‹packt›



Artificial Intelligence Governance Professional (AIGP) Certification Guide

Prepare for the AIGP exam with AI governance, risk, compliance, and responsible AI frameworks



Pavan Kumar | Avnish Singh

Artificial Intelligence Governance Professional (AIGP) Certification Guide

Prepare for the AIGP exam with AI governance, risk, compliance, and responsible AI frameworks

Pavan Kumar and Avnish Singh

‹packt›

Artificial Intelligence Governance Professional (AIGP) Certification Guide

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.
The author acknowledges the use of cutting-edge AI, such as ChatGPT, with the sole aim of enhancing the language and clarity within the book, thereby ensuring a smooth reading experience for readers. It's important to note that the content itself has been crafted by the author and edited by a professional publishing team.
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 authors, 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: Sunith Shetty
Relationship Lead: Deepesh Patel
Project Manager: Aniket Shetty
Content Engineer: Szwenska Chriss
Technical Editor: Shweta Amale
Indexer: Pratik Shirodkar
Production Designer: Shantanu Zagade
Growth Lead: Varun Nair

First published: September 2026

Production reference: 1020926

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

ISBN 978-1-80808-331-0
www.packtpub.com

Contributors

About the authors

Pavan Kumar is a technology architect, author, and cybersecurity professional with experience in enterprise IT, cloud computing, cybersecurity, data security, endpoint management, and artificial intelligence governance. His expertise includes Microsoft Azure, Microsoft Intune, enterprise architecture, cloud and data security, governance, risk management, compliance, and AI technologies. He has authored professional training courses and technical publications covering cloud computing, cybersecurity, endpoint management, and Artificial Intelligence, helping technology professionals develop practical skills.

Avnish Singh is a technology professional specializing in artificial intelligence, cloud technologies, cybersecurity, and enterprise governance. He has contributed to the development of certification content and technical learning resources that help bridge the gap between emerging technologies and practical business implementation. His areas of interest include AI governance, responsible AI, regulatory compliance, enterprise risk management, and emerging global AI regulations. As a co-author, he contributed to practical explanations, governance scenarios, and exam-focused content for this book.

Table of Contents

Preface **xi**

Free benefits with your book ..... **xv**

Chapter 1: Foundations of Artificial Intelligence and AI Governance **1**

Defining AI, its types, and applications **2**

What is artificial intelligence • 2

Viewing AI as a sociotechnical system • 3

Classifying AI systems using the OECD's five dimensions • 4

Recognizing modern drivers of AI adoption • 6

Distinguishing between ANI, AGI, and ASI • 7

Artificial narrow intelligence • 7

Artificial general intelligence • 7

Artificial superintelligence • 8

Common AI use cases • 8

Introducing AI governance, principles, and frameworks **9**

Defining AI governance • 9

Distinguishing principles from frameworks • 10

Applying a comprehensive approach to AI governance • 11

Understanding AI model types and machine learning fundamentals **12**

Distinguishing AI, ML, deep learning, and generative AI • 12

Common AI models • 13

Exploring types of machine learning models • 14

Training AI systems using machine learning fundamentals • 15

Exploring transformer models, multimodal systems, and expert systems **16**

Working with transformers, NLP, and multimodal models • 16

Comparing model relationships and deployment options • 17

Understanding expert systems • 18

Managing AI platforms, infrastructure, and operational monitoring **19**

Choosing AI platforms for common use cases • 19

Understanding the AI technology stack • 20

Understanding compute infrastructure • 21
Implementing observability and monitoring • 22
Summary **23**

Chapter 2: AI Impacts on People and Responsible AI Principles 25

Understanding AI harms and their impact on people **25**
How to identify harm • 26
Mapping AI harms using a taxonomy approach • 27
Who is affected by AI harms? • 27
Understanding bias and discrimination in AI • 28
Assessing organizational harm from AI • 29
Assessing organizational risk from AI **30**
Managing security and operational risks • 30
Addressing privacy risks in AI systems • 31
Understanding business risks from AI • 32
Recognizing other emerging AI risks • 33
Applying ethical principles and frameworks **34**
Applying fair information practice principles to AI • 34
Working with OECD AI principles • 35
Identifying key ethical issues in AI • 36
Building and operationalizing responsible AI **38**
Mitigating ethical risks with foundational controls • 38
Building a culture of ethical AI • 39
Defining and achieving trustworthy AI • 40
Operationalizing responsible AI practices • 41
Summary **42**

Chapter 3: Responsible AI Governance and Risk Management 43

Building the organizational infrastructure for AI governance **44**
Tailoring AI governance to your organization • 44
Distinguishing AI developers, deployers, and users • 45
Building policies for oversight and accountability • 46
Clarifying roles and responsibilities • 47
Securing leadership support for AI governance • 49
Choosing governance structures and models • 49

Engaging stakeholders and aligning with organizational strategy 51
Identifying and engaging key stakeholders • 51
Designing role-specific training and awareness programs • 52
Aligning AI and risk strategy • 53
Designing and conducting AI risk assessments 54
Building scalable AI assessment processes • 55
Assessing risk through context and outcomes • 56
Scoring risk using probability and severity • 57
Applying NIST frameworks for AI risk management 58
Applying the NIST AI Risk Management Framework • 58
Using the NIST ARIA program for model evaluation • 60
Summary 62

Chapter 4: Governing AI Development 63

Governing the AI development lifecycle 63
Navigating the AI system development lifecycle • 64
Defining the business problem and use cases • 64
Determining the scope of an AI system • 65
Conducting impact assessments • 66
Engaging stakeholders in development • 68
Testing, monitoring, and operationalizing AI systems 69
Testing and validating AI system behavior • 70
Assessing baseline performance, robustness, and stability • 70
Testing for fairness and safety • 70
Validating operational readiness • 71
Managing and monitoring AI systems post-launch • 71
Maintaining documentation for AI operations • 73
Developing a communication plan • 74
Governing data across the AI pipeline 75
Applying data governance questions • 76
Tracing data lineage and provenance • 77
Gathering data with quality and format controls • 78
Wrangling and preparing data using the five V's • 80
Applying privacy-enhancing technologies in data preparation • 81
Governing model features and conformity documentation 83

Designing features for responsible model development • 83
Creating model cards and conformity documentation • 84
Summary 86

Chapter 5: Governing AI in Production 87

Establishing deployment readiness 88
Evaluating and updating privacy and security policies • 88
Understanding deployment requirements • 89
Governing vendor and third-party dependencies 91
Governing proprietary AI models • 91
Managing third-party risk • 92
Evaluating vendor and open-source agreements • 94
Applying deployment governance in practice • 95
Monitoring and responding to operational risk 96
Conducting periodic performance and safety assessments • 97
Monitoring for risks, mitigation, and model drift • 98
Managing incident response and downstream consequences • 100
Ensuring accountability and transparency 101
Meeting disclosure and transparency obligations • 101
Managing incidents, issues, and risks • 102
Auditing AI systems and demonstrating accountability • 104
Selecting deployment architectures and model patterns 105
Choosing a deployment environment • 105
Deploying in the cloud • 106
Deploying on-premises • 106
Deploying at the edge • 106
Selecting a model use pattern • 107
Deploying retrieval-augmented generation safely • 108
Governing agentic AI architectures in deployment • 109
Controlling tool access and permissions • 110
Implementing human oversight and confirmation gates • 110
Defending against prompt injection • 110
Building audit trails and recovery plans • 111
Summary 111

Chapter 6: The EU AI Act 113

Establishing the scope and actors of the EU AI Act **113**
Applying EU AI Act essentials to governance practice • 114
Determining when exemptions remove regulatory obligations • 115
Identifying provider, deployer, importer, and distributor roles • 115
Applying AI literacy and risk classification principles **117**
Developing role-based AI literacy for safe deployment • 117
Classifying AI systems using the four-tier risk framework • 118
Navigating obligations across the four risk tiers **119**
Recognizing and responding to prohibited AI practices • 119
Governing high-risk AI as a provider or deployer **120**
Fulfilling importer and distributor obligations for high-risk AI **121**
Meeting Article 50 transparency obligations for limited-risk AI • 122
Maintaining proportionate governance for minimal-risk AI **123**
Governing general-purpose AI and building operational compliance **124**
Managing obligations and systemic risk for general-purpose AI models • 125
Producing technical documentation and conducting assessments • 126
Preparing and maintaining technical documentation • 126
Implementing record-keeping for traceability and auditing • 127
Conducting conformity and impact assessments • 127
Designing human oversight, transparency, and quality management • 128
Preparing for enforcement and understanding penalty structures • 129
Summary **130**

Chapter 7: Other Laws and Standards Related to AI 133

Navigating the global AI regulatory landscape **133**
Understanding alignment and dissonance across jurisdictions • 134
Building AI governance under U.S. federal guidance • 135
Translating state privacy rules into product controls • 136
Implementing city and state employment and consumer AI controls **137**
Auditing automated hiring tools under NYC local law 144 • 137
Applying Illinois disclosure requirements for AI video interviews • 138
Managing high-risk AI and algorithmic discrimination under Colorado law • 139
Embedding privacy principles and governance best practices **140**

Applying core privacy principles across the AI lifecycle • 141
Implementing practical recommendations for providers and deployers • 142
Embedding trustworthiness and documentation for AI providers • 142
Establishing oversight and monitoring for AI deployers • 143
Applying international AI frameworks and ISO standards 143
Examining South Korea's AI basic law • 144
Establishing shared AI terminology with ISO/IEC 22989 • 144
Building an AI management system under ISO/IEC 42001 • 145
Defining context, scope, and leadership responsibilities • 146
Planning, operating, and continuously improving the AIMS • 146
Conducting AI system impact assessments with ISO/IEC 42005 • 147
Summary 148

Chapter 8: Unlock Your Exclusive Benefits 151

Unlock this Book's Free Benefits in 3 Easy Steps 152

Other Books You May Enjoy 156

Index 159

Preface

Artificial Intelligence is rapidly transforming the way organizations operate, make decisions, and deliver products and services. From generative AI assistants and intelligent automation to predictive analytics and autonomous systems, AI has moved from research laboratories into everyday business operations. As AI capabilities continue to expand, organizations face an equally important challenge: ensuring these technologies are developed, deployed, and governed responsibly.

