IEC 42001 standard. NIST AI RMF explains what activities should exist to manage AI risk in a trustworthy manner. ISO/IEC 42001 explains how to institutionalize those activities through policies, procedures, roles, records, audits, and continual improvement:

NIST AI RMF function (NIST)	Key focus of NIST AI RMF	How it is addressed in the ISO/IEC 42001 standard
Govern	Establishing policies, roles, accountability, culture, and oversight for AI risk management	ISO/IEC 42001 requires organizations to establish governance structures, AI policies, defined roles and responsibilities, risk management processes, and oversight mechanisms to support effective and responsible AI governance (clause / Annex A reference: 4.1, 5.1, 5.2, 5.3, 6.1, 9.3, A.3.2).
Map	Understanding the AI system, its context, intended purpose, stakeholders, and potential impacts	ISO/IEC 42001 requires organizations to identify AI use cases, interested parties, internal and external issues, and AI-related risks and opportunities to understand the AI system and its operating context (clause / Annex A reference: 4.1, 4.2, 6.1.2, 6.1.4).
Measure	Assessing and analyzing AI risks, impacts, performance, bias, and effectiveness of controls	ISO/IEC 42001 requires organizations to conduct AI risk assessments and impact assessments, establish controls and performance criteria, and monitor AI systems throughout their lifecycle (clause / Annex A reference: 6.1.2, 6.1.3, 6.1.4, 8.1, 8.4, A.5.2–A.5.5, A.6.2.4, A. 6.2.6).

NIST AI RMF function (NIST)	Key focus of NIST AI RMF	How it is addressed in the ISO/IEC 42001 standard
Manage	Risk treatment, controls implementation, monitoring, incident handling, continual improvement	ISO/IEC 42001 requires organizations to implement risk treatment and operational controls, monitor AI systems, conduct performance evaluations and audits, manage incidents, implement corrective actions, and continually improve the AI management system (clause / Annex A reference: 6.1.3, 8.1, 9.1, 9.2, 10.1, 10.2, A. 6.2.6, A.8.4).

Table 11.2: Mapping NIST AI RMF with AIMS

When ISO/IEC 42001 is implemented properly, the intent and objectives of NIST AI RMF are inherently covered without additional effort.

Summary

AI governance requires understanding the difference between mandatory regulations and voluntary frameworks. Regulations define what organizations must comply with, while frameworks and standards provide practical ways to meet those requirements.

In this chapter, we discussed that organizations should first understand their legal obligations and then use appropriate governance structures to implement, monitor, and demonstrate compliance. We used a comparative analysis to understand the regulations organizations must comply with and the frameworks and standards that provide practical ways to meet those requirements.

We briefly revisited the principles and frameworks discussed in the previous chapters. ISO/IEC 42001 provides a structured, auditable approach to AI governance, while OECD provides high-level principles, and NIST AI RMF provides risk management guidance. The first responsibility of an AI GRC professional is to identify applicable laws and regulatory requirements. They can then use the appropriate framework to implement and manage those requirements effectively. Next, we examined how to select the right frameworks based on organizational needs.

We also discussed why AI GRC professionals prefer ISO over other frameworks: ISO/IEC 42001 can help operationalize both OECD principles and NIST AI RMF activities through defined policies, roles, controls, documentation, and continual improvement.

In the next chapter, we will build on this foundation by explaining how roles, responsibilities, committees, escalation mechanisms, and decision-making authority should be defined for effective AI oversight.

12

AI Governance Structure

So far, we have seen how AI is becoming more influential in core processes such as customer decision-making, recruitment, fraud detection, and automated monitoring. As a result, organizations must move beyond high-level principles and establish a clear governance structure to control how AI systems are designed, used, monitored, and retired. Without a defined structure, AI policies and ethical guidelines remain ineffective, as there is no clear ownership or escalation mechanism.

This chapter explains how to design an AI governance structure that fits within an organization's existing governance and risk management framework. It covers roles, committees, decision-making processes, and oversight mechanisms in a practical and scalable manner.

Specifically, we will explore the objective of establishing an AI governance structure, the different governance models that organizations can adopt, and the key considerations involved in implementing AI governance. We will also examine common pitfalls organizations may encounter when implementing AI governance and discuss the roles and responsibilities of key stakeholders involved in governing AI systems.

Objectives of an AI governance structure

Organizations need to have a proper AI governance structure to address the common themes emphasized across the NIST AI RMF, OECD AI Principles, EU AI Act, and ISO/IEC 42001, including:

- **Ethical AI**: Ensure that AI systems are used in a responsible and controlled manner. An effective AI governance structure offers clear accountability for AI-related decisions and risks. Governance also ensures that decisions related to AI approval, risk

acceptance, and exception handling are made by the appropriate authority and are properly documented.

- **Risk management**: AI systems introduce unique risks such as bias, lack of transparency, unintended automation of decisions, and model drift. A governance structure ensures these risks are identified early, assessed consistently, and monitored throughout the life of the AI system. This allows organizations to prevent issues rather than react to incidents after harm has occurred.
- **Regulatory and legal compliance**: Emerging regulations and standards increasingly require organizations to demonstrate oversight, traceability, and accountability for AI systems. A defined governance structure enables organizations to meet these requirements by establishing formal roles, committees, approval processes, and reporting mechanisms.
- **Supports trust and sustainability**: When AI systems are governed transparently and consistently, stakeholders such as customers, regulators, employees, and partners gain confidence in how AI is used. This trust is essential for scaling AI responsibly and ensuring long-term business value.

Clear objectives provide direction, but effective governance also requires a structure that offers authority and accountability. Let us now look at how an organization can establish an AI governance structure.

Governance model

An effective AI governance model should clearly define who provides strategic direction, who makes AI-related decisions, who owns individual AI systems, and who provides independent oversight. The governance structure should build on the organization's existing governance and risk management arrangements rather than creating a completely separate structure.

A clear definition of roles and responsibilities is essential for effective AI governance. Without explicit ownership, AI risks remain unmanaged and accountability becomes unclear. Every role involved in the design, deployment, and oversight of AI systems must understand its responsibilities across the AI lifecycle. Let us understand some important roles for AI governance:

- **Board of directors and senior management**: The board and senior management provide strategic oversight of AI use within the organization. Their responsibilities include setting the direction for AI adoption, approving AI governance policies, defining risk appetite, and ensuring alignment with organizational values and regulatory expectations. High-risk and business-critical AI systems should be visible at this level.

- **AI governance or AI steering committee**: The AI governance or AI steering committee is responsible for translating the board's strategic direction into operational governance. Its responsibilities include approving AI governance policies, reviewing high-risk AI use cases, monitoring significant AI risks, and ensuring alignment with organizational objectives and risk appetite. This committee often includes senior representatives from business, IT, risk management, legal, compliance, and data functions.
  The organization does not necessarily need to create a new committee or department for AI governance. An existing risk, technology, compliance, or ethics committee may be given additional AI responsibilities. The important point is that authority, ownership, oversight, escalation, and independent assurance are clearly defined.
- **AI GRC Officer**: The AI GRC Officer is responsible for coordinating and operationalizing AI governance across the organization. This role acts as a central point of contact for AI-related governance, risk, and compliance activities. The AI GRC Officer supports the development and maintenance of AI governance policies, standards, and procedures, and ensures their consistent implementation.
  The AI GRC Officer facilitates AI risk assessments, maintains the AI system inventory and AI risk register, and supports governance committees with accurate reporting and analysis. This role also monitors regulatory developments, supports compliance activities, and coordinates with business, technical, legal, and risk functions. While the AI GRC Officer does not own AI systems, the role plays a key role in enabling effective oversight and accountability.

  The responsibilities of the AI GRC Officer may be assigned to an existing GRC, risk, compliance, information security, or similar role. In smaller organizations, particularly SMBs, one individual may perform multiple responsibilities, provided that appropriate segregation of duties, independence, and accountability are maintained.
- **AI system owner**: The AI system owner is accountable for a specific AI system or use case. This role is responsible for defining the business purpose, ensuring compliance with governance requirements, and overseeing the system throughout its lifecycle. The AI system owner acts as the primary point of accountability for AI-related risks, incidents, and regulatory issues.
- **Business and AI development teams**: Business users and development teams are responsible for designing, building, configuring, and operating AI systems in line with approved requirements. Their responsibilities include implementing controls related to data quality, model performance, bias mitigation, explainability, and security. They are also responsible for monitoring AI behavior and reporting deviations or issues.

- **IT and information security teams**: IT and information security teams are responsible for providing a secure and resilient technical environment for AI systems. Their responsibilities include access controls, infrastructure security, logging and monitoring, incident response, and integration of AI systems into existing security processes.
- **Risk management and compliance teams**: Risk management and compliance team provide second-line oversight for AI systems. They are responsible for identifying and assessing AI-related risks, reviewing high-risk use cases, monitoring compliance with internal policies and external regulations, and supporting risk treatment decisions. These functions guide the business function but do not own AI systems.
- **Legal and privacy function**: The legal and privacy function ensures that AI systems comply with applicable laws, regulations, and contractual obligations. This includes data protection, intellectual property considerations, consent management, and regulatory interpretation. They support governance decisions by assessing legal exposure related to AI use cases.
- **Internal audit**: This team provides independent assurance over the effectiveness of the AI governance structure. Its responsibilities include reviewing governance processes, role clarity, decision-making mechanisms, and lifecycle controls. Internal audit does not manage AI systems but evaluates whether governance is operating as intended and identifies areas for improvement.

Clear role definitions, supported by formal documentation and reporting lines, ensure that AI governance is operational, auditable, and sustainable. We can further prevent duplication and gaps in responsibility, and support effective coordination throughout the AI lifecycle by documenting a **RACI (Responsible, Accountable, Consulted, and Informed) matrix**.

RACI matrix for AI governance

A RACI matrix can be used to clearly define who is responsible for performing an AI governance activity, who is accountable for the outcome, who should be consulted, and who should be kept informed. This helps avoid gaps or duplication in responsibilities across the AI lifecycle.

RACI stands for:

- **R – Responsible:** Performs the activity
- **A – Accountable:** Owns the final outcome and decision
- **C – Consulted:** Provides expertise or advice
- **I – Informed:** Needs to be kept informed of the outcome

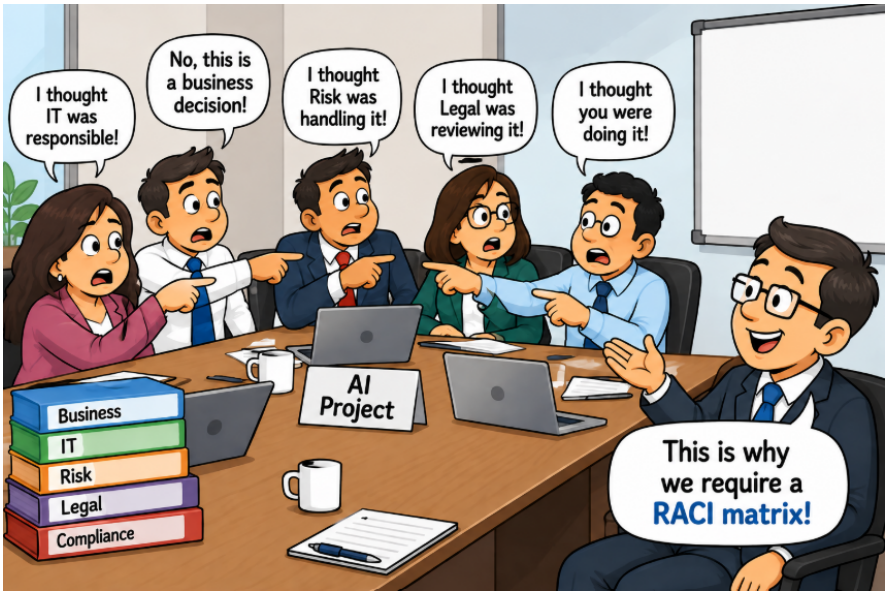


Figure 12.1: RACI matrix – role clarity

The following table demonstrates a sample AI governance RACI matrix:

Activity	Board	AI Committee	AI GRC Officer	AI System Owner	Development Team	Risk / Compliance	Legal / Privacy	Security	Internal Audit
Approve AI Governance Policy	A	R	C	I	I	C	C	C	I
Establish AI Governance Structure	A	R	R	C	I	C	C	C	I
Maintain AI Inventory	I	I	R	A	C	C	I	C	I
Define AI System Purpose	I	I	C	A	R	C	C	I	I
Conduct AI Risk Assessment	I	C	R	A	C	C	C	C	I
Conduct AI Impact Assessment	I	C	R	A	C	C	C	C	I
Regulatory/Legal Assessment	I	C	R	C	I	C	A/R	I	I
Security Assessment	I	I	C	A	C	C	I	R	I
Approve High-Risk AI System	A	R	C	C	C	C	C	C	I
Implement AI Controls	I	I	C	A	R	C	C	R	I
Monitor AI System	I	C	C	A	R	C	C	R	I
Report Significant AI Risks	I	A	R	C	C	C	C	C	I
Manage AI Incidents	I	C	R	A	R	C	C	R	I
Approve Risk Appetite and Risk Acceptance	A	R	C	C	C	C	C	C	I
Periodic Governance Review	A	R	R	C	C	C	C	C	C
Independent AI Governance Audit	I	I	C	C	I	C	C	C	A/R

Table 12.1: AI governance RACI matrix

For example, suppose a bank wants to deploy an AI-based loan approval system. The AI system owner would be accountable for the system, while the business and development teams would be responsible for implementing it. The AI GRC Officer would coordinate the risk and impact assessments, risk and compliance would provide oversight, legal and privacy would review regulatory and privacy considerations, and IT and security would assess technical risks. If the system is classified as high risk, the AI governance committee would review and approve its deployment.

In *Table 12.1*, for the **Regulatory/Legal Assessment**, the AI GRC Officer is marked **R (Responsible)**, meaning that the officer performs or coordinates the assessment process, while **Legal and Privacy** is marked **A/R (Accountable and Responsible)**, meaning that it provides the legal assessment and remains accountable for the final legal outcome. If there is a disagreement, such as between the business owner and Legal or Risk regarding whether the AI system can be deployed, the issue should be escalated through the defined governance and decision-making process, such as to the AI governance committee or senior management, depending on the organization's governance structure.

The RACI matrix should be customized to the organization's existing governance structure. Not every organization needs all the roles shown above, and some responsibilities may be combined. The key objective is to ensure that every important AI governance activity has clear ownership and that no critical responsibility is left undefined.

Implementing AI governance structures

In practice, AI governance structure answers four fundamental questions: who owns the AI system, who approves its use, who monitors its behavior, and who is accountable when something goes wrong. Without clear answers to these questions, AI initiatives tend to operate in silos, increasing the risk of regulatory non-compliance, ethical failures, and unmanaged operational risks.

The AI GRC Officer plays an important coordinating role in making the governance structure operational. The AI GRC Officer should take the following actions for implementing AI governance:

- **Define AI governance roles:** Work with senior management to document the responsibilities of the board, AI governance committee, AI system owners, business teams, technical teams, risk, legal, compliance, privacy, security, and internal audit
- **Establish approval processes:** Define how AI use cases are submitted, assessed, approved, rejected, or escalated before deployment. For example, high-risk AI systems may require approval from the AI governance committee.