Effective AI governance is no longer limited to legal compliance or ethical discussions. It has become an essential business capability that enables organizations to manage risk, protect individuals, maintain trust, and comply with an evolving global regulatory landscape. Whether implementing internal AI solutions or integrating third-party models, organizations need structured governance practices that balance innovation with accountability.

This book provides a practical, exam-focused guide to **Artificial Intelligence Governance**. It explains how governance frameworks, organizational processes, risk management, ethical principles, and regulatory requirements work together throughout the AI lifecycle. Each chapter combines conceptual understanding with practical examples and real-world scenarios, allowing readers to understand not only what governance requires but also why it matters and how it is applied in practice.

The content has been carefully structured to support professionals preparing for the **IAPP Artificial Intelligence Governance Professional (AIGP)** certification while also serving as a practical reference for organizations building responsible AI governance programs. Throughout the book, concepts are introduced progressively from AI fundamentals and governance principles to risk management, development governance, deployment governance, and regulatory compliance under the **EU AI Act**.

The book focuses on practical decision-making, governance thinking, and scenario-based learning. Each topic explains how governance decisions influence the complete lifecycle of AI systems, helping readers develop the analytical skills required for certification examinations as well as real-world governance responsibilities.

Who this book is for

This book is intended for professionals responsible for governing, managing, developing, deploying, or overseeing Artificial Intelligence systems within their organizations.

It is particularly valuable for:

- Professionals preparing for the **IAPP Artificial Intelligence Governance Professional (AIGP)** certification.
- Privacy, compliance, governance, risk, and legal professionals expanding into AI governance.
- Security architects, risk managers, auditors, and enterprise governance teams responsible for AI oversight.
- AI product managers, solution architects, data scientists, and software engineers who need to understand governance obligations throughout the AI lifecycle.
- Business leaders and decision makers responsible for implementing responsible AI practices across their organizations.

Although no advanced machine learning expertise is required, readers should have a basic understanding of business technology concepts and organizational governance processes.

What this book covers

Chapter 1, Foundations of Artificial Intelligence and AI Governance, introduces artificial intelligence, explains how AI systems operate as socio-technical systems, explores AI model types and machine learning fundamentals, and establishes the governance principles that support responsible AI adoption.

Chapter 2, AI Impacts on People and Responsible AI Principles, examines how AI affects individuals and organizations, explores harm taxonomies, bias, discrimination, privacy, organizational risk, and introduces the ethical principles that underpin responsible AI governance.

Chapter 3, Responsible AI Governance and Risk Management, explains how organizations build governance structures, assign roles and responsibilities, engage stakeholders, perform AI risk assessments, and apply governance frameworks such as the NIST AI Risk Management Framework.

Chapter 4, Governing AI Development, explores governance throughout the AI development lifecycle, including problem definition, use-case selection, impact assessments, data governance, testing, documentation, and lifecycle management practices that support trustworthy AI development.

Chapter 5, Governing AI in Production, focuses on operational governance after deployment, including deployment readiness, vendor and third-party governance, operational monitoring, incident response, accountability, transparency, and deployment architecture decisions that influence ongoing AI risk management.

Chapter 6, The EU AI Act, provides a practical guide to the European Union's AI regulatory framework, explaining its scope, risk-based classification model, governance obligations, provider and deployer responsibilities, AI literacy requirements, and compliance expectations for modern AI systems.

Chapter 7, Other Laws and Standards Related to AI, surveys AI regulations beyond the EU AI Act, covering U.S. federal and state requirements, city-level employment AI rules, core privacy principles, South Korea's AI Basic Law, and the ISO/IEC standards that support international AI governance.

To get the most out of this book

To gain the benefit from this book, you should have a general understanding of information technology, business processes, and organizational governance concepts. Prior knowledge of machine learning is helpful but not essential, as the book explains AI concepts from a governance perspective rather than a data science perspective. As you progress through each chapter, focus on understanding how governance principles apply across the entire AI lifecycle, as these relationships form the foundation for both real-world AI governance programs and **AIGP certification** scenarios. Working through the chapters sequentially will provide the strongest understanding, since each chapter builds upon concepts introduced earlier in the book.

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 *Artificial Intelligence Governance Professional (AIGP) Certification Guide*, 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/1808083318`

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 comes with free benefits to support your learning. 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: Keep your invoice handy. Purchases made directly from Packt don't require one

1

Foundations of Artificial Intelligence and AI Governance

Governing AI responsibly starts with understanding AI. Before you can assess the risk, apply a framework, or set a policy, you need a shared vocabulary and a clear mental model of how AI systems are built, trained, and deployed. Without that foundation, governance discussions tend to become either jargon-heavy or too abstract to put into practice.

This chapter builds that foundation. By the end of it, you will be able to define AI and describe it as a socio-technical system, classify AI systems using a structured framework, distinguish between major model types and training approaches, and map the technology stack that supports AI in production. These skills underpin every governance decision covered in the rest of the book.

In this chapter, we will cover the following topics:

- Defining AI, its types, and its applications
- Introducing AI governance, principles, and frameworks
- Understanding AI model types and machine learning fundamentals
- Exploring transformer models, multimodal systems, and expert systems
- Managing AI platforms, infrastructure, and operational monitoring

Note

Download 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*.

Defining AI, its types, and applications

This section provides the foundation for the rest of the chapter. It explains how AI functions within organizations, introduces the **Organization for Economic Co-operation and Development (OECD)** framework for classifying AI systems, and distinguishes between the three broad categories of AI.

What is artificial intelligence

You walk into a meeting, and someone says, *"We should use AI for this."* Within minutes, everyone is talking about something different. One person is thinking of a chatbot, another imagines automating a manual task, while someone else is talking about predictive analytics. Without a shared definition, teams can end up approving projects without agreeing on what they're building. That's why it's important to define AI before discussing governance.

AI didn't begin with chatbots. Early AI systems relied on explicit rules, with developers encoding expertise through *if-then* logic. Machine learning changed that approach by enabling systems to learn patterns from data instead of following only hand-written rules. This shift has important governance implications. When behavior is learned from data rather than programmed directly, reviewing the code alone is no longer enough to understand how a system will behave.

The question of what makes a machine "intelligent" has been debated for decades. In the 1950s, Alan Turing asked whether a machine should be considered intelligent if people could not reliably distinguish its responses from those of a human. Although AI has evolved considerably since then, Turing's question remains relevant because people often judge AI by its behavior rather than by how it reaches its conclusions. That perception influences how readily people trust and rely on AI systems.

For the purposes of AI governance, AI can be defined as a machine's ability to perform tasks that typically require human intelligence, such as recognizing patterns, making predictions, generating content, or taking actions based on data. While these capabilities may resemble human judgment, they are the result of computation.

Most AI systems can be understood by looking at four elements: the technology, its objective, its autonomy, and its output:

- **Technology:** refers to the model and the supporting tools that enable the system to perform its task.
- **Objective:** defines what the system is intended to achieve, such as detecting fraud, forecasting demand, or summarizing customer feedback. When objectives are poorly

defined, organizations often measure the wrong outcomes, creating the illusion of success while overlooking unintended harm.

- **Autonomy:** refers to the extent to which a system can operate without requiring human intervention at every step. Some AI systems only provide recommendations for people to review, while others can trigger actions automatically, such as blocking a payment or rejecting an application. As autonomy increases, so does the need for robust oversight.
- **Output:** is what the system produces, whether that's a score, a prediction, a recommendation, or generated content. Governance isn't concerned only with the output itself but with how that output is used. A recommendation can become policy, a risk score can determine who receives a service, and a generated summary can become the official record. The impact comes from how people incorporate AI outputs into decision-making.

Although AI systems may operate with varying levels of autonomy, humans remain accountable throughout their lifecycle. Humans define objectives, select training data, determine success metrics, adjust thresholds, retrain models, and decide when systems should be modified or retired. Governance makes these responsibilities explicit instead of allowing accountability to disappear once a model is deployed.

Within organizations, AI is commonly used for tasks such as image recognition, defect detection, fraud detection, cybersecurity monitoring, demand forecasting, customer personalization, conversational assistants, recommendation systems, and operational optimization. These applications often feel routine, but they still influence important decisions at scale. Whether an AI system recommends products, routes deliveries, detects fraudulent transactions, or generates customer communications, organizations need appropriate governance to manage risks such as bias, lack of transparency, automation errors, and excessive reliance on AI-generated outputs.

Viewing AI as a sociotechnical system

An AI system can perform well during testing and still fail soon after deployment. The model may behave exactly as expected, but people may interpret its outputs differently, rely on them too heavily, or use them in workflows it was never designed to support. That's because AI doesn't operate in isolation; it functions within a socio-technical system.

A socio-technical system combines both the technical components of AI and the human environment in which it operates. The technical side includes the model, the data it uses, and the surrounding software. The social side includes the people who use the system, their roles, training, organizational culture, incentives, and day-to-day ways of working. Together, these elements determine how AI behaves in practice.

In most organizations, an AI system produces an output, someone reviews it, and another person acts on it. Decisions about who sees the output first, whether it can be challenged, and what happens when people disagree all influence the outcome. If there are no effective review or escalation points, errors can spread quickly through a workflow.

Consider a résumé-ranking system introduced to help recruiters prioritize applications. Initially, it simply provides recommendations. Over time, however, recruiters may begin treating the ranked list as the default hiring order because it saves time. If the model learned from historical hiring data containing bias, that bias can be carried forward into future recruitment decisions. Without anyone intending it, a decision-support tool has become a decision-making tool.

This illustrates why AI risks often emerge gradually rather than immediately after deployment. As people become familiar with a system, they may rely on it more, skip verification steps, or stop documenting their own judgment because "the system already checked it." Generative AI makes this risk more pronounced because confident, well-written responses can appear authoritative even when they are incorrect. Individual mistakes are usually recoverable, but repeated reliance on inaccurate outputs can gradually undermine the quality of decisions.

Organizational incentives also shape how AI is used. Teams measured primarily at speed are more likely to accept AI outputs without scrutiny, whereas teams rewarded for quality, accuracy, and accountability are more likely to validate results and escalate concerns when something appears wrong. Governance helps establish these expectations by defining where human oversight is required and how AI-generated outputs should be reviewed before decisions are made.

For this reason, effective governance extends beyond evaluating model performance. It also considers how AI changes human behavior and how work is performed. Controls such as review checkpoints, audit logs, clear accountability, and escalation procedures should be built into everyday workflows instead of being added after problems occur.

Understanding AI as a socio-technical system shifts the focus from the model alone to the broader environment in which it operates. To assess these systems consistently, organizations need a common way to describe and compare them. The Organization for Economic cooperation and Development (OECD) provides one such approach through its five-dimensional framework for classifying AI systems.

Classifying AI systems using the OECD's five dimensions

If two teams both say, *"We're using AI,"* they may still be talking about systems with very different risk profiles. One might be an internal forecasting model with limited business impact, while another influences hiring or healthcare decisions. Before deciding what

safeguards are needed, organizations need a consistent way to describe the AI system they are assessing.

The Organization for Economic cooperation and Development (OECD) provides one such framework. Instead of focusing on technical complexity, it classifies AI systems across five practical dimensions that help organizations understand how a system operates and where its risks are likely to arise.

The first dimension looks at who is affected by the AI system, directly or indirectly. This may include customers, employees, students, patients, or the public. It also considers scale and vulnerability, since systems that affect vulnerable groups often require greater scrutiny. Starting with the people affected helps identify concerns around privacy, fairness, safety, and human rights early in the governance process.

The second dimension considers the economic context in which the system operates. An AI model used in healthcare, finance, education, or public services carries different expectations and risks than the same model used for internal business planning. Context also helps determine whether the AI is simply supporting human decisions or playing a more direct role in making them.

The third dimension examines the data and inputs the system relies on. This includes structured records, text, images, audio, sensor data, and third-party information. Reviewing data sources helps identify issues such as poor data quality, biased historical records, missing provenance, or the use of sensitive information. Since AI systems learn from data, understanding the data is often the first step in understanding their limitations.

The fourth dimension describes the AI model itself. This may be a statistical model, decision tree, neural network, or large language model. Different model types have different levels of explainability, testing requirements, and operational risks. The goal is not to make every stakeholder a technical expert, but to ensure that governance and assurance activities are appropriate for the technology being used.

The fifth dimension focuses on the system's task and output. Does it classify, score, recommend, forecast, detect anomalies, or generate content? Equally important is how those outputs are used. A recommendation intended to support human judgment can effectively become a decision if people begin accepting it without review. Understanding both the output and its role in the workflow helps determine the level of oversight required.

Together, these five dimensions provide a concise profile of an AI system: who it affects, where it operates, what data it relies on, what type of model it uses, and what it produces. This shared description helps stakeholders communicate consistently, document AI systems clearly, and select governance controls that are appropriate for their intended use.

Classification is not a one-time exercise. AI systems evolve as new users, data sources, and business processes are introduced. Revisiting these five dimensions whenever a system change occurs helps organizations identify new risks and adjust governance before those changes lead to unintended consequences.

Recognizing modern drivers of AI adoption

AI adoption accelerated because several enabling technologies matured at the same time, reducing the cost, effort, and time required to build and deploy AI systems. As these barriers fell, organizations could experiment more easily and move successful projects into production much faster. This section examines the main drivers behind that growth and explains why AI adoption often outpaces governance.

One of the biggest drivers has been **cloud computing**. Instead of investing in large amounts of on-premises hardware, organizations can access computing power and storage on demand. This makes it easier to launch pilot projects, scale successful solutions, and support AI workloads without significant upfront investment.

At the same time, organizations have accumulated far more digital data. Business processes, customer interactions, system logs, and operational records now provide AI systems with large volumes of information that they can learn from. As data has grown in both quantity and variety, AI has expanded beyond structured tables to work with text, images, audio, video, and sensor data. This has enabled applications such as document summarization, visual inspection, speech recognition, and content generation, while also increasing the importance of data quality, privacy, provenance, and representativeness.

Advances in AI models have also made systems far more capable of understanding language, identifying patterns, and generating human-like content. As a result, AI can now draft documents, summarize information, translate languages, answer questions, and support many knowledge-intensive tasks. These capabilities make AI more useful across a wider range of business functions, but they also increase the risk that fluent, confident outputs will be accepted without sufficient verification.

Progress in computing hardware has further accelerated adoption. More powerful processors and specialized AI hardware have made it practical to train and deploy increasingly sophisticated models at scale. Capabilities once limited to research environments are now routinely embedded in business applications, making AI a practical technology for day-to-day operations.

Another important factor is the growing accessibility of AI tools and platforms. Many cloud providers and software vendors now offer pre-built AI services, templates, and integrations that allow organizations to develop AI-enabled applications without building every

component from scratch. This lowers the technical barrier to adoption and enables a much broader range of teams to incorporate AI into their work.

Business pressure has also accelerated AI adoption. As competitors use AI to improve customer experience, reduce costs, and increase efficiency, organizations face increasing pressure to respond quickly. This can shorten delivery timelines and encourage rapid deployment, sometimes before governance processes have fully matured.

These developments have made AI easier to adopt, easier to scale, and easier to integrate into everyday business operations than many previous technologies. While this creates significant opportunities for innovation, it also means that governance must evolve at the same pace. Effective governance should enable organizations to adopt AI confidently while maintaining appropriate controls over data, accountability, and acceptable use.

Next, we'll examine the three broad categories of AI, **Artificial Narrow Intelligence (ANI), Artificial General Intelligence (AGI), and Artificial Superintelligence (ASI)**, and discuss why understanding these distinctions is important for AI governance.

Distinguishing between ANI, AGI, and ASI

Not all AI systems are designed to do the same kind of work. Researchers generally describe AI in three broad categories: Artificial Narrow Intelligence (ANI), Artificial General Intelligence (AGI), and Artificial Superintelligence (ASI). Understanding these categories helps distinguish what AI can do today from what remains a research goal or a theoretical possibility.

Artificial narrow intelligence

Artificial Narrow Intelligence (ANI), sometimes called **Weak AI**, is the only form of AI in widespread use today. It is designed to perform specific tasks, such as recognizing speech, recommending products, detecting fraud, or generating text. Although ANI can outperform humans in narrowly defined tasks, it cannot transfer its capabilities to unrelated problems or reason beyond the objectives for which it was developed.

Everyday examples include voice assistants, recommendation engines, spam filters, facial recognition, and navigation systems. Each performs a well-defined function but cannot operate outside its intended purpose.

The strength of ANI lies in its speed, accuracy, and ability to process large volumes of data. Its limitation is that it lacks general reasoning, self-awareness, and the ability to apply knowledge across different domains.

Artificial general intelligence

Artificial General Intelligence (AGI), often called **Strong AI**, refers to a hypothetical system capable of learning, reasoning, and solving problems across many different domains in the

same way humans can. Rather than requiring separate systems for different tasks, AGI would be able to transfer knowledge from one problem to another and adapt to unfamiliar situations.

An AGI system could, in theory, learn a new language, diagnose diseases, write software, and plan complex projects using the same underlying intelligence.

Despite significant research, AGI does not currently exist. Building systems with human-level reasoning, adaptability, and understanding remains one of the largest challenges in AI research, raising both technical and governance questions.

Artificial superintelligence

Artificial Superintelligence (ASI) describes a theoretical AI system whose capabilities exceed those of humans across virtually every intellectual task. Such a system could outperform experts in scientific discovery, engineering, medicine, creativity, and strategic decision-making.

Because ASI remains hypothetical, discussions focus primarily on its potential opportunities and risks. While it could contribute to solving complex global challenges, it also raises profound questions about control, alignment with human values, and long-term safety.

At present, ASI remains a subject of research and debate rather than existing technology.

Common AI use cases

Most AI systems deployed today are examples of Artificial Narrow Intelligence and are designed for specific business tasks.

Common organizational use cases include:

- **Recognition:** Identifying faces, objects, text, or patterns
- **Event Detection:** Detecting fraud, cyber threats, or anomalies
- **Forecasting:** Predicting demand, sales, or resource needs
- **Personalization:** Tailoring recommendations and user experiences
- **Interaction Support:** Chatbots and virtual assistants
- **Optimization:** Improving schedules, routes, and resource allocation
- **Recommendation Systems:** Suggesting products, content, or actions

Each use case presents different governance challenges. For example, hiring tools influence employment decisions and therefore require much stronger oversight than movie recommendation systems. Regulations such as *New York City's Local Law 144* demonstrate how governance expectations increase as the potential impact on individuals grows.