- **Maintain an AI system inventory:** Ensure that all AI systems and use cases are identified and recorded, including the system owner, purpose, risk level, data used, and deployment status.
- **Establish clear accountability:** Ensure that every AI system has a named owner who is accountable for its business purpose, risks, compliance, and lifecycle management.
- **Integrate AI risk management:** Ensure that AI risks are incorporated into the organization's existing risk management processes rather than being managed separately.
- **Define escalation mechanisms:** Establish clear rules for when an AI-related issue must be escalated to management, the AI governance committee, legal, compliance, security, or other functions. For example, a serious AI incident affecting customers should have a defined escalation path.
- **Coordinate AI risk assessments:** Ensure that appropriate risk and impact assessments are performed before deployment and when significant changes are made to an AI system.
- **Establish monitoring and reporting:** Define what information should be monitored and reported to management, such as significant AI incidents, risk status, compliance issues, and changes in high-risk AI systems.
- **Coordinate stakeholder involvement:** Ensure that relevant functions are involved in AI decisions based on the nature and risk of the AI system.
- **Maintain governance documentation:** Ensure that policies, procedures, approvals, assessments, committee decisions, exceptions, and other governance records are properly documented.
- **Review governance effectiveness:** Periodically assess whether the governance structure is working effectively and identify areas requiring improvement.
- **Keep governance aligned with change:** Update roles, responsibilities, processes, and controls when the organization introduces new AI systems, changes existing systems, or faces new regulatory or business requirements.

The AI GRC Officer should act as the coordinator and facilitator of AI governance, not as the owner of every AI decision. Business and AI system owners remain accountable for their systems, while the AI GRC Officer ensures that the organization has the right governance processes, oversight, documentation, and escalation mechanisms in place.

All the above-mentioned points should be covered as well-defined processes in the AI governance policy.

> **Note**
>
> We have created a ready-to-use AI governance policy, *Template 1*, that captures the above-mentioned points in this book's GitHub repository. To access the repository link, follow the steps in the *Download the example code files* section in the *Preface*.

Let us now look at the common pitfalls that organizations should avoid when implementing AI governance.

Common pitfalls in implementing AI governance

Organizations may establish an AI governance structure but still face challenges if roles, processes, and oversight mechanisms are not implemented effectively. The following are some common pitfalls that an AI GRC Officer should watch for:

- **Creating a separate governance structure unnecessarily:** Establishing a completely new AI committee when an existing risk, compliance, technology, or ethics committee could oversee AI may create duplication and additional complexity.
- **Unclear ownership of AI systems:** An organization may establish an AI governance committee but fail to assign a specific owner to individual AI systems. Every AI system should have a clearly identified owner who is accountable for its lifecycle and risks.
- **Making the AI GRC Officer the system owner:** The AI GRC Officer should coordinate and facilitate governance activities but should not become the owner of every AI system. Accountability should remain with the relevant business or AI system owner.
- **Governance without decision-making authority:** A committee may exist on paper but have no clearly defined authority to approve, reject, restrict, or suspend AI use cases. The committee's authority should be formally documented.
- **Lack of clear escalation mechanisms:** Employees may identify an AI risk or incident but do not know whom to inform or when escalation is required. Clear escalation criteria and reporting channels should be established.
- **Treating governance as a one-time activity:** AI governance should continue after deployment because AI systems can change due to new data, model updates, or changes in the business environment.
- **Ignoring existing GRC processes:** Creating separate AI processes without considering existing risk management, compliance, security, and privacy processes can result in duplication and inconsistent governance. AI governance should be integrated with existing organizational arrangements.

- **Insufficient involvement of supporting functions:** AI decisions may be made only by business or technology teams without involving legal, privacy, compliance, risk, or information security when their expertise is required.
- **Weak documentation and evidence:** Organizations may have governance activities but fail to maintain evidence of approvals, risk assessments, committee decisions, exceptions, monitoring, and corrective actions. This can make governance difficult to demonstrate during audits or regulatory reviews.
- **No independent assurance:** Internal audit should not own or operate AI governance. Its role is to independently evaluate whether governance processes, responsibilities, decision-making mechanisms, and lifecycle controls are operating effectively.

The objective is not to create the largest governance structure. The objective is to create a structure where ownership, authority, oversight, escalation, and assurance are clear and actually work in practice.

Summary

In this chapter, we discussed how effective AI governance requires more than policies and committees. We explored the importance of establishing clear ownership, decision-making authority, oversight, and accountability, while integrating AI governance into existing organizational governance and GRC structures.

We also discussed how organizations can make governance practical by maintaining an AI inventory, assigning accountable system owners, establishing approval and escalation mechanisms, conducting risk and impact assessments, and monitoring and reporting AI-related risks. A RACI matrix can further clarify governance roles and responsibilities. Controls and documentation should be periodically reviewed and updated as AI systems, risks, business requirements, and regulatory expectations evolve.

Since a governance structure alone is not sufficient without clear policies and procedures, in the next chapter, we will discuss the key AI policy and procedure documents needed to operationalize governance, ensure consistent implementation, and support regulatory and audit requirements.

13

AI Governance Documents

As **artificial intelligence** (**AI**) becomes deeply embedded in business processes, decisions, and customer interactions, organizations can no longer rely on informal guidelines or ad hoc controls to manage AI. Effective AI governance requires clear, documented, and enforceable policies and procedures that define how AI is designed, used, monitored, and controlled across the organization.

This chapter focuses on translating high-level AI principles and governance frameworks into practical policy and procedural documents. While ethics statements and strategy documents explain why AI should be used responsibly, policies and procedures explain what is allowed, who is responsible, and how controls are implemented in day-to-day operations. Without these documents, AI governance remains theoretical and difficult to enforce.

In this chapter, we will discuss the importance of documented policies and procedures in establishing and maintaining effective AI governance. We will examine the document hierarchy and the important AI governance documents that organizations should establish. The chapter will also explain the key contents of an **AI Governance Policy** and outline documentation requirements under different regulatory requirements, standards, and AI governance frameworks.

Importance of documented AI policies and procedures

As AI becomes increasingly integrated into business processes and decision-making, organizations need clear and documented policies and procedures to ensure AI is used responsibly and consistently. Policies establish management expectations and boundaries, while procedures translate those requirements into practical steps. The following image illustrates the importance of clear policies and procedures for governing AI-related activities:

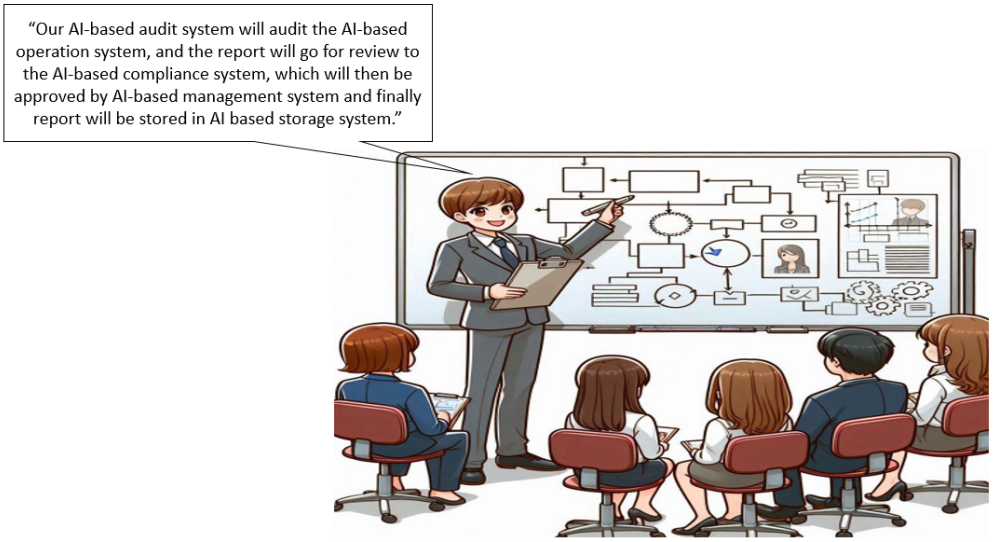


Figure 13.1: Integrated AI

Key reasons for documenting AI policies and procedures include:

- **Define clear expectations**: Establish what AI use is permitted, restricted, or prohibited within the organization
- **Clarify roles and accountability**: Define who is responsible for developing, approving, using, monitoring, and overseeing AI systems

- **Standardize AI governance**: Ensure that different business and technical teams follow a consistent approach when managing AI systems
- **Support risk management**: Establish requirements for identifying, assessing, treating, and monitoring AI-related risks throughout the AI lifecycle
- **Ensure regulatory compliance**: Help organizations address applicable AI regulations, privacy requirements, contractual obligations, and internal policies
- **Enable consistent decision-making**: Define processes for AI use-case approval, risk acceptance, exceptions, escalation, and deployment
- **Provide audit evidence**: Maintain documented approvals, assessments, monitoring records, decisions, and other evidence needed to demonstrate effective governance
- **Improve accountability and transparency**: Ensure that AI-related decisions and actions can be traced to responsible individuals or functions
- **Support continuous improvement**: Policies and procedures can be reviewed and updated as AI technology, business use cases, risks, and regulatory requirements change

Understanding the importance of policies and procedures raises an important question: *how should these documents be organized?*

Document hierarchy

Effective AI governance requires more than a single policy. Organizations should establish a clear hierarchy of documents that translates management expectations into practical activities and provides evidence that governance requirements are being followed.

A simple and practical AI governance hierarchy is:

Policy → Procedures → Guidelines → Templates and Checklists → Records and Evidence.

See *Figure 13.2*:



Figure 13.2: AI documentations

Let us now discuss these documents in detail:

- **Policy**: The policy establishes management's intent and the organization's overall requirements for AI governance. It defines the scope, objectives, principles, roles, responsibilities, and key governance requirements. It answers the question: *What is required and why?*
- **Procedures**: Procedures translate policy requirements into practical and repeatable steps. They define how activities such as AI use-case approval, risk assessment,

monitoring, incident management, and escalation should be performed. They answer the question: *How should the requirement be implemented?*

- **Guidelines**: Guidelines provide practical recommendations and good practices for applying policies and procedures. They allow flexibility where a single approach may not be appropriate for every AI system or business situation.
- **Templates and Checklists**: Templates and checklists provide standardized tools to support implementation. Examples include AI use-case assessment forms, AI impact assessment templates, risk assessment forms, vendor assessment checklists, and monitoring checklists. They help ensure that important governance activities are performed consistently.
- **Records and Evidence**: Records and evidence demonstrate that governance requirements have actually been implemented. Examples include approved use cases, completed assessments, approval records, monitoring reports, incident records, committee minutes, and review reports. These records support accountability, audits, and regulatory reviews.

This hierarchy helps translate AI governance from management intent into operational practice and demonstrable evidence. It also provides a consistent structure for integrating AI governance with the organization's existing governance, risk, and compliance processes.

A policy provides the overall direction, but effective AI governance requires several key supporting documents as we will examine next.

Important AI governance documents

An effective AI governance program is supported by a structured set of documents that translate governance requirements into practical actions and demonstrable evidence. Each document serves a specific purpose and helps ensure that AI systems are properly assessed, approved, monitored, and controlled throughout their lifecycle. Let us now understand different documents for AI GRC.

AI governance policy

The **AI governance policy** establishes management's overall expectations for the responsible and controlled use of AI. It defines the scope of AI governance, roles and responsibilities, risk management requirements, approval processes, human oversight, security, compliance, incident management, and monitoring requirements. The policy should apply to internally developed AI systems as well as third-party, vendor-provided, and generative AI tools.

Use case: A bank's AI governance policy may require that every high-risk AI system undergo an AI risk assessment and AI impact assessment before deployment and receive approval from the AI governance committee.

AI use case assessment and approval form

Before an AI system is developed or purchased, the organization should document the proposed use case. The assessment should capture the business purpose, intended users, data involved, expected benefits, potential impact, and initial risk classification. This document provides the starting point for determining whether an AI use case should proceed and what level of governance is required.

Use case: A business department proposes using AI to screen loan applications. The form records the purpose, types of customer data used, proposed decision-making role of AI, expected benefits, and whether the system could affect customers' access to credit.

AI system inventory

An **AI system inventory** provides a central record of AI systems and use cases across the organization. It should identify information such as the system name, business purpose, system owner, provider, data used, risk classification, deployment status, and lifecycle status. The inventory should include third-party and embedded AI systems, not only internally developed models.

Use case: An organization may maintain an inventory containing a loan approval AI classified as high risk, a customer service chatbot classified as limited risk, and an employee productivity assistant classified as low risk.

AI impact assessment

An **AI impact assessment** evaluates how an AI system may affect individuals, groups, society, and other stakeholders. It considers areas such as fairness, privacy, human rights, safety, legal compliance, organizational impact, and stakeholder impact. The assessment should be proportionate to the risk and context of the AI system.

Use case: For an AI recruitment system, the assessment may examine whether the system could discriminate against certain groups, whether applicants can challenge decisions, whether personal data is appropriately used, and whether human review is available.

Note

For your ready reference, we have created a template for AI impact assessment. The same is available as *Template 9* in the GitHub repository. To access the repository link, follow the steps in the *Download the example code files* section in the *Preface*.

AI risk assessment

The **AI risk assessment** identifies, analyzes, evaluates, and treats risks associated with an AI system. It considers risks such as technical failure, bias, legal non-compliance, security threats, operational problems, and reputational damage.

Use case: For a loan approval AI, a risk may be that the model produces biased loan approval recommendations because historical lending data contains discriminatory patterns. The organization can assess the likelihood and impact, determine the risk level, and define mitigation measures such as bias testing, human review, and model monitoring.

AI risk register

The **AI risk register** is used to record and track identified AI risks. It normally includes the risk description, cause, likelihood, impact, risk rating, risk owner, treatment strategy, controls, and residual risk. It should be treated as a living document because AI risks may change as data, models, and use cases evolve.

Use case: A risk register may record bias in loan decisions as a high risk with fairness testing and human review as mitigation measures, with the Credit Head assigned as the risk owner.

> **Note**
>
> For your ready reference, we have included a template of a risk register based on Annex A controls of the ISO 42001 standard. It is available as *Template 11* in the GitHub repository.

AI vendor and third-party assessment

Organizations should assess the risks associated with AI systems, models, data, or services obtained from third parties. The assessment should consider the vendor's security, privacy, transparency, compliance capabilities, contractual obligations, and ability to support ongoing monitoring. The AI governance policy should also establish requirements such as audit rights, data protection obligations, and performance expectations for third-party AI systems.

AI risks vary depending on the type of AI system:

- In-house AI may have more development and model-related risks
- SaaS AI, such as ChatGPT, may have more vendor, privacy, and data security risks
- Agentic AI may have additional risks because it can independently make decisions and perform actions

Therefore, the AI risk assessment should consider the type and level of AI use.

Use case: Before adopting a third-party generative AI service, the organization may assess whether customer data is used to train the vendor's models, where the data is stored, what security controls are implemented, and whether the contract provides appropriate audit and data protection rights.

AI model and system documentation

AI model and system documentation records important information about how an AI system is designed, developed, tested, deployed, and maintained. It should include the system's purpose, data sources, model characteristics, limitations, testing results, validation, and significant changes. Such documentation supports transparency, explainability, accountability, and regulatory or audit requirements.

Use case: Documentation for a fraud detection model may include its purpose, data sources, features used, model version, accuracy results, known limitations, testing performed, and approval date.

AI monitoring and review records

AI governance does not end when an AI system is approved and deployed. Organizations should maintain evidence of ongoing monitoring of performance, bias, security, model drift, incidents, and compliance. The AI governance policy should require periodic monitoring and review of AI systems.

Use case: A monthly monitoring report for a credit model may include accuracy, error rates, fairness indicators, significant changes in data, model drift, customer complaints, and incidents.