Understanding how AI systems differ is the first step toward governing them effectively. In the next section, we'll introduce AI governance and examine how organizations establish the principles, processes, and oversight needed to manage AI responsibly throughout its lifecycle.

Introducing AI governance, principles, and frameworks

This section introduces the governance layer that surrounds every AI system. It explains what AI governance is, how it differs from one-time compliance, distinguishes between principles and frameworks, and outlines the elements of a comprehensive governance approach. By the end of this section, you will be able to explain AI governance, distinguish it from compliance, and identify the core components of an effective governance program.

Defining AI governance

AI governance is how an organization establishes policies, assigns responsibilities, and maintains oversight so that AI systems are developed, deployed, and used safely and responsibly. It covers the entire AI lifecycle, including use-case approval, data and model decisions, testing, deployment, monitoring, incident response, and retirement. The goal is not perfection but a governance approach that is consistent, transparent, and easy to explain.

Governance differs from compliance. Compliance demonstrates that specific requirements have been met at a particular point in time, whereas governance provides continuous oversight as AI systems evolve. A system that passes an initial review can still become risky as data changes, models are updated, or new use cases emerge. Governance ensures that oversight continues throughout the system's lifecycle.

The purpose of governance is to reduce avoidable harm. That includes unfair outcomes, privacy breaches, security misuse, operational failures, and reputational damage caused by AI systems behaving in unexpected or poorly understood ways. By establishing clear responsibilities and review processes, organizations can identify issues early rather than responding only after problems become public.

AI governance is not the responsibility of a single team. Product teams understand the business problem and user experience. Engineering teams design, build, and test the system. Privacy, security, legal, risk, and compliance teams provide oversight within their areas of expertise. Governance brings these groups together, so that decisions remain consistent and accountability is clear throughout the AI lifecycle.

In practice, governance means having clear approval processes, documentation standards, monitoring activities, and escalation paths. Teams know what they can deploy, what requires additional review, and what evidence must be documented before systems move into

production. As AI systems evolve, governance ensures that updates are treated as managed changes instead of informal modifications.

Governance also builds trust. Customers trust AI systems when outcomes are consistent and explanations are available. Employees trust AI systems more readily when they feel safe raising concerns about outputs and confident that those concerns will be heard. Regulators and auditors trust them when organizations can demonstrate how decisions were made and how risks are managed. Governance turns responsible AI from an aspiration into a repeatable practice.

Many organizations align their governance programs with established risk frameworks. For example, the **National Institute of Standards and Technology (NIST)** *AI Risk Management Framework (AI RMF) 1.0,* published in January 2023, organizes AI risk management around four functions: Govern, Map, Measure, and Manage. You don't need to memorize these functions, but it's useful to recognize that AI governance is supported by well-established frameworks rather than ad hoc practices.

Distinguishing principles from frameworks

The terms *principles* and *frameworks* are often used interchangeably, but they serve different purposes. Principles describe the values an organization wants AI to reflect, while frameworks provide the processes and controls needed to put those values into practice. Understanding the distinction makes governance easier to communicate and apply.

Principles express the outcomes an organization wants to achieve. Common examples include fairness, privacy, transparency, safety, accountability, and human oversight. They help leaders communicate expectations and guide decision-making when detailed rules do not exist. Principles by themselves don't tell teams what actions to take.

Frameworks translate those principles into repeatable practices. They define the activities, documentation, evidence, and controls needed to manage AI throughout its lifecycle. Review processes, testing requirements, monitoring activities, and approval workflows are all examples of how frameworks make governance operational and consistent across projects.

Transparency provides a useful example. As a principle, transparency means people should know when AI is involved and have a reasonable understanding of how decisions are made. A framework puts that principle into practice by requiring disclosures, documenting system behavior, and defining how explanations are provided to users and stakeholders. The principle explains *why* transparency matters; the framework explains *how* it is achieved.

Neither principles nor frameworks are sufficient on their own. Principles without frameworks remain aspirations, while frameworks without principles become checklists followed without understanding their purpose. Combining the two creates governance that is both meaningful and practical.

One useful approach is to map each principle to concrete governance practices. For example, fairness may require testing across different groups, privacy may require data minimization and access controls, transparency may require documentation and user disclosures, accountability may require clearly assigned ownership and escalation procedures. Making these connections helps organizations apply governance consistently across different AI systems.

Many organizations use the NIST AI RMF 1.0 as a common structure because it provides a shared language for organizing governance activities.

Applying a comprehensive approach to AI governance

A comprehensive approach to AI governance looks beyond the model itself. It considers the entire AI system, including the use case, the data, the model, the workflow, the people involved, and what happens after deployment. Looking at these elements together helps organizations identify risks that may not be visible when each component is considered in isolation.

A strong governance approach begins with the use case. What problem is the AI intended to solve, and what decision will it influence? Clear objectives help define what success looks like, what outcomes are acceptable, and what risks need to be managed. When the use case is poorly defined or keeps changing, testing and oversight become much more difficult.

The next consideration is context. The same AI model can present very different risks depending on how it is used. For example, a summarization tool used to draft internal notes requires a different level of oversight than one used to generate formal customer communications. Context includes who is affected, the potential impact of errors, and whether people can reasonably identify and correct mistakes before they cause harm.

Data is another critical governance consideration. The quality, representativeness, provenance, and sensitivity of the data directly influence how an AI system performs. Biased or incomplete data can produce biased or unreliable outcomes, making data governance just as important as model governance.

Governance should also consider the model itself. Different model types vary in their explainability, predictability, and operational risks. Instead of applying the same controls to every system, organizations should match the level of testing and oversight to the capabilities and risks of the model being used.

Testing should evaluate more than accuracy. AI systems should be assessed under realistic operating conditions, including unusual inputs, edge cases, and situations where failures could have significant consequences. Scenario-based testing often provides a better understanding of operational risk than benchmark scores alone.

A comprehensive approach also examines the workflow surrounding the AI system. Risk often arises not because the model is technically incorrect, but because people rely on its outputs without sufficient review. Governance should therefore define where human oversight is required, how outputs are reviewed, and when decisions should be escalated.

Clear ownership is equally important. Organizations should identify who is responsible for the use case, the model, monitoring, and incident response. Clearly assigning these responsibilities improves accountability and allows issues to be addressed quickly when they arise.

Finally, governance continues after deployment. AI systems evolve as data changes, models are updated, and business needs shift. Ongoing monitoring helps detect issues such as performance degradation, model drift, emerging bias, and operational failures before they become significant problems.

In the next section, we'll examine common AI models and explore the governance questions that each model type raises.

Understanding AI model types and machine learning fundamentals

This section introduces the AI concepts that are most relevant to governance. It explains the relationship between artificial intelligence, machine learning, deep learning, and generative AI, explores common AI model types and machine learning approaches, and examines how training decisions influence system behavior and governance risk. By the end of this section, you will be able to identify the main AI model types and explain how training choices shape both system capability and governance risk.

Distinguishing AI, ML, deep learning, and generative AI

Artificial intelligence (AI), machine learning (ML), deep learning, and generative AI are closely related terms, but they describe different concepts. Understanding how they relate to one another helps you communicate more precisely and apply governance controls that match the capabilities and risks of different AI systems.

Artificial intelligence (AI) is the broad field of building systems that perform tasks typically associated with human intelligence, such as recognizing patterns, making predictions, understanding language, or solving problems. AI includes many approaches, ranging from rule-based systems to machine learning techniques.

Machine learning (ML) is a subset of AI in which systems learn patterns from data instead of relying entirely on manually programmed rules. It is widely used for tasks such as classification, forecasting, recommendation, and anomaly detection. When organizations refer to "training a model," they are usually talking about machine learning.

Deep learning is a specialized form of machine learning that uses multilayered neural networks to learn complex patterns from data. It has driven significant advances in computer vision, speech recognition, and natural language processing. Although these models often deliver strong performance, they are generally more difficult to interpret than simpler machine learning models.

Generative AI refers to systems that create new content, such as text, images, audio, video, or software code. Unlike predictive models that classify or forecast, generative models produce original outputs based on patterns learned during training. These capabilities make them valuable for drafting, summarization, creative work, and conversational applications, while also introducing risks such as hallucinations, misinformation, and inappropriate content generation.

Describing both the technology and the use case helps avoid confusion. For example, "a machine learning model that forecasts demand" provides much more useful information than simply saying "an AI tool." Clear descriptions also make it easier to select appropriate testing, monitoring, and governance controls.

In general, predictive machine learning systems emphasize validation, fairness assessment, and performance monitoring, while generative AI systems often require additional safeguards such as content review, usage restrictions, and controls for sensitive information.

Common AI models

AI systems are built using different types of models, each designed for particular tasks. Model choice influences explainability, performance, maintenance, and the level of governance required. Understanding the major model families helps practitioners ask informed questions without needing deep technical expertise.

Statistical models identify relationships in structured data and are commonly used for prediction and classification tasks. Because their behavior is often easier to explain, they can simplify governance and auditing. However, they remain sensitive to biased data and poorly defined objectives.

Decision tree models make predictions by following a sequence of decision rules. Their structure is generally easier to interpret than more complex approaches, making them useful when explainability is important.

Neural networks excel at recognizing complex patterns in images, speech, language, and other high-dimensional data. Their complexity often makes them more difficult to interpret, so organizations typically rely on thorough testing and monitoring to evaluate their behavior.

Language models process and generate text. They can summarize information, answer questions, draft documents, and support conversational interfaces. Because they may produce

convincing but inaccurate responses, they are commonly deployed with human review processes and usage guidelines.

Generative image, audio, and video models create original media that support creative, educational, and design workflows. These systems also introduce risks such as misinformation, impersonation, and copyright concerns, making content controls and disclosure policies particularly important.

Choosing the right model involves balancing performance, explainability, operational complexity, and business requirements. Instead of selecting the most sophisticated model available, organizations should choose the model that best fits the intended use case and its associated risks.

Regardless of the model type, a few simple questions help guide governance discussions: What data does the model use? What does it produce? How is it tested? How will its performance be monitored after deployment?

Exploring types of machine learning models

Machine learning models learn in different ways depending on the type of data available and the problem they are designed to solve. Understanding these learning approaches helps governance teams ask the right questions during development, testing, and deployment.

Supervised learning uses labeled examples to learn relationships between inputs and known outcomes. It is commonly used for classification and prediction tasks. Because the model learns from human-provided labels, the quality and consistency of those labels directly affect model performance.

Unsupervised learning identifies patterns in data without labeled outcomes. It is often used for clustering similar records or detecting anomalies. Since these patterns are not predefined, they should be interpreted carefully and validated against business objectives.

Reinforcement learning improves through trial and error by receiving rewards or penalties for its actions. It is commonly applied to systems that make sequences of decisions over time. The design of the reward function is particularly important because it determines the behavior the model learns.

Semi-supervised learning combines a small amount of labeled data with a much larger set of unlabeled data. This approach reduces the effort required to create training data while still improving model performance.

Machine learning models are also designed for different prediction tasks. **Classification** predicts categories, such as "fraud" or "not fraud," while **regression** predicts continuous values, such as sales forecasts or risk scores.

Many business applications rely on **decision tree models** and **ensemble methods**, which often provide a useful balance between predictive performance and interpretability. Neural networks, by contrast, are better suited to complex tasks involving language, images, and speech, although they generally require more extensive validation because their internal decision-making is less transparent.

Different learning approaches require different forms of oversight, but the underlying principle remains the same: governance controls should be proportionate to the potential impact and uncertainty of the system.

Training AI systems using machine learning fundamentals

Training is the process through which a machine learning model learns patterns from data. The choices made during training, including the data used, the objectives defined, and the evaluation methods selected, have a direct influence on both model performance and governance risk.

Unlike traditional software, where developers explicitly define every rule, machine learning systems learn behavior from examples. During training, the model repeatedly adjusts its internal parameters to improve its predictions. As a result, model behavior depends not only on the algorithm itself but also on the data and objectives used during training.

The quality of the training data is therefore critical. Incomplete, outdated, biased, or unrepresentative data can produce unreliable or unfair outcomes because the model learns from the patterns present in that data.

Many machine learning models also rely on **labeled data**, where humans identify the correct outcome for each training example. Those labels reflect human judgment and may also reflect human bias. Poor labeling practices can therefore lead to models that repeat inconsistent or unfair decisions.

Training objectives shape model behavior as well. Optimizing only for overall accuracy, for example, may produce poor performance for smaller groups or unusual situations. Selecting appropriate evaluation metrics helps ensure that the model supports both business objectives and responsible outcomes.

Even well-trained models can behave differently once deployed. Real-world data changes, user behavior evolves, and unexpected situations emerge. Continuous monitoring helps identify performance degradation, model drift, and new risks that were not apparent during development.

AI systems also evolve through retraining, fine-tuning, and configuration updates. Governance should treat these changes as controlled events that require review instead of assuming that the system remains unchanged after deployment.

Although machine learning automates prediction, humans remain responsible for selecting the training data, defining objectives, choosing evaluation metrics, approving deployment, and determining where AI may be used. Effective governance ensures these decisions are documented, reviewed, and supported by evidence throughout the AI lifecycle.

In the next section, we'll examine transformer models, multimodal AI systems, expert systems, and explore the governance considerations associated with each.

Exploring transformer models, multimodal systems, and expert systems

This section introduces AI architecture and deployment choices that commonly arise in governance discussions. You'll learn how transformer models power modern language systems, how multimodal models process multiple forms of data, how deployment choices influence governance responsibilities, and why expert systems remain relevant despite predating modern machine learning. By the end of this section, you'll be able to explain how these technologies differ and how those differences influence testing, explainability, and oversight.

Working with transformers, NLP, and multimodal models

Natural language processing (**NLP**), transformer models, and multimodal systems power many of today's AI applications. Although you don't need to understand their internal engineering, knowing what these systems do, and where they commonly fail, helps you apply appropriate governance controls.

Natural language processing (NLP) is the branch of AI that enables computers to understand, analyze, and generate human language. It supports applications such as translation, summarization, sentiment analysis, chatbots, and document analysis. Because organizations process large volumes of text, including emails, contracts, policies, and support tickets, NLP has become a core capability in many AI systems.

Transformers are a neural network architecture that improves language understanding by processing word relationships across an entire sequence simultaneously.

Transformer-based models are widely used for drafting content, extracting information, summarizing documents, classifying text, and answering questions. While these capabilities improve productivity, they also create a risk that plausible but incorrect outputs are accepted without verification.

Multimodal models extend these capabilities by processing more than one type of input, such as text, images, audio, or video. For example, a multimodal system might analyze a photograph, interpret a chart, or answer questions about a document containing both text and

images. These capabilities support applications in customer service, accessibility, compliance, and training, but they also increase the likelihood of processing sensitive information.

Common governance concerns include hallucinations, sensitive data exposure, prompt injection, and inappropriate content generation. Organizations typically address these risks through controls such as restricting sensitive inputs, defining acceptable use, requiring human review for high-impact outputs, and monitoring systems for recurring failures.

Ultimately, transformer and multimodal models should be governed according to how they are used rather than simply because of their underlying technology. Clear usage boundaries and well-designed review processes help organizations benefit from these models while managing their risks.

Comparing model relationships and deployment options

Organizations have many choices when selecting AI models, and those choices influence cost, transparency, operational responsibility, and governance. Understanding the trade-offs helps ensure that the selected model aligns with both business needs and risk tolerance.

Classic predictive models estimate outcomes such as fraud, customer churn, or future demand. Generative models, by contrast, create new content such as text, images, or summaries. While both have valuable applications, generative models often require stronger controls around content quality, review, and appropriate use because their outputs can appear highly convincing even when they are incorrect.

Another important consideration is whether a model is proprietary or open source. Proprietary models are typically accessed through vendor-managed services, making them easier to adopt but providing less visibility into training data, model updates, and internal implementation. Governance therefore focuses on vendor management, contractual obligations, and data handling.

Open-source models provide greater flexibility because organizations can inspect, customize, and deploy them within their own environments. However, this additional control also brings greater responsibility for security, maintenance, evaluation, and ongoing monitoring.

Organizations must also consider model size. Smaller models are often faster, less expensive, and easier to deploy privately, while larger models generally provide stronger performance across a wider range of tasks. Rather than assuming larger models are always better, organizations should select the smallest model that meets their requirements.

Deployment choices extend beyond model size. Language-only models process text, while multimodal models combine text with images, audio, or video. Because multimodal systems frequently process richer and more sensitive information, governance should account for both the type of data being processed and the potential impact of the outputs.

Finally, organizations may choose to use a model as-is or fine-tune it for specific business tasks. Fine-tuning often improves relevance and performance but also increases responsibility for model behavior, particularly when organizational data is used during training.

There is no universally "best" model. Effective governance focuses on selecting the approach that best balances capability, operational complexity, explainability, and risk for the intended use case.

Understanding expert systems

Not all AI systems rely on machine learning. Expert systems represent an earlier approach to AI that captures human expertise as explicit rules rather than learning patterns from data. Although they predate modern AI models, they continue to be used in many organizations, particularly where decisions must follow well-defined policies or regulations.

An expert system consists of two main components: a knowledge base, which stores facts, and decision rules, and an inference engine, which applies those rules to reach conclusions. Because the reasoning process follows predefined logic, expert systems are often easier to explain and audit than machine learning models.

Their effectiveness, however, depends entirely on the quality of the rules they contain. Outdated policies, incomplete knowledge, or biased decision rules can produce consistently incorrect outcomes. Unlike machine learning systems, expert systems do not automatically adapt when conditions change, making regular review and maintenance essential.

Expert systems are commonly used for eligibility assessments, policy enforcement, diagnostic support, and rule-based decision making. These applications benefit from consistency, but they can also encourage overreliance if users assume that rule-based decisions are always correct.

Governance therefore focuses on maintaining accurate rule sets, documenting changes, assigning ownership, reviewing exceptions, and monitoring outcomes to ensure that the system continues to reflect current policies and business needs. Users should also understand when expert-system recommendations can be challenged or overridden.

Although expert systems differ from machine learning, they can still influence important decisions affecting people and organizations. For that reason, they require the same governance principles of accountability, transparency, change management, and ongoing oversight.

In the next section, we'll examine AI platforms and common organizational use cases, and consider how platform choices influence AI governance.

Managing AI platforms, infrastructure, and operational monitoring

This section introduces the platforms, infrastructure, and operational practices that support AI systems in production. It explains how organizations build and deploy AI using modern platforms, explores the technology stack that supports AI applications, examines the infrastructure choices that influence performance and control, and describes the monitoring practices that help detect issues after deployment. By the end of this section, you will be able to identify the key components of an AI technology stack and explain where governance controls should be applied across infrastructure and operations.