AI incident management procedure and records

An **AI incident management procedure** defines how AI-related incidents are identified, reported, investigated, escalated, and resolved. AI incidents may include incorrect or harmful outputs, data misuse, security attacks, discriminatory outcomes, or unexpected system behavior. The procedure should also require root cause analysis and corrective action to prevent recurrence.

Use case: If a chatbot provides incorrect financial advice to customers, the incident should be recorded, investigated, escalated according to its severity, and followed by corrective actions such as model changes, additional testing, or stronger human oversight.

AI change management records

AI systems can change because of model retraining, new data, software updates, changes in algorithms, or changes in their intended use. Significant changes should therefore be assessed and approved before implementation.

Use case: A loan approval model originally used only for internal recommendations is changed so that its output automatically determines loan approval. This significant change should trigger a new risk and impact assessment and appropriate approval.

AI governance committee records

The **AI governance committee** should maintain records of important governance decisions. These may include meeting minutes, use-case approvals or rejections, risk acceptance decisions, exceptions, significant incidents, and escalations.

Use case: The committee may review a high-risk recruitment AI and record its decision to approve deployment subject to bias testing, human oversight, quarterly monitoring, and periodic reassessment.

Training and awareness records

Organizations should maintain evidence that employees and relevant stakeholders understand their AI governance responsibilities. Training should be appropriate to the person's role and may cover responsible AI, acceptable AI use, risks, prohibited uses, privacy, security, and human oversight.

Use case: Employees may receive annual AI awareness training, while AI developers receive additional training on model risk, bias testing, security, and documentation requirements.

AI governance review and audit records

Organizations should periodically evaluate whether the AI governance framework is working effectively. Reviews may include internal audits, compliance assessments, management reviews, KPI/KRI monitoring, and corrective-action tracking.

Use case: Internal audit may review whether high-risk AI systems have completed required assessments, received appropriate approvals, maintained monitoring records, and followed the approved governance process.

Among these documents, the AI governance policy provides the primary direction for AI governance. This brings us to the question: *What should an AI governance policy contain?*

Key contents of AI governance policy

An AI governance policy serves as the primary governance document. The key contents of an AI governance policy should include the elements in *Figure 13.3*:

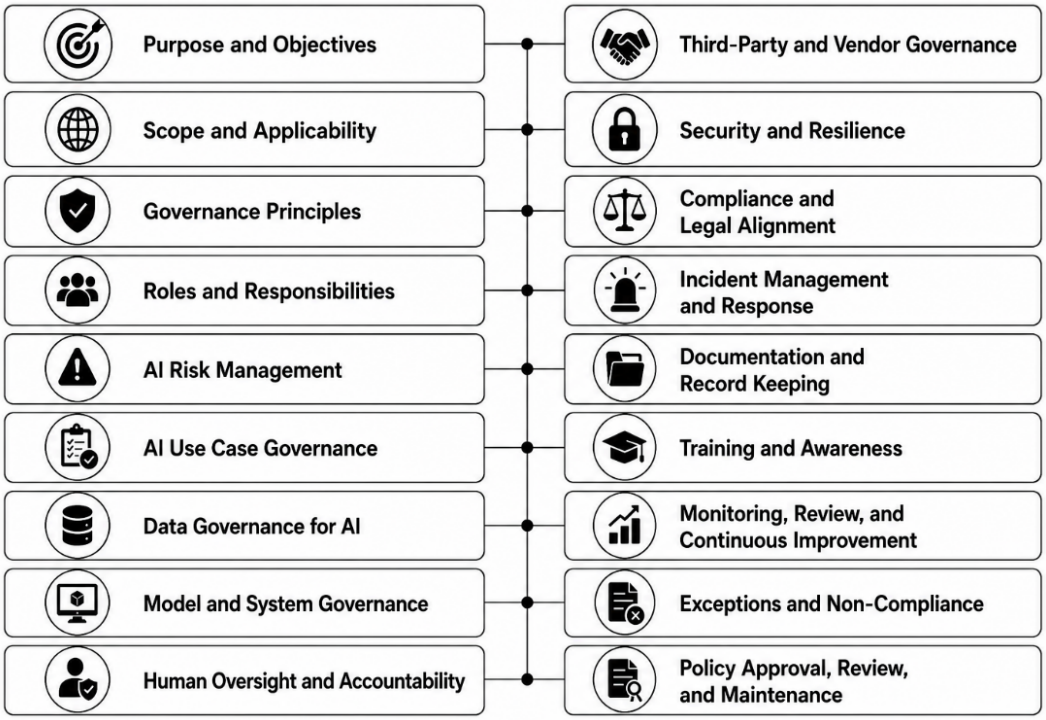


Figure 13.3: Contents of AI governance policy

Let's go through these elements in detail:

- **Purpose and Objectives**: The AI governance policy should clearly state the purpose of establishing governance over artificial intelligence within the organization. Its primary objective is to ensure that AI systems are used responsibly, safely, and in alignment with business goals. The policy should highlight objectives such as managing AI-related risks, ensuring compliance with applicable laws and regulations, protecting stakeholders, and building trust in AI-enabled decisions. It should also confirm alignment with the organization's values, ethical commitments, and overall business strategy.
- **Scope and Applicability**: The policy should define the scope of AI governance by specifying which AI systems are covered, including internally developed AI, externally

procured solutions, third-party AI services, and generative AI tools. It should clearly state that the policy applies to all relevant business units, employees, contractors, and vendors involved in AI-related activities. Coverage should extend across the full AI lifecycle, from design and development to deployment, monitoring, and eventual retirement of AI systems.

- **Governance Principles**: The AI governance policy should establish core principles that guide all AI-related decisions. These typically include responsible and ethical use of AI, transparency and explainability of AI outputs, fairness and non-discrimination, and clear accountability supported by appropriate human oversight. The policy should also emphasize security, safety, and resilience to ensure AI systems operate reliably and do not cause unintended harm.
- **Roles and Responsibilities**: Clear definition of roles and responsibilities is essential for effective AI governance. The policy should outline oversight responsibilities of the board and senior management, as well as the role of an AI governance committee or equivalent body. It should assign ownership to AI system owners and business sponsors and define the supporting roles of risk management, compliance, legal, and IT or information security teams. Responsibilities of AI developers and users should also be clearly documented to ensure accountability at all levels.
- **AI Risk Management**: The policy should require a structured approach to identifying, classifying, and managing AI-related risks. This includes risks related to ethics, data quality, bias, security, legal compliance, and operational reliability. It should define requirements for risk assessment, approval of AI systems based on risk level, and processes for risk treatment, acceptance, and escalation. AI risk management should be integrated into the organization's enterprise risk management framework rather than treated as a standalone activity.
- **AI Use Case Governance**: The policy should establish a formal process for proposing, reviewing, and approving AI use cases. This includes defining clear approval and rejection criteria based on risk, impact, and compliance considerations. AI use cases should be classified into categories such as low, medium, or high risk, with corresponding governance controls. Decisions related to use case approval should be properly documented to ensure traceability and accountability.
- **Data Governance for AI**: Data is a critical component of AI systems, and the policy should define clear data governance requirements. This includes controls over data sourcing, data quality, and data suitability for AI use. The policy should address bias, representativeness, and data minimization to reduce unfair or discriminatory outcomes. It should also specify requirements for data protection, retention, deletion, and compliance with applicable privacy and data protection laws.

- **Model and System Governance**: The policy should define governance controls for AI models and systems throughout their lifecycle. This includes requirements for model development, testing, validation, and documentation. Explainability expectations should be clearly stated, especially for high-impact AI systems. The policy should also require ongoing monitoring of performance, bias, and model drift, along with controlled processes for change management, retraining, decommissioning, and retirement.
- **Human Oversight and Accountability**: Human oversight should be a core requirement of the AI governance policy. The policy should specify when human-in-the-loop or human-on-the-loop controls are required and how decisions supported by AI can be reviewed or overridden. It should clearly state that accountability for AI-assisted decisions remains with designated individuals or roles, particularly for critical or high-impact outcomes. Escalation mechanisms should be defined for cases where AI outcomes raise significant concerns.
- **Third-Party and Vendor Governance**: The policy should address governance of AI systems provided by third parties and vendors. This includes due diligence requirements to assess vendor risk, transparency, and compliance capabilities. Contractual requirements such as audit rights, data protection obligations, and performance expectations should be clearly defined. The policy should also require ongoing monitoring of third-party AI systems to ensure continued compliance with organizational governance requirements.
- **Security and Resilience**: AI systems should be designed and operated in a secure and resilient manner. The policy should require secure design and deployment practices, including access controls, logging, and continuous monitoring. It should address protection against misuse, unauthorized access, and adversarial threats specific to AI systems. Business continuity and recovery considerations should be included to ensure AI systems can be restored or safely disabled in the event of failures or incidents.
- **Compliance and Legal Alignment**: The AI governance policy should require identification and monitoring of applicable AI-related laws, regulations, and contractual obligations. It should ensure alignment with internal policies such as information security, privacy, and risk management. The policy should also define responsibilities for regulatory reporting, disclosures, and maintaining evidence to demonstrate compliance and audit readiness.
- **Incident Management and Response**: The policy should define what constitutes an AI-related incident, including failures, unintended outcomes, data misuse, or ethical breaches. It should establish procedures for incident detection, reporting, investigation, and response. Root cause analysis and corrective actions should be

required to prevent recurrence. Where applicable, the policy should address communication with regulators, customers, and other stakeholders.

- **Documentation and Record Keeping**: Effective AI governance depends on proper documentation. The policy should specify mandatory documentation artifacts such as risk assessments, approval records, model documentation, and decision logs. It should require audit trails covering the AI lifecycle and define retention and secure storage requirements. Proper documentation supports transparency, accountability, and regulatory compliance.
- **Training and Awareness**: The policy should require AI awareness training for general employees and role-based training for developers, reviewers, and decision-makers. Periodic refresher training and competency assessments should be conducted to ensure continued understanding of AI risks and responsibilities. Employees should also be made aware of prohibited and high-risk AI uses to prevent misuse.
- **Monitoring, Review, and Continuous Improvement**: The AI governance policy requires periodic review of governance effectiveness. This includes internal audits, assessments, and monitoring of key performance indicators and key risk indicators related to AI governance. The policy should be updated based on lessons learned, incident outcomes, technological changes, and evolving regulatory expectations.
- **Exceptions and Non-Compliance**: The policy should define a formal process for requesting exceptions to AI governance requirements. Any approved deviations should be documented, justified, and time bound. The policy should also describe how non-compliance is handled, including corrective actions, disciplinary measures, and reporting to senior management where required.
- **Policy Approval, Review, and Maintenance**: Finally, the AI governance policy should clearly define policy ownership, approval authority, and review frequency. Version control and change history should be maintained to track updates over time. This ensures the policy remains current, effective, and aligned with organizational and regulatory expectations.

In practice, many organizations adopt a single enterprise-level AI governance policy supported by detailed procedures, registers, and templates. This approach reduces duplication while ensuring clarity and accountability.

Note

For your reference, ready-to-use AI governance policy and AIMS policy templates are provided as *Template 1* and *Template 4*, respectively, in this book's GitHub repository. To access the repository link, follow the steps in the *Download the example code files* section in the *Preface*.

> These templates are designed for practical implementation and can be adapted to suit different organizational contexts.
>
> The provided template is fully aligned with the requirements of the ISO/IEC 42001 standard and supports organizations in establishing a structured and compliant AI governance framework.

Next, we will examine the documentation and record-keeping expectations under major AI regulations, standards, and frameworks.

Documentation and record keeping requirements under different regulations and frameworks

As a GRC professional, it is essential to understand the basic documentation and record-keeping expectations prescribed by different AI regulations, standards, and governance frameworks. While the depth and formality of documentation may vary, most regulations and frameworks emphasize the need for traceability, accountability, transparency, and evidence of compliance across the AI lifecycle. We have discussed AI regulations and frameworks at length in the previous chapters; we will now discuss their key documentation requirements.

EU AI Act

The EU AI Act places strong and explicit documentation obligations, especially for high-risk AI systems. Organizations are required to maintain technical documentation that demonstrates compliance throughout the AI lifecycle. This includes documentation of the intended purpose of the AI system, system design, algorithms used, training and testing data characteristics, risk management measures, and human oversight arrangements.

In addition, providers must maintain logs and records to enable traceability of system behavior and decisions after deployment. Post-market monitoring records, incident reports, and corrective actions must also be documented. Documentation must be sufficiently detailed to allow regulators and notified bodies to assess compliance and must be retained for defined periods even after the AI system is withdrawn.

ISO/IEC 42001 artificial intelligence management systems

ISO/IEC 42001 takes a management-system approach to documentation and record-keeping. It requires organizations to document AI policies, objectives, roles and responsibilities, risk management processes, lifecycle controls, and monitoring activities.

Records are required to demonstrate that the AI management system is operating effectively. This includes records of risk assessments, impact assessments, training and awareness

activities, performance monitoring, incidents, corrective actions, and management reviews. Documentation under ISO/IEC 42001 is less prescriptive about technical detail compared to the EU AI Act, but it strongly emphasizes consistency, governance maturity, and continual improvement.

NIST AI Risk Management Framework (AI RMF)

The **NIST AI RMF** emphasizes documentation as a key enabler of trustworthy AI. It expects organizations to document AI system context, intended use, assumptions, limitations, risks, and mitigation measures. Documentation should support transparency, explainability, and accountability across the AI lifecycle.

Records related to risk identification, risk prioritization, testing, validation, monitoring, and incident response are critical. While NIST AI RMF is not a regulatory requirement, documentation is essential to demonstrate alignment with its core functions and to support internal assurance, audits, and stakeholder trust.

OECD AI Principles

The **OECD AI Principles** do not mandate specific documents but strongly emphasize documentation to support transparency, accountability, and responsible AI use. Organizations are expected to maintain records that explain how AI systems were designed, how risks such as bias and harm were addressed, and how human oversight is ensured.

Documentation should enable affected stakeholders and oversight bodies to understand AI decision-making processes at an appropriate level. From a GRC perspective, OECD principles act as a foundation that influences more detailed regulatory and standard-based documentation requirements.

For a GRC professional, the key is to recognize common documentation themes across regulations and frameworks and implement a harmonized documentation structure that satisfies multiple requirements simultaneously. This approach reduces duplication, improves consistency, and ensures that AI documentation and record-keeping remain practical, scalable, and audit-ready across jurisdictions and governance expectations. Let us now understand the stepwise procedure for creating AI documents.

Stepwise procedure for creating AI documents

Developing AI governance documents is only the first step. The organization must also establish a practical process to approve, implement, communicate, and maintain them. The

following roadmap can help organizations move from documented requirements to operational AI governance:

1. **Identify documentation requirements**: Start by identifying the AI governance documents required by the organization. This should include the AI governance policy, procedures, guidelines, templates, checklists, registers, and required records.
   The documentation should be based on the organization's AI use cases, risk profile, regulatory requirements, and existing GRC arrangements. Avoid creating separate documents where existing enterprise policies and procedures can be extended to cover AI.

2. **Develop the AI governance policy**: Prepare the AI governance policy as the primary governance document. It should define management expectations, scope, governance principles, roles and responsibilities, risk management, approval requirements, human oversight, security, compliance, monitoring, and other key requirements.
   The policy should be written at a level that provides clear direction without including detailed operational instructions that belong in procedures. The policy should also identify the supporting procedures, templates, and records required for implementation.

3. **Develop supporting procedures and guidelines**: Translate the policy requirements into practical procedures and guidelines. Procedures should explain how activities are performed, who performs them, what approvals are required, and what records must be maintained.
   For example, an AI Use Case Approval Procedure may define the steps for submitting a proposed AI use case, performing an initial assessment, determining its risk level, obtaining required reviews, and approving or rejecting the use case.