Choosing AI platforms for common use cases

Most organizations do not build AI systems entirely from scratch. Instead, they rely on platforms that provide tools for developing, deploying, and managing AI applications. These platforms accelerate adoption, but they also influence where data is processed, how models are managed, and what governance controls are available.

Cloud platforms provide infrastructure for training, deploying, and scaling AI systems without requiring organizations to manage every component themselves. This improves speed and flexibility but also increases dependence on the platform provider. Organizations therefore need to understand where data is processed, how access is controlled, and what responsibilities remain with the vendor.

Many AI capabilities are now embedded into everyday business applications, including document editors, coding assistants, customer service platforms, and productivity tools. While these features can improve efficiency, they also create risks if users submit sensitive information or rely on AI-generated outputs without appropriate review. Clear acceptable-use policies and role-based access controls help reduce these risks.

Platforms also differ in the visibility they provide. Features such as logging, monitoring, version management, and audit trails make it easier to understand how AI systems behave over time and investigate problems when they occur. Organizations should consider these operational capabilities alongside model performance when selecting a platform.

Because cloud services evolve continuously, platform updates can change model behavior, introduce new features, or alter performance. Treating these updates as controlled change events, and documenting significant changes, helps maintain consistent oversight.

Platform selection also affects how data moves between internal systems and external services. Understanding data flows, retention policies, and access controls is essential for protecting sensitive information and meeting regulatory obligations.

Ultimately, platform decisions should be driven by the intended use case. A drafting assistant, for example, requires different controls from an AI system used for hiring or fraud detection. Matching platform capabilities to business risk helps organizations apply governance proportionately.

Before adopting an AI platform, organizations should ask a few practical questions: Where is the platform hosted? Who can access it? What logging and monitoring capabilities are available? How are updates managed? What happens if the platform becomes unavailable? These questions often reveal operational risks before deployment.

Understanding the AI technology stack

The AI technology stack describes the layers that work together to build, deploy, and operate an AI system. Although the model is often the most visible component, effective governance depends on understanding the supporting layers that enable it.

The data layer includes data collection, storage, preparation, and management. Because model quality depends on data quality, governance begins with ensuring that data is accurate, representative, appropriately sourced, and handled securely.

The model layer covers training, fine-tuning, evaluation, and version management. Governance focuses on documenting how models were developed, how they were tested, and what limitations they have before they are deployed.

The deployment layer includes APIs, applications, user interfaces, and access controls that make AI systems available to users. These components influence how people interact with AI and can significantly affect how outputs are interpreted and acted upon.

The operations layer supports monitoring, maintenance, incident response, and ongoing updates after deployment. Since AI systems continue to evolve in production, operational oversight is essential throughout the system's lifecycle.

Security and privacy extend across every layer of the stack. Protecting data, managing access, securing integrations, and maintaining audit logs require coordinated controls rather than isolated technical solutions.

Although the technology stack describes technical components, people and processes remain equally important. Teams collect data, approve deployments, investigate incidents, and manage ongoing operations. Assigning clear ownership across these layers helps ensure accountability throughout the AI lifecycle.

Describing an AI system in terms of its data, model, deployment, operations, and supporting controls provides a shared understanding that simplifies governance, audits, and technical discussions.

Understanding compute infrastructure

Infrastructure provides computing resources that allow AI systems to operate reliably and securely. Decisions about compute, storage, networking, and software influence not only system performance but also privacy, security, operational resilience, and governance.

Compute provides the processing power needed to train and run AI models. Training large models often requires specialized hardware, while production systems require sufficient capacity to deliver consistent performance. Decisions about where processing occurs also influence data exposure and regulatory obligations.

Storage holds training data, models, evaluation results, and operational logs. Organizations should establish clear policies for data retention, access, and deletion so that information is available for investigation without creating unnecessary privacy risks.

Networks transfer data between users, applications, databases, and AI services. Because sensitive information often moves across multiple environments, organizations should understand where data travels, how it is protected, and which jurisdictions may be involved.

Infrastructure can be deployed in cloud, on-premises, or edge environments. Cloud platforms offer scalability, on-premises deployments provide greater organizational control, and edge computing can reduce latency and limit data movement. The most appropriate option depends on the use case, operational requirements, and risk profile.

Supporting software, including orchestration tools, deployment pipelines, model serving, logging, and monitoring, connects these infrastructure components. Effective software engineering practices such as version control and change management improve transparency and simplify governance.

Infrastructure also supports system resilience. Organizations should understand how AI systems will continue operating during outages, failures, or service disruptions and establish fallback procedures where appropriate.

Strong security controls, including encryption, identity management, network segmentation, and continuous monitoring, should be applied consistently across AI infrastructure rather than treated as AI-specific exceptions.

From a governance perspective, the most important questions remain straightforward: Where is data processed? Where is it stored? Who has access? How are changes tracked? Clear answers provide the foundation for effective oversight.

Implementing observability and monitoring

Governance continues long after an AI system has been deployed. As data changes, users adopt new workflows, and business requirements evolve, monitoring helps organizations identify emerging risks before they become significant incidents.

Observability refers to the ability to understand how an AI system is behaving through logs, metrics, alerts, and operational signals. Without sufficient visibility, organizations cannot detect failures or evaluate whether systems continue to perform as intended.

Monitoring begins with **data integrity**. Changes in data formats, missing fields, or unexpected inputs can quietly reduce model performance long before users notice a problem. Routine validation of incoming data helps detect these issues early.

Another important consideration is **data drift**, where production data gradually differs from the data used during training. Drift is a natural consequence of changing customer behavior, business processes, and operating environments, but unmanaged drift can reduce model reliability and require retraining or recalibration.

Organizations should also monitor model performance, including accuracy, error rates, stability, and behavior in important edge cases. Monitoring metrics should reflect how the model is actually used instead of relying solely on technical benchmarks.

Monitoring should also consider fairness. Performance may vary across different groups or situations over time, even if initial testing showed acceptable results. Regular evaluation helps identify unintended disparities before they become significant.

Technical metrics alone are not sufficient. User feedback, complaint trends, manual overrides, and escalation rates often provide early signals that a system is no longer supporting business needs effectively.

When monitoring identifies significant issues, organizations should have a clearly defined incident response process that assigns ownership, documents investigations, and tracks corrective actions. Reviewing incidents helps strengthen governance and reduces the likelihood of similar problems in the future.

Monitoring is effective only when ownership is clear. Someone must review alerts, investigate anomalies, and decide when intervention is required. Connecting monitoring activities to defined responsibilities ensures that operational issues are addressed consistently.

Summary

In this chapter, we established the foundations of AI governance by introducing the core concepts, technologies, and governance practices that underpin modern AI systems. We began by defining artificial intelligence, tracing its evolution, and examining AI as a socio-technical system shaped by technology, people, and organizational processes. We also explored the OECD's AI classification framework, the factors driving AI adoption, the different categories of AI, and the most common organizational use cases.

Next, we introduced AI governance as the discipline for managing AI responsibly throughout its lifecycle. We distinguished governance principles from governance frameworks and examined how effective governance extends beyond the model itself to encompass data, deployment, monitoring, ownership, and organizational processes.

We then explored the technical foundations of AI, including the relationships between artificial intelligence, machine learning, deep learning, and generative AI. We examined common AI model types, machine learning approaches, training fundamentals, transformer and multimodal models, deployment options, and expert systems, highlighting how different technologies present different governance considerations.

Finally, we looked at the operational environment in which AI systems run. We explored AI platforms, the AI technology stack, infrastructure components, and the monitoring and observability practices that help organizations maintain reliable, secure, and accountable AI systems after deployment.

In *Chapter 2, AI Impacts on People and Responsible AI Principles*, we'll examine how AI systems affect individuals and society, explore the ethical challenges associated with AI deployment, and introduce the responsible AI principles that guide governance decisions in practice.

2

AI Impacts on People and Responsible AI Principles

Chapter 1, Foundations of Artificial Intelligence and AI Governance, introduced the foundations of AI governance by explaining what AI is, how AI systems are classified and trained, and how governance provides accountability throughout an AI system's lifecycle. Building on that foundation, this chapter explores the risks, harms, and ethical considerations that make AI governance essential.

You will learn about the harms AI can create and who they affect, the security, privacy, and business risks organizations must manage, and the ethical principles and frameworks that guide responsible AI. The chapter concludes by exploring the controls, organizational culture, and operational practices that help organizations embed responsible AI and sustain it over time.

In this chapter, we will cover the following topics:

- Understanding AI harms and their impact on people
- Assessing organizational risk from AI
- Applying ethical principles and frameworks
- Building and operationalizing responsible AI

Understanding AI harms and their impact on people

This section explains how to identify and categorize AI harms using a structured approach, examines who is affected and why some groups are more vulnerable than others, explores bias, and discrimination as common sources of harm, and assesses the organizational consequences of AI-related harms. By the end of this section, you will be able to identify AI harms, describe

who bears them, and explain how bias and discrimination connect to governance decisions made before and after deployment.

How to identify harm

Identifying harm means predicting how an AI system could negatively affect people, communities, and organizations before those outcomes occur. AI governance looks beyond obvious failures, such as security breaches, to consider less visible harms, including unfair treatment, loss of autonomy, privacy intrusion, and people being misled by overconfident outputs. The goal is to prevent harm before it occurs.