4. **Develop templates and checklists**: Create standardized templates and checklists to make the procedures easier to implement consistently. Templates should be simple enough for business and technical teams to use while capturing the information required for governance and audit purposes.

5. **Define ownership and approval authority**: Assign an owner to each governance document and clearly define who has authority to approve it. The AI governance policy may require approval by the board or senior management, while supporting procedures may be approved by the responsible function or AI governance committee, depending on the organization's governance structure.
   Document ownership should also cover responsibility for periodic review and updates.

6. **Review and approve the documents**: Before implementation, the documents should be reviewed by relevant functions such as business, AI/IT, risk, compliance, legal, privacy, and information security.
   The review should confirm that the documents are practical, consistent with existing GRC processes, and aligned with applicable regulatory requirements. Following the required reviews, the appropriate authority should formally approve the documents.
7. **Communicate and implement**: Approved documents should be communicated to employees and other relevant stakeholders. Employees should understand what the policy requires and where to find the applicable procedures, guidelines, and templates. Role-based training should be provided where necessary. For example, AI developers may require detailed guidance on model documentation and testing, while business users may primarily need guidance on acceptable AI use and approval requirements.
8. **Monitor and review**: AI governance documents should be periodically reviewed to ensure that they remain relevant and effective. Reviews should consider changes in AI technology, new AI use cases, incidents, audit findings, business requirements, and regulatory developments.
   Where weaknesses are identified, the organization should update the relevant policy, procedure, guideline, template, or control.
9. **Maintain version control**: All governance documents should have appropriate version control, including document owner, approval date, effective date, review date, version number, and change history. This ensures that employees use the current approved version and provides a clear record of how governance requirements have evolved.

The overall implementation can therefore be summarized as:

Identify → Develop → Approve → Communicate → Monitor → Update

The objective is not simply to create a collection of AI governance documents. The objective is to establish a working governance system in which policies define expectations, procedures guide implementation, templates support execution, and records provide evidence that governance is operating effectively.

Summary

In this chapter, we discussed how AI governance is effective only when organizational expectations are translated into consistent actions and supported by evidence. A policy can provide direction, but procedures, guidelines, templates, assessments, registers, and records turn that direction into repeatable governance practices. The key lesson in this chapter is that governance must be embedded throughout the AI lifecycle, with clear accountability for

assessing risks, approving use cases, protecting data, overseeing models, managing third parties, and responding to incidents.

We discussed that AI governance arrangements should not remain static. Organizations need to communicate requirements, monitor whether they are being followed, retain evidence of key decisions and controls, and periodically update their governance framework as AI technologies, business needs, risks, and regulatory expectations evolve. In this way, AI governance becomes an ongoing management discipline rather than a one-time documentation exercise.

In the next chapter, we will go through the AI lifecycle governance, explaining how governance controls are implemented from design and development to deployment, monitoring, change management, and retirement of AI systems.

14

AI Lifecycle Governance

AI systems are not one-time projects. They may be developed, purchased, configured, integrated, or consumed as third-party services. They may also be changed, monitored, and eventually retired. Each of these stages introduces different risks, responsibilities, and governance requirements. If governance is applied only at the design stage or only after deployment, important risks can easily be missed.

AI lifecycle governance focuses on applying governance, risk management, and compliance controls across the entire lifecycle of an AI system—from idea and data collection to development, deployment, monitoring, and decommissioning. It ensures that ethical principles, regulatory requirements, security controls, and accountability mechanisms are embedded at every stage, rather than treated as an afterthought.

The following image highlights why governance should extend across the entire AI lifecycle rather than being treated as a one-time activity:

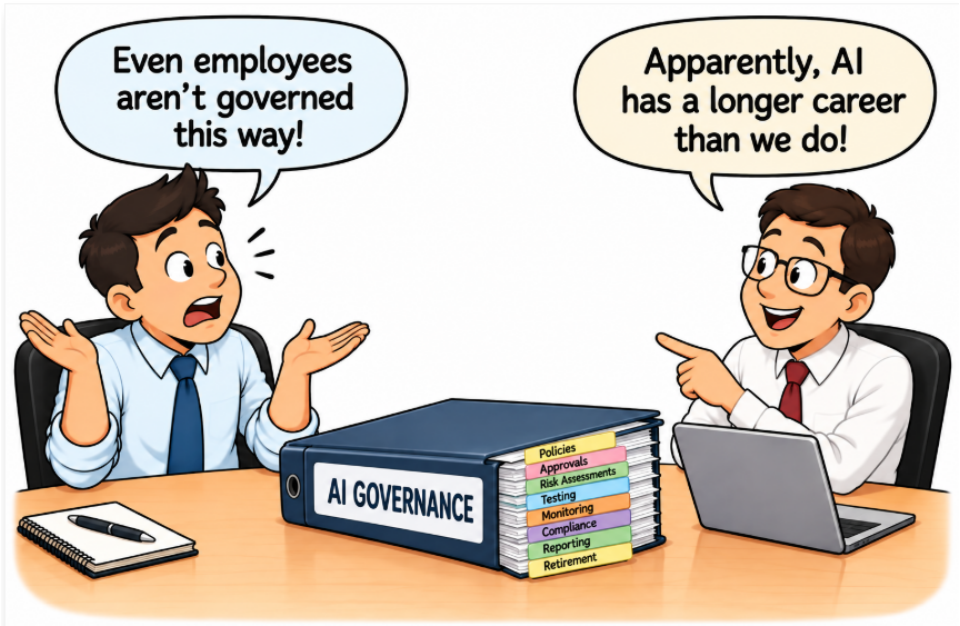


Figure 14.1: Governance across the entire AI lifecycle

This chapter explains how organizations can define clear governance checkpoints, roles, documentation, and controls for each phase of the AI lifecycle. The objective is to help organizations manage AI risks proactively, maintain trust, and demonstrate compliance, while still enabling innovation and business value.

AI lifecycle stages — a governance perspective

To govern AI effectively, an organization must first understand the end-to-end lifecycle of an AI system. AI governance is not limited to model development; it starts much earlier and continues even after the system is taken out of use. Each stage of the lifecycle introduces specific risks and therefore requires defined governance controls.

From an AI GRC perspective, the lifecycle can be broadly divided into seven stages. They are an illustrative governance model and are not a universal lifecycle structure. Different organizations and frameworks may use different lifecycle representations. The stages can therefore be adapted based on the organization's AI systems, risks, and governance requirements:

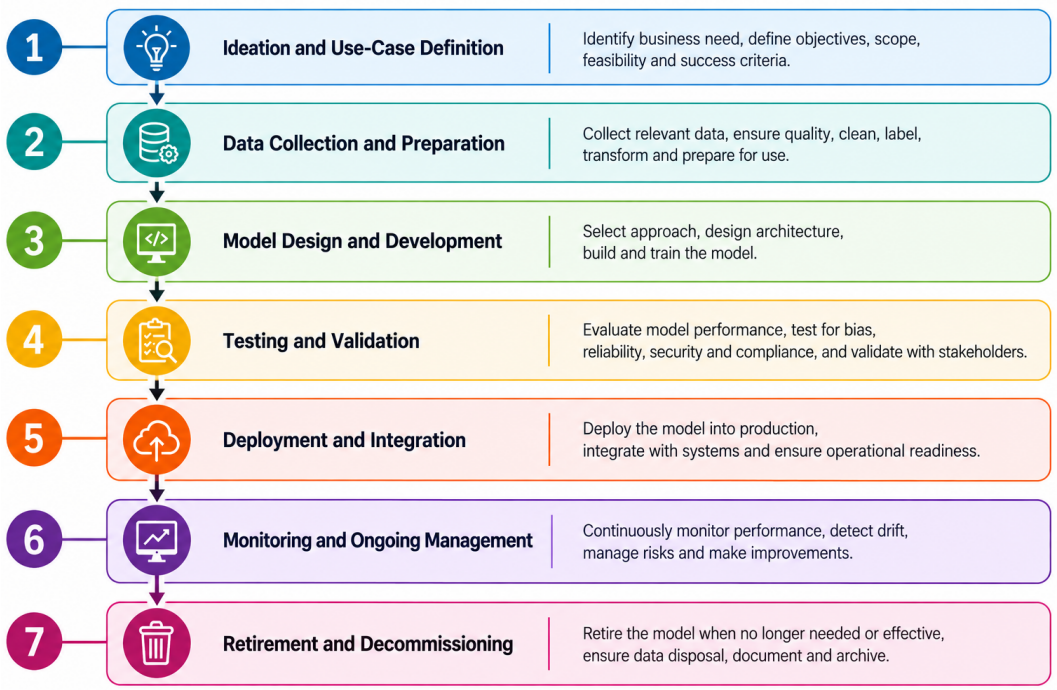


Figure 14.2: AI lifecycle

Let's go through these stages in detail.

Ideation and use-case definition

This stage focuses on identifying the business problem and deciding whether AI is appropriate for the business use case. The following are some of the governance activities at this stage:

- Conducting a formal use-case proposal and approval process
- Assessing AI suitability versus non-AI alternatives
- Classifying initial AI risk (for example low, medium, high)
- Considering ethical and societal impacts
- Aligning with organizational strategy and risk appetite
- Obtaining approval from the designated governance authority for high-risk use cases

Use case: Let's say a bank proposes using AI to assess loan applications and provide loan approval recommendations. The business team first documents the proposed use case and evaluates whether AI is appropriate compared with traditional rule-based credit assessment. The organization then performs an initial risk classification, considering factors such as potential financial impact, fairness, discrimination, and customer impact. The use case is also assessed against the organization's business strategy and risk appetite. If classified as high risk, the proposal is submitted to the designated AI governance committee for approval before development begins.

Data collection and preparation

AI systems depend heavily on data quality and legality. Following are some of the governance activities at this stage:

- Identifying and approving data sources
- Assigning data ownership and stewardship
- Conducting privacy and data protection assessments
- Validating consent, purpose limitation, and lawful use
- Identifying bias and mitigation measures
- Enforcing data security and access controls

Use case: A healthcare provider that uses a medical diagnostic AI system should identify and approve reliable sources such as clinical records, diagnostic images, and laboratory results, while assigning ownership and stewardship responsibilities. Privacy and data protection requirements should be assessed, and consent, lawful use, and purpose limitations should be validated before patient data is used. The data should also be assessed for potential bias—for example, whether certain patient populations are underrepresented—and appropriate mitigation measures should be implemented. Finally, strong security and access controls should be enforced to protect sensitive patient information.

Model design and development

During development, governance ensures that models are designed responsibly. This includes documentation of algorithms, training methods, assumptions, explainability requirements, and security controls. Key governance controls at this stage include:

- Documenting model objectives and intended use
- Obtaining approval for algorithms and training approaches
- Defining explainability and transparency requirements
- Following secure development practices

- Documenting assumptions, limitations, and dependencies
- Maintaining segregation of duties between development and validation

Use case: An educational institution using an AI system to grade student assignments should document the model's objectives and intended use, such as assisting teachers in evaluating assignments against approved grading criteria. The algorithms and training approach should be reviewed and approved, with explainability requirements defined so that teachers can understand the basis for grades assigned by the AI. Secure development practices should be followed to protect student submissions and personal information. The organization should also document assumptions, limitations, and dependencies, such as the AI's difficulty in evaluating creative or unconventional answers. Finally, development and validation should be performed by separate individuals or teams to provide independent assurance of the model's accuracy, fairness, and reliability.

Testing and validation

Before deployment, the AI system must be tested for accuracy, bias, robustness, and unintended outcomes. Pre-deployment validation is a major checkpoint, but **test, evaluation, verification, and validation** (**TEVV**) continues during development, deployment, monitoring and after material changes. Following are some of the governance activities at this stage:

- Conducting independent testing and validation
- Testing accuracy, robustness, and bias
- Conducting stress testing for abnormal or edge cases
- Validating against the approved use-case scope
- Reviewing and approving risk mitigation actions
- Obtaining formal sign-off prior to production deployment

Use case: A car company making driverless cars that use AI systems should conduct independent testing and validation to verify that the system performs reliably under real-world driving conditions. Accuracy, robustness, and potential bias should be tested across different road conditions, weather, traffic situations, and environments. Stress testing should include abnormal and edge cases, such as sudden pedestrian crossings, unexpected obstacles, sensor failures, and poor visibility. The system should be validated against its approved use-case scope, and identified risks should have documented and approved mitigation actions. Finally, formal sign-off should be obtained from the designated authority before the system is deployed for production use.

Deployment and integration

This stage focuses on controlled release into production. Key governance controls at this stage include:

- Obtaining deployment approval and following the change management process
- Implementing role-based access controls and logging
- Establishing human oversight and intervention mechanisms
- Communicating with users and providing transparency
- Confirming that deployment matches the approved design
- Obtaining go-live authorization from responsible stakeholders

Use case: If a bank decides to use an AI chatbot to answer customer queries, it must obtain deployment approval and ensure that the chatbot has passed the required change management process. Role-based access controls and logging should be implemented to restrict administrative access and maintain an audit trail of chatbot activities. Human intervention mechanisms should be established for complex, sensitive, or high-risk customer queries. Customers should be informed when they are interacting with an AI chatbot and provided with clear options to contact a human representative. Before going live with the chatbot, the organization should confirm that the deployed chatbot matches the approved design, including its data sources, capabilities, and restrictions. Finally, formal go-live authorization should be obtained from the responsible business, technology, risk, and governance stakeholders.

Monitoring and ongoing management

AI systems can change behavior over time due to data drift, model updates or evolving usage. Following are some of the governance activities at this stage:

- Ongoing performance and drift monitoring
- Conducting periodic bias and fairness reviews
- Establishing incident detection, reporting, and response processes
- Retraining and model update approval process
- Conducting regular compliance and control effectiveness reviews
- Reassessing risk classification periodically and also after material changes

Use case: An organization using an AI system to detect spam messages should continuously monitor detection accuracy and model drift to ensure that legitimate messages are not increasingly classified as spam. Periodic reviews should assess whether the system unfairly flags messages from particular users, languages, or communication channels. An incident detection and response process should address issues such as large numbers of legitimate

messages being blocked or malicious messages bypassing detection. Retraining and model updates should require approval before implementation, particularly when new types of spam are identified. Regular compliance and control-effectiveness reviews should confirm that the system continues to operate as intended. Finally, the organization should periodically reassess the risk classification if the system's purpose, data sources, or impact change.

Retirement and decommissioning

When an AI system is no longer needed, governance ensures secure decommissioning, data disposal, documentation retention, and lessons-learned reviews:

- Obtaining formal decommissioning approval
- Securely shutdown of AI system and interfaces
- Retaining or disposing of data in accordance with policy
- Preserving required audit and compliance records
- Conducting and documenting a lessons-learned review
- Updating the AI inventory and risk registers

Use case: For an AI system used to assist software development, the organization should obtain formal approval before decommissioning the system and securely shut down the AI tool, its integrations, and associated interfaces. Data such as source code, prompts, and user information should be retained or securely disposed of in accordance with organizational policies. Required audit trails, approval records, and compliance documentation should be preserved. A lessons-learned review should be conducted to capture issues and improvements identified during the system's use. Finally, the AI system inventory and relevant risk registers should be updated to reflect its retirement.

Effective AI lifecycle governance is not possible without clear ownership and accountability. Each stage of the AI lifecycle involves different activities, risks, and decisions, and therefore requires defined roles and responsibilities. Without this clarity, governance controls may exist on paper but fail in practice. We already discussed the RACI chart for AI GRC in *Chapter 12*.

From an AI GRC perspective, roles should be assigned to ensure that risk identification, compliance checks, ethical considerations, and approval decisions are consistently applied throughout the lifecycle. Responsibilities should not rest solely with technical teams; business, legal, risk, and compliance functions must also be involved.

Third-party and vendor AI lifecycle governance

AI lifecycle governance also applies when an AI system is provided by a third-party vendor or SaaS provider.

As a service receiver, the organization should obtain reasonable assurance that the service provider follows appropriate AI lifecycle governance processes. This assurance may be obtained through due diligence, direct audits, independent third-party audit reports, ISO/IEC 42001 certification, or clearly defined contractual requirements. The organization should consider the following activities:

- Conduct vendor due diligence before procurement.
- Assess security, privacy, and compliance requirements at periodic intervals.
- Define clear AI-related requirements in the contract.
- Ensure the service provider has appropriate AI lifecycle and change management processes.
- Monitor the AI service and vendor performance on an ongoing basis.

Summary

In this chapter, we discussed each stage of the AI lifecycle in detail. We explored how effective governance begins by clearly defining the AI use case and confirming that AI is appropriate for the intended purpose. We then examined the importance of using appropriate, lawful, secure, and representative data; developing AI models with documented objectives, controls, and clear accountability; and independently testing and validating systems before deployment.

We also discussed that AI governance is not a one-time approval activity. It is a continuous process that ensures risks, responsibilities, controls, and decisions remain appropriate throughout the AI system's entire lifecycle.

The next chapter focuses on the security of AI systems. It examines how organizations can protect AI systems from threats, misuse, and unintended harm, and how security and safety controls integrate with AI governance to ensure resilient, trustworthy, and responsible use of AI.

15

Security of AI Systems

Consider a bank that uses an AI model to detect fraudulent transactions. The model is trained using historical transaction data and continuously receives new transaction data to improve its ability to identify suspicious patterns. An employee with authorized access to transaction data deliberately manipulates certain transaction records so that the AI system does not recognize them as fraudulent. The bank may already have strong traditional security controls, such as user authentication, access controls, network security, antivirus protection, and logging. However, these controls may not detect whether the data being provided to the AI system has been deliberately manipulated. As a result, the employee's actions could influence the model's behavior and allow fraudulent transactions to pass through the fraud detection process. This highlights why traditional security controls may need to be supplemented with AI-specific controls.

AI security is not limited to traditional IT security controls. It also includes protecting data used to train AI models, safeguarding models from manipulation or misuse, securing AI infrastructure, and ensuring that AI outputs cannot be easily exploited or tampered with.

In this chapter, we will focus on the practical aspects of securing AI systems across their lifecycle. This chapter highlights why AI security is a critical pillar of AI governance, what makes AI security different from traditional system security, and how organizations can design and implement effective, risk-based security controls to protect AI systems in real-world environments.

Understanding the unique security risks of AI systems

AI systems must also protect data, models, algorithms, and outputs. Each of these components can be targeted in different ways, creating new attack surfaces that organizations must understand and manage. Though security risks are primarily managed by the CISO, as an AI

GRC professional, you need to have a basic understanding about AI security risks. Let us understand some of AI systems' specific risks:

- **Data poisoning**: Data poisoning occurs when an attacker intentionally introduces incorrect, misleading, or malicious data into the dataset used to train or fine-tune an AI model.
  For example, an attacker adds fraudulent transactions to a fraud-detection training dataset and labels them as legitimate. The AI model may then learn that certain fraudulent patterns are normal.

  This can have impacts such as inaccurate predictions, biased decisions, reduced model reliability, financial losses, and reputational damage.

- **Adversarial attacks**: Adversarial attacks involve manipulating input data in a way that causes an AI model to produce an incorrect result, even though the manipulation may appear insignificant to a human.
  For example, an attacker places a small sticker or mark on a road sign, causing an AI-based camera system to incorrectly identify the sign. This can have impacts such as incorrect decisions, safety incidents, and operational disruption.

  This can have impacts such as incorrect decisions, security failures, safety incidents, financial losses, and operational disruption.

- **Prompt injection**: Prompt injection occurs when an attacker provides specially crafted instructions to manipulate a generative AI system into ignoring its intended instructions or performing actions that it was not designed to perform.
  For example, an attacker enters a malicious instruction into an enterprise chatbot connected to internal documents and attempts to make the chatbot reveal confidential information.

  This can have impacts such as unauthorized disclosure of information, bypassing security controls, misuse of AI capabilities, and compromise of connected systems.

- **Model theft and extraction**: Model theft occurs when an attacker attempts to obtain, replicate, or reconstruct a proprietary AI model through unauthorized access or repeated interaction with the model.
  For example, an attacker sends thousands of carefully designed queries to a company's AI service and uses the responses to develop a similar model.

  This can have impacts such as loss of intellectual property, competitive advantage, proprietary technology, revenue, and research investment.

- **Model inversion**: Model inversion occurs when an attacker analyzes the outputs of an AI model to infer sensitive information about the data used to train the model.

For example, an attacker repeatedly queries a healthcare AI model and analyzes its responses to determine information about individuals whose medical records may have been used during training.

This can have impacts such as exposure of sensitive information, privacy breaches, regulatory violations, and loss of customer trust.

- **AI supply chain attacks**: AI systems often depend on third-party models, datasets, software libraries, APIs, plugins, cloud services, and other components. A compromise of any of these components can introduce risks into the AI environment.
  For example, an organization downloads a third-party AI model that has been modified to contain malicious code or a hidden backdoor.

  This can have impacts such as unauthorized access, data theft, manipulated AI outputs, malware infection, and compromise of other organizational systems.

- **Denial-of-service and resource exhaustion**: AI systems, particularly large AI models, can require significant computing resources to process requests. Attackers may exploit this by sending excessive or computationally expensive requests.
  For example, an attacker sends a large number of complex requests to a generative AI service, consuming excessive computing resources and making the service slow or unavailable to legitimate users.

  This can have impacts such as service disruption, increased infrastructure costs, degraded performance, reduced availability, and loss of business operations.

Let us now understand some best practices for protecting the data, models, and relevant infrastructure.

Protecting data, models, and infrastructure

AI systems depend on three critical components: data, AI models, and the underlying infrastructure. Infrastructure refers to the computing, storage, networks, platforms, software, security, and other technologies needed to run an AI system throughout its lifecycle. A weakness in any one of these components can affect the confidentiality, integrity, availability, privacy, and reliability of the AI system. Therefore, organizations should establish appropriate security controls to protect each component throughout the AI lifecycle.

Protecting AI data

Data is one of the most important assets in an AI system because it is used to train, test, validate, and operate AI models. Organizations should protect data from unauthorized access, modification, disclosure, and destruction. For example, an organization uses customer data to train an AI-based loan assessment model. If unauthorized employees can access or modify this

data, sensitive customer information could be exposed or the model could learn from inaccurate information. Organizations should consider controls such as:

- **Data classification**: Identify and classify data used by AI systems according to its sensitivity and business importance. For example, customer financial information used to train a loan assessment model should be classified as sensitive or confidential information.
- **Access control**: Restrict access to AI datasets to authorized users, applications, and AI systems based on business requirements. For example, only the data science team and approved AI applications should be able to access a customer training dataset.
- **Encryption**: Encrypt sensitive AI data both at rest and in transit to reduce the risk of unauthorized disclosure.
- **Data quality**: Validate data before it is used for training, testing, or operation to identify inaccurate, incomplete, outdated, or inconsistent information. For example, before training a loan approval model, the organization should check whether customer income, repayment history, and other information are complete and accurate.
- **Data masking and anonymization**: Mask, anonymize, or pseudonymize personal and sensitive information where appropriate. For example, a testing environment may use masked customer names and account numbers instead of actual customer information.
- **Data lineage**: Maintain information about the source, ownership, purpose, processing, and use of datasets. For example, the organization should be able to identify where a training dataset came from, who approved it, and which AI models use it.
- **Data retention**: Define how long AI-related data should be retained and securely dispose of data when it is no longer required. For example, customer records that are no longer required for an approved AI purpose should be securely deleted according to the organization's retention requirements.
- **Data monitoring**: Monitor access to and modification of important datasets to identify suspicious activity. For example, an alert can be generated if a user who normally accesses a small dataset suddenly downloads the entire training dataset.
- **Backup and recovery**: Maintain appropriate backups of critical datasets to support recovery from accidental deletion, corruption, or malicious modification. For example, if a training dataset is accidentally deleted or corrupted, the organization should be able to restore an approved version.

Organizations need to have well-defined roles and responsibilities to implement the above-mentioned processes. We already discussed the RACI matrix for AI GRC in *Chapter 12*.

The next section discusses key controls for protecting AI models.

Protecting AI models

AI models can represent significant intellectual property and may contain proprietary logic, configurations, learned patterns, and information derived from training data. Organizations should consider following controls:

- **Access control**: Restrict access to AI models and model repositories to authorized personnel. For example, only authorized data scientists and system administrators should be able to download or modify a production AI model.
- **Version control**: Maintain approved versions of AI models and track changes made during development, testing, and deployment. For example, if version 3.2 of a model is approved for production, the organization should be able to identify exactly which model version is currently deployed.
- **Model validation**: Verify the integrity of the model using appropriate security controls, such as trusted model sources, repository access controls, cryptographic hashes or digital signatures where available, and security testing. These measures help identify unauthorized modification or potential compromise.
- **Model security testing**: Assess models for security weaknesses such as adversarial attacks, backdoors, model extraction, and unexpected behavior. For example, security testers may attempt to manipulate inputs to determine whether the model can be easily fooled.
- **Change management**: Require appropriate authorization and review before significant changes are made to production models. For example, a data scientist should not directly replace a production model without testing, approval, and documented change management.
- **Model backup**: Maintain secure copies of approved models to support recovery from unauthorized modification or model corruption. For example, if the production model is compromised, the organization should be able to restore a previously approved model.
- **Model retirement**: Establish procedures to securely remove obsolete, unsupported, or compromised models. For example, an outdated AI model that is no longer supported should be removed from production and its access permissions revoked.
- **Output validation and sanitization**: Validate and sanitize AI-generated outputs before they are passed to downstream systems, browsers, or business processes. For example, AI-generated content should be checked for malicious code, unsafe instructions, or inappropriate content before being used by another application.

Above procedures provides reasonable assurance about model integrity throughout the lifecycle. Now let us discuss best practices for protecting the infrastructure that supports AI systems.

Protecting AI infrastructure

AI systems depend on infrastructure such as software, servers, cloud platforms, databases, networks, containers, GPUs, storage systems, and other supporting technologies:

- **Asset inventory**: Maintain an inventory of infrastructure supporting AI systems, including servers, cloud resources, databases, GPUs, containers, networks, and storage. For example, the organization should know which cloud servers, databases, GPUs, and APIs support each production AI system.
- **Secure configuration**: Configure AI infrastructure according to approved security baselines and hardening standards. For example, unnecessary services and ports should be disabled on a server hosting an AI application.
- **Vulnerability management**: Regularly identify, assess, prioritize, and remediate vulnerabilities affecting AI infrastructure. For example, if a vulnerability is identified in the operating system supporting an AI application, it should be assessed and remediated according to the organization's vulnerability management process.
- **Patch management**: Apply security patches and updates to operating systems, databases, applications, libraries, and supporting components. For example, a security patch released for the database supporting an AI application should be evaluated and applied within the organization's defined timeframe.
- **Identity and access management**: Ensure that only authorized users and systems can access AI infrastructure. For example, only approved administrators should have administrative access to the production server hosting an AI application.
- **Multi-factor authentication**: Use MFA, particularly for privileged accounts and administrative access. For example, an administrator accessing a cloud environment hosting AI workloads should be required to use MFA.
- **Least privilege**: Provide users, applications, and AI agents only the permissions required to perform their intended functions. For example, an AI chatbot that only needs to read customer information should not have permission to delete or modify customer records.
- **Network segmentation**: Separate AI infrastructure from other critical environments where appropriate to limit the impact of a security compromise. For example, production AI servers can be placed in a separate network segment from employee workstations.

- **Endpoint protection**: Protect servers, workstations, and other endpoints used to develop, deploy, and operate AI systems. For example, developer laptops used to access AI model repositories should have appropriate endpoint protection and security controls.
- **Logging and monitoring**: Record and monitor authentication, administrative activities, system changes, and other security-relevant events. For example, the organization should record who accessed a production AI server and what administrative changes were made.
- **Backup and recovery**: Establish appropriate backup and recovery mechanisms for critical AI infrastructure and configurations. For example, infrastructure configurations can be backed up so that an AI environment can be rebuilt after a major system failure.
- **Availability protection**: Implement capacity management, redundancy, and resilience controls to reduce the risk of service disruption. For example, a critical AI service may use redundant infrastructure so that a single server failure does not make the service unavailable.

The AI GRC professional should work with the CISO, security, technology, data, privacy, and business teams to ensure that AI-specific security risks are monitored and documented throughout the AI lifecycle. Let's see what their responsibilities look like.

Role of the AI GRC professional in the security of AI systems

While the primary responsibility for security rests with the CISO and the information security function, AI systems introduce risks that are not always fully addressed by traditional security programs. The AI GRC professional plays a complementary role by ensuring that AI-specific security risks and regulatory requirements are explicitly identified, governed, and adhered to, rather than being assumed to be covered by existing controls:

- **Clarifying shared responsibility with the CISO**: The CISO remains accountable for overall cybersecurity strategy, controls, and incident response. The AI GRC professional works alongside the CISO to identify gaps where AI systems create new attack surfaces, risk scenarios, or compliance obligations that require additional or adapted controls.
- **Ensuring coverage of AI-specific security risks**: AI GRC professionals ensure that risks are explicitly assessed and addressed. Their role is to make sure these risks are not overlooked within generic IT security assessments.
- **Ensuring adherence to regulatory and legal requirements**: AI GRC professionals help identify applicable requirements and advise on their integration into controls and

processes. This includes aligning security controls with requirements related to data protection, cybersecurity, sector-specific regulations, and emerging AI regulations, and ensuring evidence of compliance is maintained.

- **Oversight of AI security assessments and reviews**: AI GRC professionals oversee AI risk assessments, security reviews, and impact assessments at key lifecycle stages, ensuring that identified security and regulatory issues are documented, tracked, and remediated in coordination with the CISO's team.
- **Supporting incident management and regulatory reporting**: During AI-related security incidents, AI GRC professionals support the CISO by ensuring proper escalation, documentation, and regulatory or contractual notifications are carried out in a timely and compliant manner.
- **Driving continuous improvement of AI security governance**: By analyzing incidents, audit findings, regulatory changes, and emerging threats, the AI GRC professional continuously improves AI security governance to keep it aligned with both evolving technologies and regulatory expectations.

In practice, the CISO owns security execution, while the AI GRC professional ensures that AI-specific risks and regulatory requirements are fully covered, visible, and governed within the organization's overall security framework.

Summary

In this chapter, we learned that AI security goes beyond traditional IT security because AI systems introduce additional risks related to data, models, prompts, APIs, outputs, infrastructure, and third-party components. We understood that AI risks are not limited to only IT risks. We learned about AI-specific security risks such as data poisoning, adversarial attacks, prompt injection, model theft, model inversion, supply chain attacks, and denial-of-service or resource exhaustion.

We discussed that an AI GRC professional has an important role in identifying and assessing AI-specific security risks and working with the CISO and security teams to implement appropriate controls. This includes ensuring that security risks are documented, controls are monitored, incidents are managed, and the effectiveness of controls is periodically reviewed. Even with strong preventive controls, AI-related security incidents can still occur.