A common mistake is to start with the algorithm and ask what could go wrong technically. A stronger approach begins with the context. First, define the system's purpose, who it affects, and the decisions or influence it creates. For example, a generative AI assistant that drafts emails, a model that scores loan applicants, and a tool that monitors employee productivity can all create different types of harm, even if they use similar technology. Understanding the context makes harm identification more realistic and aligned with how the system will be used.

A repeatable process makes harm identification more consistent. First, map the AI lifecycle stage, such as data collection, training, deployment, or monitoring. Next, identify the people who may be affected, including customers, employees, children, or other vulnerable groups. Then consider what could go wrong across core harm categories such as privacy, fairness, safety, security, transparency, and operational reliability. Finally, connect each identified harm to its potential impact, ownership, and appropriate controls so responsibility is clearly assigned.

Imagine a hospital deploying an AI system to prioritize which patients receive follow-up calls. Harm identification would look beyond model accuracy. It would also consider whether biased historical data could deprioritize certain patient groups, whether sensitive health information could be exposed through vendor access, and whether staff might become over-reliant on the system and miss urgent cases. This example shows why AI harms are socio-technical: the way people interact with the system can either reduce or amplify risk.

Effective harm identification is supported by evidence. That evidence may include stakeholder feedback, user research, model testing results, privacy reviews, and records of the decisions made and why they were made. Regulators and auditors increasingly expect organizations to demonstrate that they have carried out structured, risk-based assessments instead of relying on broad statements about responsible AI. Documentation makes harm identification transparent and defensible.

Mapping AI harms using a taxonomy approach

Once potential harms have been identified, the next step is to classify them consistently. A taxonomy provides a shared language that helps different teams assess risks, select controls, and document decisions using the same framework.

Privacy harms occur when personal data is collected, used, shared, or inferred in ways that people do not expect, have not consented to, or cannot reasonably control. AI systems can create harm even without exposing personal data directly because models may infer sensitive information or use existing data in unexpected ways. The consequences can include legal liability, reputational damage, financial loss, stigma, and unfair treatment.

A privacy harms taxonomy helps organizations move beyond vague concerns such as "this feels risky" to clearly defined risks. For example, a system might infer a person's health condition from shopping behavior or use information collected for one purpose to make decisions unrelated to the original purpose. Classifying harms in this way supports consistent risk assessment, control selection, and documentation.

Common privacy harms include collecting more personal data than necessary, using data beyond its original purpose, inferring sensitive information from seemingly ordinary data, and exposing personal information through AI-generated outputs or explanations.

For example, a retail company deploys a generative AI assistant trained on historical customer service chats. Those conversations contain addresses, purchase histories, and occasionally medical information shared by customers. Applying a privacy taxonomy highlights risks such as retaining sensitive information longer than necessary or reproducing personal details in AI-generated responses. These risks can then be mapped to controls such as data minimization, redaction, access controls, output filtering, and incident response procedures.

Managing privacy harms begins with clear purpose limitation and data minimization. Organizations should also understand where training data came from, whether it was collected for an appropriate purpose, who can access it, and whether the model has been tested for memorization or unintended information disclosure. A structured taxonomy helps organizations identify privacy risks consistently, apply appropriate controls, and demonstrate how those decisions were made.

Who is affected by AI harms?

AI harms are not limited to the people who directly interact with a system. They can also affect individuals who never use the AI, as well as the organizations responsible for deploying it. Understanding who may be affected is essential when assessing the potential impact of an AI system.

Some people interact directly with AI, such as employees using a generative AI assistant or customers using an AI chatbot. Others are affected indirectly, including job applicants screened by automated systems or citizens whose access to public services is influenced by algorithmic decisions. There are also bystanders whose personal information may be captured incidentally, such as family members mentioned in conversations or individuals appearing in images used to train models.

The level of harm often depends on a person's ability to understand, question, or challenge the system. Children, older adults, patients, migrants, and other vulnerable groups may experience greater consequences because they have fewer opportunities to appeal decisions or reduce their exposure to risk. Vulnerability and power imbalance should therefore be considered during every impact assessment.

Organizations can also be affected. AI-related harms may lead to regulatory investigations, legal action, reputational damage, customer churn, and operational disruption. A biased lending model, for example, may disadvantage applicants while simultaneously exposing the bank to enforcement action and financial penalties.

A practical approach is to map the AI value chain by identifying who provides the data, who develops the model, who deploys it, who relies on its outputs, and who is ultimately affected by its decisions. This also helps identify second-order effects, such as changes in behavior, incentives, or levels of surveillance that may not be immediately obvious.

Understanding who is affected helps organizations design appropriate controls, transparency measures, and avenues for appeal that reflect the real impact of AI systems on both people and organizations.

Understanding bias and discrimination in AI

Bias and discrimination are among the most widely discussed AI harms because they can directly affect people's opportunities, rights, and access to services. Bias occurs when an AI system produces systematically unfair outcomes for certain individuals or groups. While bias can arise from technical issues, it is often rooted in the data, assumptions, or processes used to design and deploy the system.

Bias can enter an AI system at multiple stages. Historical data may reflect existing inequalities, training data may not represent all groups equally, or developers may unintentionally introduce assumptions into the model or its design. Bias can also emerge after deployment if user behavior or real-world conditions change. Understanding where bias originates is the first step toward reducing its impact.

Consider an AI system used to screen job applicants. If it is trained on historical hiring data from an organization that has consistently underrepresented certain groups, the model may learn to favor the same patterns, even if protected characteristics are removed from the

training data. This illustrates why fairness depends not only on the algorithm but also on the quality and representativeness of the data used to train it.

Managing bias requires more than testing a model once before deployment. Organizations should evaluate performance across different groups, examine whether outcomes are consistently fair, and investigate unexpected differences. Bias testing should be supported by human review, particularly where AI influences decisions about employment, lending, healthcare, education, or other high-impact areas.

Fairness also depends on context. A model that performs well overall may still disadvantage a particular group if error rates differ significantly. Organizations therefore need to define what fairness means for a specific use case and select appropriate measures to assess it. Documentation of these decisions helps demonstrate that fairness has been considered systematically rather than assumed.

Bias cannot always be eliminated completely, but it can be identified, measured, and reduced through appropriate data practices, testing, monitoring, and human oversight. Recognizing bias as an ongoing governance challenge helps organizations improve AI systems throughout their lifecycle and avoids treating fairness as a one-time exercise.

Assessing organizational harm from AI

AI harms affect more than individuals. They can also disrupt organizations by damaging trust, increasing costs, creating legal exposure, and affecting long-term business performance. Understanding these organizational harms helps leaders appreciate why AI governance is a business responsibility as well as a technical one.

Reputational harm is often one of the most immediate consequences of AI failure. Biased decisions, unsafe outputs, privacy breaches, or inaccurate information can quickly attract public attention and reduce customer confidence. Rebuilding trust is often far more difficult than preventing the incident in the first place.

Legal and regulatory harm can arise when AI systems fail to meet applicable laws, contractual obligations, or industry standards. Organizations may face investigations, financial penalties, litigation, or increased regulatory scrutiny if AI systems produce unlawful or discriminatory outcomes. Strong governance, documentation, and oversight help demonstrate that risks have been identified and managed appropriately.

Operational harm occurs when AI systems disrupt normal business activities. Model failures, inaccurate predictions, unreliable automation, or excessive reliance on AI can reduce productivity, interrupt services, and create additional work for employees. Effective monitoring, clear escalation procedures, and human oversight help organizations respond before these issues affect business operations.

AI can also create strategic harm. Poor governance decisions may lead to investments in systems that cannot adapt to changing regulations, business needs, or customer expectations. Organizations that overlook accountability, transparency, or long-term maintainability may find it difficult to scale AI responsibly or demonstrate compliance as governance requirements evolve.

These harms are often interconnected. A privacy incident can lead to regulatory action, damage customer trust, and disrupt operations at the same time. Looking at organizational harms collectively helps decision-makers understand the wider consequences of AI failures rather than treating each risk in isolation.

Understanding organizational harms reinforces why AI governance extends beyond technical performance. Effective governance protects people, supports regulatory compliance, strengthens organizational resilience, and helps ensure that AI systems continue to deliver value over time. The harms covered in this section explain what can go wrong for people and organizations. The next section examines the specific risks those harms create across security, privacy, and business operations.

Assessing organizational risk from AI

This section examines the risks organizations face when deploying AI systems. It explores the security, operational, privacy, and business risks associated with AI, along with broader AI-specific risks such as hallucinations, over-reliance, and adversarial attacks. By the end of this section, you will be able to identify the main categories of AI risk and explain why AI requires governance controls that extend beyond those used for traditional software risk management.

Managing security and operational risks

AI systems introduce security and operational risks that resemble those found in traditional software but behave differently in practice. They are probabilistic, depend on evolving data, and can be influenced by user inputs in unpredictable ways. For example, a generative AI tool can be manipulated through prompt injection, while a machine learning model can silently degrade over time as real-world conditions change. As a result, organizations need controls that account for continuous change not just static testing.

Key security risks include data leakage, unauthorized access to models or training data, and attacks that manipulate model behavior. A chatbot might expose sensitive information because it was trained on internal documents or because users successfully bypass its safeguards. Supply chain risks also arise when third-party models or open-source libraries introduce vulnerabilities. Mitigating these risks requires strong access controls, secure configuration, vendor due diligence, red teaming, and robust data handling practices.

Operational risks include instability, degraded performance, and model drift. Drift occurs when the real world changes and the model's assumptions no longer hold, for example, when fraud patterns evolve or customer language changes. Without ongoing monitoring, these issues may remain unnoticed until they begin affecting business outcomes. Organizations should monitor key performance indicators, define thresholds for retraining or rollback, and assign clear ownership for continuous oversight.