The next chapter focuses on how organizations detect, respond to, and learn from AI incidents to minimize impact and strengthen governance over time.

16

AI Incident Management

We have taken the example of a bank using an AI system to approve loan applications to understand a lot of concepts in the previous chapters. Building on the same example, consider the AI system being used by the bank has been tested and approved for production, and the bank has established normal IT and cybersecurity controls. After several months, the bank notices that the AI system is rejecting applications from a particular customer group at a much higher rate than expected. Investigation shows that a change in the data used by the AI system has affected its decisions. The issue may now involve not only technical performance but also fairness, customer impact, regulatory compliance, and reputational risk.

How should the bank respond? Who should declare the issue as an AI incident? Who should investigate it? When should senior management, Legal, Compliance, or the AI Governance Committee be informed? Should the AI system continue operating while the investigation is underway?

This example shows why **AI incident management** is different from **traditional IT incident management**. Traditional incident response assumes predictable systems. AI systems are different because their outputs can vary. A single safety gap can therefore produce many harmful outputs before the issue is detected and reviewed. An AI incident may involve technical problems, but it can also affect individuals, compliance, fairness, trust, and the organization's reputation.

AI incident management therefore needs a structured process for identifying, reporting, assessing, responding to, documenting, and resolving AI-related incidents:

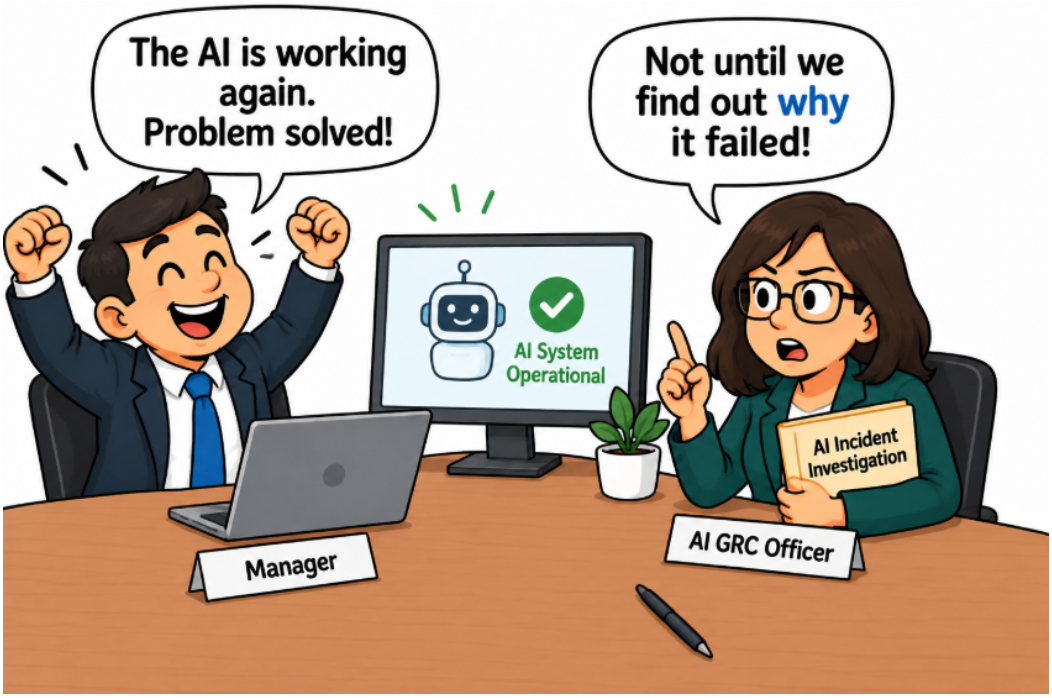


Figure 16.1: AI incident management

The objective of incident management is not only to address the immediate problem but also to identify the root cause, implement corrective actions, and use the lessons learned to strengthen **AI governance**.

In this chapter, we will discuss the criteria for identifying and classifying AI incidents and explore an **AI incident management framework** for effectively detecting, reporting, assessing, responding to, and learning from AI-related incidents. We will examine the role and responsibilities of an AI GRC professional in AI incident management, including incident oversight, coordination, risk assessment, reporting, and follow-up. The chapter will also introduce an **AI incident register** for systematically recording and tracking incidents and examine common pitfalls organizations may face when managing AI incidents. Finally, we will provide an **AI incident management checklist** to help organizations establish a consistent and effective approach to managing AI-related incidents.

Incident criteria

Not every issue related to an AI system should be treated as an AI incident. A clear definition of what constitutes an AI incident is essential to ensure timely reporting, consistent handling, and appropriate escalation.

An AI incident is any event or condition where an AI system behaves in a way that causes, or has the potential to cause, harm, noncompliance, or significant risk. This harm may be technical, ethical, legal, operational, or reputational in nature.

Clear criteria help employees and stakeholders recognize incidents early and ensure they are reported and managed through the formal incident management process rather than being handled informally or ignored. The criteria should be aligned with the organization's risk appetite, regulatory obligations, and AI governance framework. The following are some examples which demonstrate how criteria can be defined:

Area	Suggested Threshold
AI system availability	Downtime > 30 minutes for a critical AI system
Customer impact	≥ 10 customers materially affected
Financial impact	Direct loss ≥ USD 10,000
AI performance	Performance decreases by >10% from the approved threshold
Error rate	Error rate increases by >20% from the approved baseline
Data exposure	Any unauthorized disclosure of sensitive or regulated data
Security compromise	Any confirmed unauthorized access to an AI model, training data, or AI infrastructure
Prompt injection	Any successful prompt injection that bypasses a security or access control
Regulatory breach	Any confirmed or suspected material regulatory violation
Unauthorized AI deployment	Any AI system deployed without required approval

Area	Suggested Threshold
High-risk control failure	Any failure of a mandatory control
Third-party failure	Critical third-party AI service unavailable for > 30 minutes
Data poisoning	Any confirmed manipulation of training or validation data
Model manipulation	Any confirmed unauthorized modification of a production model

Table 16.1: Incident criteria

These thresholds should be treated as illustrative starting points and customized according to the organization's size, industry, AI use cases, risk appetite, contractual obligations, and applicable laws and regulations. While defining thresholds, the organization should consider:

- **Criticality of the AI system:** Critical or high-risk AI systems should generally have lower thresholds. For example, a small performance decrease in a medical AI system may require immediate investigation.
- **Severity of impact:** Consider the potential impact on individuals, customers, safety, finances, privacy, compliance, and reputation. A single affected person may be enough to trigger an incident if the potential harm is serious. For example, an AI system used for loan approval incorrectly rejects a single customer because of a discriminatory model as it may have significant implications.
- **Data sensitivity:** Unauthorized access to sensitive, personal, or regulated data may require immediate escalation regardless of the number of records affected.
- **Regulatory requirements:** Consider applicable laws and regulations and their reporting requirements. These requirements should take precedence over internal thresholds.
- **Security and AI-specific risks:** Consider events such as prompt injection, data poisoning, model manipulation, unauthorized model changes, or loss of human oversight.
- **Potential impact:** Do not wait for actual harm if there is a credible possibility of significant harm. For example, unauthorized access to a production AI model should be investigated even when no misuse has yet been confirmed.
- **Recurring or systemic issues:** Multiple small incidents may indicate a larger control weakness and should be assessed collectively.

Identifying an incident is only the starting point. Once an AI-related event is identified, the organization needs a structured process to assess its impact, assign responsibilities, respond appropriately, and ensure that the incident is properly resolved and documented. Let us now discuss the incident management framework.

AI incident management framework

AI incident management should not operate as a standalone or parallel process. Ideally, it should be integrated with the organization's existing incident management framework used for IT, cybersecurity, data privacy, and operational risks. This integration ensures consistency, faster response, and better governance oversight. Considering the emerging risk and regulated impact, it is advisable to treat AI incidents as high-priority items. AI incident management policy and procedure documents should at least include the aspects shown in *Figure 16.2*:



Figure 16.2: Incident management framework

We will now look at each aspect of the framework in detail:

- **Incident Definition:** Clearly define what qualifies as an AI-related incident.
- **Roles and Responsibilities:** Assign teams or individuals who will detect, review, and respond to incidents. Example: System owner investigates the incident, while Legal decides if it needs to be reported.

- **Incident Reporting Process:** Describe how staff or users can report suspected AI-related incidents. Example: Use a centralized reporting platform or email channel such as ai.incident@company.com.
- **Incident Management Procedures:** Define clear incident management procedures along with well-structured responsibilities and accountabilities. For example:
    - **Initial assessment and triage:** Identify and record the incident, assess its severity and potential impact, and determine whether immediate containment or escalation is required.
    - **Investigation and root cause analysis:** Investigate what happened, identify the affected AI system, data, users, or processes, and determine the underlying cause of the incident.
    - **Remediation and response:** Contain the incident and take corrective actions to reduce or eliminate its impact. This may include restricting system use, correcting data or models, applying technical controls, or temporarily suspending the AI system.
    - **Regulatory reporting:** Determine whether the incident must be reported to regulators, customers, affected individuals, or other stakeholders based on applicable legal and regulatory requirements.
    - **Post-incident review and lessons learned:** Review the incident, evaluate the effectiveness of the response, identify what could be improved, and update controls, procedures, or risk assessments as necessary.
    - **Documentation:** Maintain records of the incident, investigation, decisions, actions taken, communications, approvals, and lessons learned to support accountability, auditability, and future improvement.
- **Incident Logging and Categorization:** Requirement to maintain a record of each incident with details such as time, description, type, impact level, and resolution steps.
- **Communication and Escalation:** Define how incidents are escalated internally and how users are informed.
- **Third Party Roles and Responsibilities:** When creating an AI incident response plan, it is essential to identify third-party tools integrated into the system. If an incident occurs, such as a malfunction, security breach, or misuse of the AI system, those third parties may also be impacted or legally require notification.

> **Note**
>
> For your ready reference, a detailed incident management procedure is included in AIMS Policy, *Template 4*, in this book's GitHub repository. To access the repository link, follow the steps in the *Download the example code files* section in the *Preface*.

The AI GRC professional plays an important role in coordinating governance, risk, compliance, escalation, and documentation throughout the incident lifecycle. Let us now discuss their roles.

Role of the AI GRC professional in AI incident management

The AI GRC professional plays a central role in ensuring that AI incidents are managed in a structured manner. While technical teams focus on fixing models or systems, the AI GRC professional ensures that risk and compliance requirements are met throughout the incident lifecycle. Key responsibilities of the AI GRC professionals include:

- **Incident identification and declaration:** The AI GRC professional defines clear criteria for what constitutes an AI incident and ensures these criteria are communicated across the organization. They support teams in recognizing when an issue should be formally declared as an AI incident and not handled informally.
- **Incident classification and severity assessment:** The AI GRC professional ensures that AI incidents are classified correctly and assigned appropriate severity levels. This includes assessing regulatory, operational, and reputational impact in line with the organization's risk appetite.
- **Governance and escalation oversight:** The AI GRC professional ensures timely escalation of AI incidents to the appropriate governance bodies such as the AI governance committee, senior management, legal, or compliance teams. They confirm that escalation decisions are documented and justified.
- **Regulatory and compliance coordination:** For incidents with legal or regulatory implications, the AI GRC professional coordinates with compliance and legal teams to assess notification requirements. They help ensure that regulatory reporting is accurate, timely, and aligned with applicable laws and standards.
- **Incident documentation and record keeping:** The AI GRC professional ensures that complete and accurate records are maintained for each AI incident. This includes

incident details, decisions taken, root cause analysis, corrective actions, and supporting evidence for audits or reviews.

- **Post incident review and lessons learned:** After incident closure, the AI GRC professional supports post incident reviews and lessons learned exercises. They ensure that gaps in controls, policies, or processes are identified and addressed.
- **Continuous improvement of AI governance:** Based on incident trends and findings, the AI GRC professional recommends updates to AI policies, risk assessments, lifecycle controls, and training programs. This helps strengthen the overall AI governance framework and reduce the likelihood of future incidents.

Effective incident management requires more than responding to individual events. The organization should maintain a consistent record of incidents so that management can track their status, identify recurring problems, and monitor corrective actions in the form of an incident register.

Incident register

Organizations should maintain a common Incident Register for recording and tracking incidents across different areas, including IT, cybersecurity, privacy, operational, and AI-related incidents. AI incidents should be recorded in the same register, with an appropriate incident category or classification to identify AI-related events.

The register can include fields such as:

Field	Information to Record
Incident ID	Unique incident reference
Date and Time	Date and time of occurrence or detection
Incident Category	IT, Cybersecurity, Privacy, AI, Operational, etc.
Incident Type	Security breach, system failure, data issue, AI output issue, etc.
Affected System	System or AI system involved
Incident Description	Description of what happened
Severity	Critical, High, Medium, or Low

Field	Information to Record
Business Impact	Operational, financial, customer, regulatory, or reputational impact
Incident Owner	Person responsible for managing the incident
Initial Action	Immediate action taken
Root Cause	Cause identified during investigation
Corrective Action	Action required to prevent recurrence
Action Owner	Person responsible for corrective action
Target Date	Expected completion date
Regulatory Reporting	Whether notification was required
Status	Open, Under Investigation, Remediated, or Closed
Closure Date	Date of closure
Lessons Learned	Improvements identified

Table 16.2: Incident register

For AI-related incidents, the organization can use an additional AI-specific classification or tag rather than creating a separate register. For example, an incident involving prompt injection can be classified as **Cybersecurity – AI**, while an unfair AI decision can be classified as **AI – Fairness**.

This approach provides a consolidated view of organizational incidents while allowing management and the AI GRC Officer to filter and analyze AI-related incidents separately. It also avoids maintaining duplicate incident management processes and supports integration with the organization's existing incident management framework.

Having an incident management process does not guarantee that it will work effectively. Certain weaknesses can cause delays, incomplete responses, or repeated incidents. Let us examine the common pitfalls that organizations should avoid.

Common pitfalls in AI incident management

Organizations may have an incident management process in place but still struggle to manage AI incidents effectively. The following are common pitfalls that the AI GRC professional should watch for:

- **Not clearly defining an AI incident:** If the organization does not establish clear criteria, employees may not know when an AI-related issue should be formally reported and escalated.
- **Creating a separate AI incident process:** AI incidents should generally be integrated with the existing IT, cybersecurity, privacy, and operational incident management processes rather than creating a completely separate process.
- **Delaying incident reporting:** Employees may wait until the cause or impact is fully understood before reporting an issue. Incidents should be reported as soon as they meet the organization's defined criteria so that assessment and containment can begin quickly.
- **Focusing only on technical impact:** AI incidents can also create ethical, legal, regulatory, customer, and reputational impacts. These should be considered when assessing incident severity.
- **Unclear roles and responsibilities:** If it is not clear who investigates, who decides whether an incident requires regulatory notification, and who approves closure, response can be delayed. Responsibilities should be clearly documented.
- **Poor escalation:** Significant AI incidents should be escalated to the appropriate governance bodies, senior management, Legal, Compliance, CISO, or other relevant functions. Failure to escalate an incident at the right time can increase its impact.
- **Incomplete incident records:** Failing to document the incident, decisions, investigation, root cause, corrective actions, and supporting evidence makes it difficult to demonstrate effective incident management.
- **Closing incidents without addressing the root cause:** Resolving the immediate problem does not necessarily eliminate the underlying risk. The organization should identify why the incident occurred and address the weaknesses that allowed it to happen.
- **Not tracking corrective actions:** Corrective actions should have clear owners and target dates and should be followed until completion. Otherwise, the same weakness may cause another incident.

- **Ignoring third-party responsibilities:** AI incidents may involve external AI models, APIs, datasets, or service providers. Contracts and incident procedures should clearly define notification, investigation, cooperation, and escalation responsibilities.
- **Treating incident closure as the end of the process:** Incident management should result in lessons learned and improvements to policies, risk assessments, controls, monitoring, and training.

The AI GRC professional should periodically review the incident register for recurring AI-related incidents, overdue corrective actions, and repeated control failures. The objective is not simply to close incidents but to use incident information to strengthen AI governance and reduce the likelihood of recurrence.

AI incident management checklist

The following checklist can be used by the AI GRC professional to assess whether the organization has established an effective AI incident management process:

☑ AI incident definition and criteria are documented

☑ Numerical thresholds for incident severity are defined

☑ AI incidents are integrated with the common organizational Incident Management process

☑ Employees have a clear mechanism to report suspected AI incidents

☑ AI incidents are recorded in the common Incident Register

☑ AI incidents are categorized and assigned an appropriate severity level

☑ Incident ownership and responsibilities are clearly defined

☑ Initial assessment and triage procedures are established

☑ AI incidents are escalated according to defined thresholds

☑ CISO, Legal, Compliance, Privacy, Risk, and other relevant functions are involved when required

☑ Procedures for containment and immediate response are defined

☑ Root cause analysis is performed for significant incidents

☑ Regulatory and contractual notification requirements are assessed

☑ Third-party responsibilities and notification requirements are defined

☑ Corrective actions are documented, assigned, and tracked to completion

☑ Supporting evidence and incident records are retained

☑ Incidents are formally closed after required actions are completed

☑ Post-incident reviews are performed for significant incidents

☑ Lessons learned are documented

☑ AI risk assessments and controls are updated where required

☑ Recurring incidents and trends are periodically analyzed

☑ AI incident management procedures are periodically reviewed and updated

This checklist complements the incident management framework by helping the AI GRC professional verify that the process is not only documented but also implemented and operating effectively.

Summary

In this chapter, we learned that AI incident management provides a structured approach to identify, report, assess, respond to, and resolve AI-related incidents. We saw that organizations should define clear incident criteria and thresholds, integrate AI incidents into the common incident management process, and establish clear roles, escalation, documentation, and corrective actions.

We explored the role an AI GRC professional plays in coordinating governance, risk, compliance, and regulatory requirements throughout the incident lifecycle. We discussed that organizations should also maintain a common incident register, conduct root cause analysis, track corrective actions, and use lessons learned to improve AI governance and reduce recurring risks.

So far, we understand that incident management focuses on responding to issues after they occur. In the next chapter on AI Audit, we will learn that organizations must also proactively evaluate whether AI controls, processes, and governance mechanisms are working as intended.

17

AI System Audits

Imagine a hospital that uses an AI system to assist doctors in identifying patients who may be at high risk of developing a serious medical condition. The hospital has taken AI governance seriously and followed all the steps we've discussed so far. It has approved an AI governance policy, defined roles and responsibilities, completed an AI risk assessment, documented the model and training data, performed security and privacy assessments, established human oversight, implemented performance monitoring, and defined an AI incident management process. The system was also properly tested and approved before deployment. On paper, everything appears to be in place.

However, there is one critical gap: **independent oversight**. The AI system has never been independently reviewed to determine whether the established governance, risk management, and control mechanisms are actually operating effectively.

The hospital relies on its own clinical, technology, and operational teams to monitor the system. While these activities are important, they do not provide the same level of independent assurance. Without independent oversight, weaknesses in controls may remain undetected, particularly when those responsible for operating the system are also responsible for assessing its effectiveness.

For example, the system may appear to be performing accurately based on its overall accuracy metrics, while an independent review could identify that its performance is significantly lower for a particular patient group. Similarly, an independent audit may identify weaknesses in how AI-related incidents are reported, how model changes are approved, or how human intervention is documented.

This scenario demonstrates why an independent AI systems audit is an important part of AI governance. It provides objective assurance that AI governance, risk management, data handling, security, fairness, explainability, human oversight, and other controls are working as intended.

Unlike traditional IT audits, AI audits require a deeper examination of areas such as type of AI (LLM, Agentic etc.), training data quality, model behavior, explainability, human oversight, and continuous monitoring mechanisms. Auditors must evaluate both technical controls and governance practices to ensure that AI risks are identified, documented, and appropriately mitigated. This chapter explores how AI systems can be audited effectively, what key control areas should be reviewed, and how organizations can establish a practical and repeatable approach to provide assurance over the responsible and compliant use of AI. Before examining how an AI system should be audited, it is important to establish who should perform the audit.

In this chapter, we will discuss the responsibilities of the audit function in providing independent assurance over AI governance and controls. We will examine the role of the AI GRC professional in planning, coordinating, and supporting AI system audits, as well as the key audit areas that should be considered when evaluating AI systems. The chapter will also cover the different types of AI audits, including internal and external audits, and explain how each type contributes to effective AI governance and assurance. Let us first understand the responsibility of the audit function.

Responsibility of audit

Ideally, the AI GRC Officer should not act as the auditor of the AI systems they are responsible for governing. The AI GRC Officer may coordinate the audit, provide documentation, explain governance processes, and track remediation, but the actual audit should be performed by an independent internal audit function or an appropriately independent external assessor.



Figure 17.1: AI audit independence

This separation helps avoid a conflict of interest. If the AI GRC Officer designs or coordinates the governance controls and then audits those same controls, there is a risk of self-review. Independent audit provides objective assurance to management and the AI governance committee that the controls are appropriately designed and operating effectively.

Responsibility for AI systems audits should rest with the **Chief Audit Officer** (**CAO**) through the internal audit function. The internal audit team should include an auditor with appropriate AI knowledge and skills. Where the organization does not have sufficient internal expertise, it may engage an independent external assessor or specialist to perform the AI audit. The CAO should ensure that the external assessor is sufficiently independent and appropriately qualified for the scope and complexity of the AI system.

Although the audit itself should be performed independently, the AI GRC professional has an important role in supporting the audit process. Let us now look at how the AI GRC professional can facilitate the audit while maintaining the required separation from the audit function.

Role of AI GRC professional in AI system audits

While the internal audit function or independent assessors typically conduct the audit, the AI GRC professional plays a critical role in ensuring that AI-specific risks, governance expectations, and regulatory requirements are appropriately addressed throughout the audit lifecycle. AI system audits require coordination across technical, business, risk, and compliance functions. The AI GRC professional acts as a bridge between auditors, AI development teams, business owners, and senior management to ensure that the audit is both effective and aligned with organizational objectives:

- **Ensuring AI risk coverage:** The AI GRC professional helps identify AI-specific risks that must be included within the audit scope, such as bias, explainability, model drift, and ethical considerations. They ensure that these risks are documented in the risk register and that appropriate controls exist to mitigate them.
- **Aligning audit scope with regulatory and framework requirements:** AI systems may be subject to multiple regulations, standards, and governance frameworks. The AI GRC professional ensures that the audit scope considers applicable requirements, internal policies, and industry best practices. This includes verifying that documentation, monitoring, and accountability mechanisms are aligned with regulatory expectations.
- **Facilitating audit planning and coordination:** The AI GRC professional coordinates between auditors and various stakeholders, including AI developers, data owners, security teams, and business users. They help ensure that relevant documentation, evidence, and access to systems are available during the audit.

- **Supporting control design and documentation:** Prior to and during the audit, the AI GRC professional ensures that governance controls, procedures, and records related to the AI lifecycle are properly documented. Well-maintained documentation improves audit readiness and reduces the likelihood of audit observations.
- **Monitoring remediation activities:** If gaps or weaknesses are identified, the AI GRC professional tracks remediation actions, assigns ownership, and ensures that corrective measures are implemented within defined timelines. They also validate that implemented controls effectively address the identified risks.
- **Promoting continuous improvement:** AI system audits should not be treated as one-time activities. The AI GRC professional analyzes audit findings, identifies recurring issues, and drives improvements in governance processes, risk management practices, and control effectiveness across future AI initiatives.
- **Supporting management reporting and assurance:** The AI GRC professional consolidates audit outcomes and presents meaningful insights to senior management.

In practice, while primary accountability for implementing and operating AI system controls remains with system owners, technology leadership, and the CISO, the AI GRC professional ensures that AI-related governance, risk, compliance, and ethical requirements are consistently embedded, monitored, and auditable across the organization.

The next question is what the auditor should actually examine. An AI systems audit should cover both governance and technical controls and should focus on the risks that could affect the reliability, security, fairness, and compliance of the AI system. Let us now discuss the key audit areas in AI systems.

Key audit areas in AI systems

An AI systems audit should go beyond checking whether policies and documents exist. The auditor should determine whether controls are appropriately designed, implemented, and operating effectively. The audit should cover both technical and governance aspects of the AI system, including data, model performance, security, privacy, fairness, explainability, human oversight, and monitoring. Let us understand some of the AI-specific risks to be audited.

AI governance and oversight

The auditor should review whether appropriate AI governance mechanisms are established and operating effectively. This includes reviewing the AI governance policy, defined roles and responsibilities, AI inventory, risk assessments, approval processes, and oversight by the AI governance committee or equivalent body. The auditor should verify a sample of AI systems to confirm that ownership, risk assessment, approval, and governance requirements have

actually been completed. Evidence may include the AI policy, AI inventory, RACI matrix, risk assessments, approval records, and governance committee minutes.

Data quality and integrity

One of the primary risk areas is data quality and integrity. If training or testing data is incomplete, inaccurate, outdated, or biased, the AI model may produce incorrect or unfair outcomes. Auditors should verify whether appropriate controls exist for data sourcing, validation, cleansing, and classification.

Bias and fairness

AI models can unintentionally discriminate against certain groups if the training data reflects historical biases. Auditors should review whether fairness testing, bias detection mechanisms, and periodic reviews are implemented to minimize discriminatory outcomes.

Transparency and explainability

Lack of transparency is a common concern in AI systems. Some models operate as *black boxes*, making it difficult to understand how decisions are made. This creates challenges in regulated environments where justification of automated decisions is required. Auditors should assess whether explainability tools and sufficient documentation are available.

Security risks in AI systems

AI systems may be vulnerable to threats such as data poisoning, adversarial attacks, unauthorized model access, or model theft. The audit should confirm that appropriate access controls, encryption, monitoring, and secure development practices are implemented.

Privacy and regulatory compliance

Privacy and regulatory compliance are critical, especially where personal or sensitive data is used. Auditors should verify adherence to applicable privacy laws, consent requirements, and data minimization principles.

Model performance and reliability

AI models may degrade over time due to changes in data patterns, commonly known **as model drift**. Continuous monitoring, performance testing, and periodic retraining controls should be reviewed to ensure consistent accuracy and reliability.

Human oversight and accountability

AI systems should not operate without defined ownership, escalation mechanisms, and the ability for humans to intervene when necessary. Auditors must confirm that responsibilities are clearly assigned and that appropriate oversight exists.

AI incident management

The auditor should assess whether AI-related incidents are identified, reported, investigated, escalated, and resolved through the organization's common incident management process. The audit should review incident criteria, severity thresholds, the incident register, root cause analysis, corrective actions, and closure records. For significant incidents, the auditor should verify whether lessons learned resulted in improvements to AI controls, policies, or risk assessments.

For each audit area, the auditor should follow a consistent approach: understand the control, obtain evidence, test the control, identify exceptions, assess the risk, document the finding, agree on corrective action, and follow up on remediation. This ensures that the audit focuses not only on whether controls are documented, but also on whether they have been implemented and are operating effectively.

> **Note**
>
> A detailed AI systems audit checklist is also provided as *Template 11* for practical use during the audit. To access the template, follow the steps in the *Download the example code files* section in the *Preface*.

Types of AI audits

An AI GRC Officer should understand the different types of AI audits to ensure that the appropriate audit approach is applied based on the organization's needs, risk profile, and available expertise. AI audits can generally be performed internally by the organization's Internal audit function or externally by an independent third party:

- **Internal AI audit:** An internal AI audit is performed by the organization's internal audit function. It provides independent assurance to management and the AI governance committee on whether AI governance, risk management, and controls are appropriately designed and operating effectively. The internal audit team should have sufficient AI knowledge and expertise to evaluate AI-specific risks.
- **External AI audit:** An external AI audit is performed by an independent third-party assessor or specialist. Organizations may use external auditors when specialized AI expertise is not available internally, when greater independence is required, or when the scope and complexity of the AI system requires specialized assessment.

An organization implementing ISO/IEC 42001 should conduct internal audits to determine whether its AI management system conforms to the organization's requirements and the

standard. If the organization wants to become ISO/IEC 42001 certified, the certification audit is performed by an independent external certification body.

The AI GRC Officer may coordinate and support either type of audit by providing relevant documentation, facilitating access to stakeholders and systems, explaining governance processes, and tracking remediation. However, the AI GRC Officer should not audit the controls they are responsible for governing, as this could create a self-review or conflict-of-interest situation.

Summary

In this chapter, we learned why independent AI systems auditing is essential for effective AI governance and explored the key areas that should be audited. We also examined the difference between internal and external audits.

AI systems audit provides independent assurance that AI governance, risk management, and controls are appropriately designed and operating effectively. The audit should cover key areas such as governance, data quality, fairness, explainability, security, privacy, model performance, human oversight, and incident management. The responsibility for the audit should ideally rest with the Chief Audit Officer through an independent internal audit function, supported by AI expertise or an independent external specialist where required. The AI GRC professional should support the audit by coordinating with stakeholders, providing evidence, and tracking remediation, but should not audit the controls they govern. A practical AI audit should focus not only on whether controls are documented, but also on whether they are implemented, tested, and operated effectively.

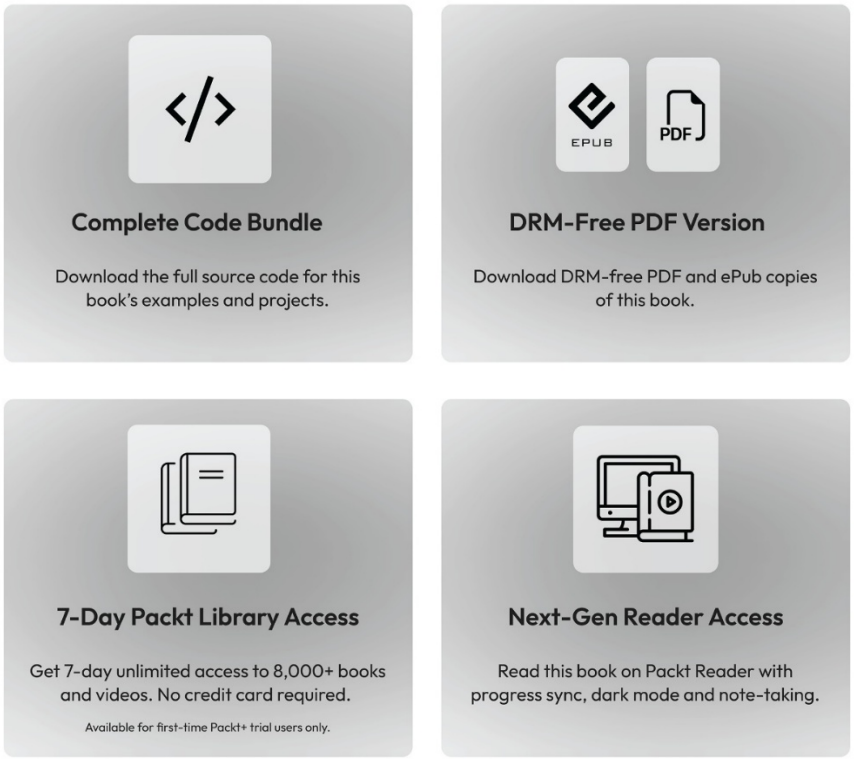
This chapter marks the end of this book. At this point, you are aware of how organizations can establish a practical approach to AI governance and manage AI responsibly throughout its lifecycle. We explored AI governance structures, responsible AI principles, risk management, AI impact assessments, and key AI governance frameworks and regulations. We also learned how ISO/IEC 42001 can provide a structured management system for AI governance.

To make the implementation of these concepts easier for you, the book provides practical templates, checklists, and tools that organizations can adapt to their own AI governance programs.

18

Unlock Access to the Code Bundle and the PDF Version

Your copy of this book includes the following exclusive benefits:



Follow this guide to unlock them. The process takes only a few minutes and only needs to be completed once.

Unlock this book's free benefits in three easy steps

Step 1

Have your purchase invoice ready for *step 3*. If you have a physical copy, scan it using your phone and save it as a PDF, JPG, or PNG.

For more help on finding your invoice, visit `https://www.packtpub.com/unlock-benefits/help`.

> **Note**
>
> **Note**: If you bought this book directly from the Packt website, no invoice is required. After *step 2*, you can access your exclusive content right away.

Step 2

Scan the QR code or go to `packtpub.com/unlock`.

On the page that opens (similar to *Figure 18.1* on desktop), search for this book by name and select the correct edition.

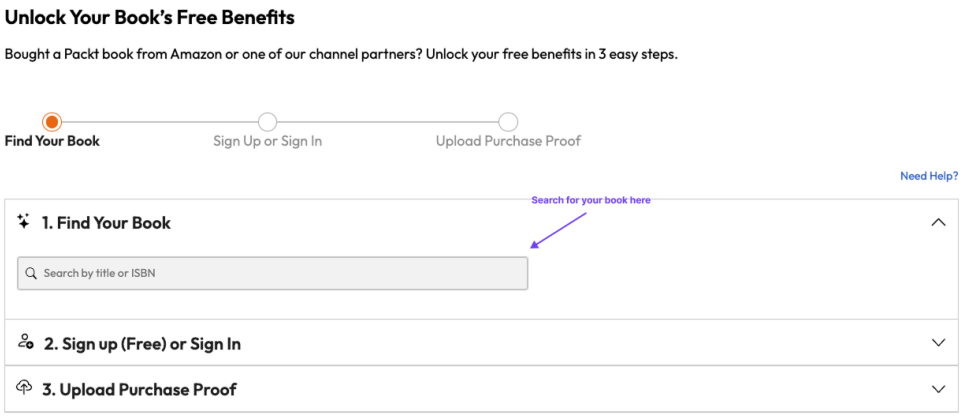


Figure 18.1 – Packt unlock landing page on desktop

Step 3

After selecting the book, sign in to your Packt account or create one for free. Then, upload your invoice (PDF, PNG, or JPG, up to 10 MB). Follow the on-screen instructions to finish the process.

Need help?

If you get stuck and need help, visit `https://www.packtpub.com/unlock-benefits/help` for a detailed FAQ on how to find your invoice and more. This QR code will take you to the help page.

Note

Note: If you are still facing issues, reach out to `customercare@packt.com`.



packtpub.com

Subscribe to our online digital library for full access to over 7,000 books and videos, as well as industry leading tools to help you plan your personal development and advance your career. For more information, please visit our website.

Why subscribe?

- Spend less time learning and more time coding with practical eBooks and Videos from over 4,000 industry professionals
- Improve your learning with Skill Plans built especially for you
- Get a free eBook or video every month
- Fully searchable for easy access to vital information
- Copy and paste, print, and bookmark content

At www.packtpub.com, you can also read a collection of free technical articles, sign up for a range of free newsletters, and receive exclusive discounts and offers on Packt books and eBooks.

Other Books You May Enjoy

If you enjoyed this book, you may be interested in these other books by Packt:



AI Under Attack

Kris Kimmerle, David Okeyode

ISBN: 9781806119936

- Identify AI-specific risks and clearly communicate them to business teams
- Defend models, data, RAG, and agents from threats like poisoning, prompt injection, jailbreaking, and data exfiltration
- Design resilient cloud/MLOps with Zero Trust, supply chain security, and isolation
- Build secure APIs, apps, and agents with strong auth, validation, and safe tool use
- Apply AI-focused GRC, alignment checks, bias mitigation, monitoring, and incident response
- Translate complex concepts into actionable steps, using threat intel and collaboration for lasting security



(ISC)2 Certified in Cybersecurity (CC) Exam Guide

Hemang Doshi

ISBN: 9781807301934

- How to apply the CIA triad, authentication, and privacy concepts
- Basic risk assessment and choosing suitable cybersecurity methods
- How to classify administrative, technical, and physical controls
- Business continuity, disaster recovery, and incident response
- The implementation of least privilege, RBAC, and the segregation of duties
- How to recognize common network attacks and apply core safeguards
- Encryption, data handling, and logging to support operations

Packt is searching for authors like you

If you're interested in becoming an author for Packt, please visit `authors.packt.com` and apply today. We have worked with thousands of developers and tech professionals, just like you, to help them share their insight with the global tech community. You can make a general application, apply for a specific hot topic that we are recruiting an author for, or submit your own idea.

Share your thoughts

Now you've finished *AI Governance in Practice*, we'd love to hear your thoughts! Scan the QR code below to go straight to the Amazon review page for this book and share your feedback or leave a review on the site that you purchased it from.



https://packt.link/r/1807300811

Your review is important to us and the tech community and will help us make sure we're delivering excellent quality content.

Index

A

AI GRC Officer **149**

AI GRC professional **189**
- role, in AI incident management 197, 198
- role, in AI system audits 205, 206

AI Impact Assessment (AIIA) **117**
- benefits 118
- conducting 121, 122
- key areas 118 – 121
- requirements 117, 118

AI RMF crosswalks **92**

AI RMF playbook **91**

AI RMF profiles **91**
- Current Profile 89
- Target Profile 89

AI Risk Management Framework (AI RMF) **48, 54**

AI audits
- external AI audit 208
- internal AI audit 208

AI change management records **165**

AI deployers **111**

AI deployment options **13**
- cloud deployment 13
- edge deployment 14
- on-premises deployment 14

AI documents
- creating, stepwise procedure 171 – 173

AI elements **5**
- algorithm 6
- data 6
- model 6

AI governance **19**
- implementing, pitfalls 154, 155
- KPIs 24, 25
- policies 22, 23
- RACI matrix 150 – 152
- roles 148 – 150
- roles and responsibilities 21, 22

AI governance committee records **165**

AI governance documents **161**

AI governance frameworks
- EU AI Act 138
- ISO/IEC 42001 137
- NIST AI RMF 137
- OECD AI principles 138
- overview 137, 138

AI governance policy **161, 162**
- key contents 166 – 169

AI governance review **165**

AI governance structure
- implementing 152, 153
- objectives 147, 148

AI governance themes **56 – 58**

AI impact assessment **162**

AI incident management
- AI GRC professional, role 197, 198
- checklist 201
- framework 192, 195, 196
- pitfalls 200
- procedure 164

AI lifecycle governance **175**

AI lifecycle stages **176**
- AI lifecycle governance 182
- data collection and preparation 178
- deployment and integration 180
- ideation 177
- model design and development 178, 179
- monitoring and ongoing management 180, 181

retirement and decommissioning 181
testing and validation 179
third-party vendor 182
use case definition 178

AI model and system documentation 164

AI models
decision tree model 11
large language models (LLMs) 12
multimodal AI model 12
natural language processing (NLP) model 12
personalization model 12
recognition AI model 12
robotic model 12
types 11

AI monitoring and review records 164

AI providers 111

AI risk assessment 125, 163
mitigation/treatment 133, 134
requirements 127, 128
risk analysis 131
risk evaluation 133
risk identification 129, 130
steps 128

AI risk register 163

AI risks
ethical risks 126
legal and regulatory risks 126
operational risks 127
reputational risks 127
security risks 126
technical risks 126
types 126

AI security 183
data, protecting 185, 186
infrastructure, protecting 188
models, protecting 187, 188

AI security risks
adversarial attacks 184
AI supply chain attacks 185
data poisoning 184
denial-of-service and resource exhaustion 185
model inversion 184
model theft and extraction 184
prompt injection 184

AI system audits
AI GRC professional, role 205, 206

AI system inventory 162

AI system owner 149

AI systems 4, 20, 104, 105
classification, determining 109 – 111
high risk AI 106, 107
key audit areas 206
limited risk AI 108
minimal risk AI 108
prohibited AI 106
unique security risks 183 – 185

AI use case assessment and approval form 162

AI vendor and third-party assessment 163

AI-based systems
governing 21
versus normal system 2

AIMS structure, under ISO/IEC 42001 63
Annex A 65, 66
four annexes 64, 65
Statement of Applicability (SoA) 66
ten clauses 63, 64

Artificial Intelligence Management System (AIMS) 48, 53, 61, 77, 127, 137

accountability 29, 30
practical implementation 32, 33
under EU AI Act 30, 31

algorithm 6

algorithm types 7
reinforcement learning algorithms 9
semi-supervised learning algorithms 8
supervised learning algorithms 7
unsupervised learning algorithms 7, 8

artificial intelligence (AI) 1, 2
elements 5
normal system, versus AI-based system 2, 3

audit
responsibility 204, 205

C

Chief Audit Officer (CAO) **205**
Chief Information Security Officer (CISO) **38**
Cybersecurity Framework (CSF) **78**
certification audit process **69 – 72**
cloud deployment **13**
compliance, with laws and standard **44**
practical implementation 44
continuous monitoring and improvement **42, 43**
practical implementation 43

D

Data Protection Impact Assessments **128**
data **6**
decision tree model **11**
deep learning **10**
distributors **111**
document hierarchy **159**
guidelines 161
policy 160
procedures 160
records and evidence 161
templates and checklists 161
documented AI policies and procedures
importance 158, 159

E

European Union Artificial Intelligence Act (EU AI Act) **49, 50, 140, 170**
applicability 104
complaint and redress mechanism 112
enforcement 112 – 114
fines 112 – 114
global impact 103
obligations by role 111
penalties 112 – 114
practical implementation 114, 115
roadmap
reference link 104
edge deployment **14**
ethical risks **126**
ethical use **41**
practical implementation 42
explainability **35, 36**
practical implementation 36

F

fairness **33**
practical implementation 34
financial services company case study, ISO/IEC 42001 **67**
AI governance policy, creating 67
certification audit 68
controls, maintaining 68
gap analysis, performing 67
SoA, preparing 68
frameworks **48**

G

General Data Protection Regulation (GDPR) **50 – 52, 117, 128**
Govern function **80**
accountability and responsibility, establishing 81
AI risk culture, establishing 82
diversity and equity, ensuring 82
policies and processes, developing 81
stakeholders, engaging 82
third-party AI risks, managing 83
third-party failures, preparing 83
governance **15, 27**
and regulations 18
elements 16, 17
governance model **148, 149**
guidelines **18, 19**

H

high risk AI systems **106, 107**
human oversight **40**
practical implementation 40

I

IEEE 7000 series **55**
ISO
advantages 140, 141
ISO/IEC 27001 certificate, leveraging **72**
governance process, aligning 73
internal audit processes, aligning 74
management review process, aligning 74
risk management processes, aligning 73
scope, reviewing 73
ISO/IEC 42001
scenarios, for using 139, 140
ISO/IEC 42001 artificial intelligence management systems **170**
ISO/IEC 42001 certification
importance 72
ISO/IEC 42001 implementation **66, 67**
financial services company case study 67
pitfalls 68, 69
ISO/IEC 42001 standard
NIST AI RMF mapped 143, 144
OECD principles mapped 141 – 143
IT governance **19**
Incident Register **198, 199**
Incident criteria **193 – 195**
Information Security Management System (ISMS) **72**
Institute of Electrical and Electronics Engineers (IEEE) **55**
International Electrotechnical Commission (IEC) **53, 61**
International Organization for Standardization (ISO) **53, 61**
international standards 62
standard 62
importers **111**

K

key audit areas, AI systems
AI governance and oversight 206
AI incident management 208
bias and fairness 207
data quality and integrity 207
human oversight and accountability 207
model performance and reliability 207
privacy and regulatory compliance 207
security risks 207
transparency and explainability 207

L

labels **7**
large language model (LLM) **13, 29**
legal and regulatory risks **126**
limited risk AI **108**

M

Manage function
AI risks, managing 87
benefits, maximizing 87
impacts, minimizing 87
post-deployment monitoring 88
third-party AI risks, monitoring 88
Map function **83**
AI capabilities 84
AI impacts, assessing 84
AI system, categorizing 84
benefits and costs 84
context, establishing 83

risks and benefits, mapping across AI system 84

Measure function
AI risk metrics, applying 85
AI risks, tracking over time 86
AI trustworthiness, evaluating 85
measurement effectiveness, assessing 87

machine learning (ML) 9
deep learning 10
neural network 10

management 17

minimal risk AI 108

model 6

model drift 207

multi-modal models 12

N

NIST AI RMF components 78 – 80
Govern function 80
Manage function 87
Map function 83
Measure function 85

NIST AI Risk Management Framework (RMF) 77, 78, 137, 171
components 78
implementing 88
scenarios, for using 139

NIST AI RiskManagement Framework (RMF) 117

NIST Privacy Framework, NIST SP 800-53 78

National Institute of Standards and Technology (NIST) 77, 78

natural language processing (NLP) model 12

neural network
uses 10

neurons 10

non-discrimination 33

normal system
versus AI-based system 2, 3

O

OECD AI principles 54, 55, 95, 96, 117, 171
background and evolution 94
mapping, to AI governance frameworks 100 – 102
practical applications, in organizations 97 – 100
scenarios, for using 138, 139
scope and applicability 95

Organization forEconomic Co-operationand Development (OECD) 54, 94

on-premises deployment 14

operational risks 127

operations 17

P

personalization model 12

policies 18

privacy and data protection 36, 37

procedures 18

prohibited AI 106

R

RACI matrix
for AI governance 150 – 152

Responsible AI principles 28, 29
accountability 29, 30
compliance with laws and standards 44
continuous monitoring and improvement 42
ethical use 41
explainability 35, 36
fairness and non-discrimination 33
human oversight 40
privacy and data protection 36, 37
risk-based approach 41
safety and reliability 39
security and resilience 37, 38
transparency 34

recognition AI model 12

regulations **47**
reinforcement learning algorithms **9**
reliability **39**
reputational risks **127**
resilience **38**
risk analysis **131**
likelihood and impact 131, 132
techniques 131
risk evaluation **133**
risk identification
techniques 129, 130
risk management policy **134**
risk mitigation/ treatment **133**
acceptance 134
avoidance 133
reduction/mitigation 133
transfer 133
risk register **135**
risk-based approach **41**
practical implementation 41
robotic model **12**

S

safety **39**
safety and reliability
practical implementation 39
security **37**
security risks **126**
semi-supervised learning algorithms **8**
standards **48**
supervised learning algorithms **7**

T

technical risks **126**
traditional systems **20**
training and awareness records **165**
transparency **34**
practical implementation 35

U

US Executive Order **52, 53**
unsupervised learning algorithms **7, 8**

W

World Economic Forum AI Governance Playbooks (WEF AI Governance Playbooks) **55, 56**