Many operational failures stem from how people use AI rather than from technical faults. Employees may over-trust outputs, fail to escalate anomalies, or use AI tools outside approved purposes. For example, a sales team might paste customer information into an unauthorized external AI tool to save time, creating both privacy and security risks. Training, role-based access controls, and clear acceptable-use policies help reduce these risks by reinforcing appropriate behavior.

Organizations increasingly need AI-specific security testing and operational resilience planning, particularly for systems that affect safety, rights, or access to essential services. Documented threat models, monitoring plans, incident response playbooks, and evidence that controls have been implemented help demonstrate that security and operational risks are being actively managed rather than assumed.

Addressing privacy risks in AI systems

Privacy harms describe the negative outcomes people experience, such as embarrassment, discrimination, or loss of control over personal information. Privacy risks are the conditions that make those harms possible, including collecting excessive data, using it for new purposes, or exposing it through AI outputs. Focusing on risk pathways helps organizations design controls before harm occurs.

One common privacy risk begins during data collection, when teams gather more personal information than the use case requires because it may prove useful later. AI projects are especially susceptible to this because additional data often appears to improve model performance. Purpose definition, data minimization, and documented justification for each category of personal data help limit unnecessary collection.

Privacy risk also increases when data collected for one purpose is reused for training, fine-tuning, or evaluation without appropriate authorization. Even if the original collection was lawful, reuse can conflict with user expectations and increase regulatory exposure. Data lineage, documented approvals, and retention policies help ensure that datasets are not repurposed without proper review.

AI systems can also infer sensitive information from ordinary data. Purchase history, location patterns, or behavioral signals may reveal health conditions, financial circumstances, or other

personal characteristics. Privacy assessments should therefore consider not only the information that is collected but also what the model can reasonably infer.

Generative AI introduces additional disclosure risks. Models may reproduce personal information learned during training or retrieve private content through retrieval-augmented generation. Organizations can reduce these risks through redaction, carefully scoped retrieval, role-based access controls, logging, leakage testing, and clear rules governing the information used to ground responses.

Many AI deployments rely on external providers for hosting, model development, or data labeling. Privacy risk increases when vendor contracts do not clearly define data ownership, retention requirements, model training restrictions, audit rights, and breach obligations. Vendor management should therefore be treated as part of privacy governance instead of an administrative activity.

Privacy governance should be supported by evidence, including documented data inventories, approvals for training data, access logs, retention schedules, and testing for leakage and misuse. These records demonstrate that privacy controls have been implemented and maintained throughout the AI lifecycle.

Understanding business risks from AI

Business risk is the possibility that an AI system affects revenue, increases costs, disrupts operations, damages trust, or prevents an organization from achieving its strategic objectives. Many business risks originate from unmanaged ethical, privacy, security, or reliability issues, making governance an essential part of business resilience.

AI systems introduce costs beyond initial development. Compute resources, monitoring, human review, red teaming, incident response, and regulatory compliance all require ongoing investment. Organizations that budget only for building a prototype often underestimate the long-term resources needed to operate AI responsibly.

Customer-facing AI systems also create reputational risk. A single harmful output, such as incorrect advice, biased responses, or disclosure of private information, can spread rapidly and undermine public trust. Organizations can reduce this risk through content safety controls, approval processes for high-impact systems, and well-defined incident response procedures.

Where AI influences decisions about employment, lending, insurance, education, or access to services, legal and regulatory exposure increases significantly. Weak governance can lead to investigations, enforcement actions, and litigation, making transparency, documentation, and human oversight important business safeguards.

Business risk also increases when AI systems degrade over time. As data changes or user behavior evolves, model performance may decline without obvious warning signs. Monitoring

key performance metrics, establishing retraining thresholds, maintaining rollback plans, and defining escalation procedures help organizations respond before service quality deteriorates.

Strategic risks should also be considered. Organizations may become dependent on technologies that cannot meet future compliance requirements, generate sufficient evidence for audits, or support changing business needs. Selecting architectures that prioritize accountability, portability, and traceability helps reduce these long-term risks.

Business risk extends beyond financial impact. Responsible AI governance helps organizations protect trust, maintain compliance, support operational resilience, and ensure that AI investments continue to deliver value over time.

Recognizing other emerging AI risks

Privacy and fairness are central concerns in AI governance, but they do not represent the full AI risk landscape. AI systems can also introduce misinformation, manipulation, safety issues, intellectual property disputes, and adversarial attacks. Governance should account for these broader risks alongside more familiar concerns.

Generative AI can produce responses that are fluent but inaccurate, outdated, or entirely fabricated. These risks become more significant when outputs influence legal, medical, financial, or policy decisions. Organizations can reduce them by defining appropriate use cases, requiring verification for high-impact outputs, and limiting systems to approved information sources.

Over-reliance is another common risk. Under time pressure, users may accept AI outputs without sufficient verification, allowing errors to spread more quickly. Training, interface design that communicates uncertainty, and mandatory human review for high-impact decisions help reduce this risk.

AI systems can also influence behavior through personalization, ranking, and conversational persuasion. Even when outputs are technically accurate, these techniques may reduce user autonomy or disproportionately affect vulnerable individuals. Reviewing user experience patterns, providing transparency, and limiting persuasive techniques in sensitive contexts help address these concerns.

Intellectual property risks arise when models are trained on content with unclear permissions or when generated outputs closely resemble protected works. Clear data sourcing policies, contractual safeguards, and documented records of training data sources help reduce legal uncertainty.

When AI supports physical processes or critical services, reliability becomes a safety issue. Applications such as healthcare, industrial automation, and public services require stronger

testing, continuous monitoring, fallback procedures, and effective human override mechanisms.

AI systems may also be deliberately attacked or misused through prompt injection, data poisoning, or attempts to generate harmful content. Threat modeling, red teaming, secure configuration, and AI-specific incident response procedures help organizations anticipate and manage these risks.

Together, these emerging risks show why AI governance extends beyond privacy and fairness alone. Organizations need controls that anticipate misuse, support responsible deployment, and maintain trust as AI systems evolve. The next section introduces the ethical principles and frameworks that guide those governance decisions.

Applying ethical principles and frameworks

This section introduces the principles and frameworks that guide responsible AI governance. It explains how the **Fair Information Practice** Principles support responsible data use, explores the OECD AI Principles as a widely adopted governance framework, and examines the ethical issues of fairness, transparency, accountability, and safety that shape governance decisions. By the end of this section, you will be able to explain the main AI ethics frameworks and apply their principles to practical governance scenarios.

Applying fair information practice principles to AI

Fair Information Practices are foundational privacy principles that shape modern expectations for responsible data use. Because AI systems rely heavily on data, these principles remain relevant even when the focus shifts from traditional databases to model behavior. They provide a practical foundation for collecting, using, and protecting personal information throughout the AI lifecycle.

At their core, Fair Information Practices are about ensuring that people understand what happens to their information and that organizations do not use personal data in unexpected or harmful ways. This means collecting only the data that is needed, using it for clear purposes, protecting it appropriately, and giving people reasonable ways to correct errors or raise concerns. AI makes these principles more challenging because models can infer new information from existing data and reproduce it through their outputs.

Transparency means more than notifying people that AI is being used. Organizations should also be able to explain the system's purpose, its limitations, the choices made during training, and the controls used to manage it. In practice, transparency is supported through disclosures, documentation, and records that explain how important decisions were made.

Purpose limitation means data should not be reused for a new objective without appropriate justification and authorization. Data minimization means collecting and retaining only the information needed for the intended purpose. When developing AI systems, these principles guide decisions about what belongs in training data, how long that data should be retained, and whether it can be reused for future models.

Fair Information Practices also recognize that people should not be subject to decisions they cannot question or challenge. For higher-impact AI systems, organizations should establish appeal processes, correction mechanisms, and appropriate human review. Even where regulations do not explicitly require these safeguards, they help build trust and reduce complaints.

Security safeguards include measures such as access controls, encryption, monitoring, and incident response. Accountability requires clear ownership and evidence that these controls are operating as intended. Increasingly, organizations are expected to demonstrate this through logs, assessments, and audit trails instead of relying on policy statements alone.

Imagine a university uses AI to detect academic misconduct. Fair Information Practices would encourage clear notice to students, careful limits on the data used, protection of sensitive records, and a fair appeal process for disputed cases. These principles also help ensure that the system does not gradually expand into broader surveillance and that important decisions remain reviewable instead of being fully automated.

Fair Information Practices provide the privacy and accountability foundation for responsible AI. They guide how personal data is collected, reused, protected, and explained, while supporting review and redress mechanisms that reduce risks for both individuals and organizations.

Working with OECD AI principles

The Organization for Economic cooperation and Development (OECD) AI Principles are among the most widely adopted international frameworks for trustworthy AI. They provide a shared foundation for governments and organizations by promoting AI that benefits society while respecting human rights and democratic values. The principles have also influenced how many jurisdictions define AI systems and approach governance across the AI lifecycle.

The OECD AI Principles focus on five broad areas: inclusive growth and well-being; respect for human rights, democratic values, fairness, and privacy; transparency and explainability; robustness, security, and safety; and accountability. Together, these principles describe what trustworthy AI should achieve. Organizations then translate them into policies, controls, and governance practices that support responsible AI throughout the system lifecycle.

Inclusive growth and well-being encourage organizations to look beyond efficiency and consider whether AI delivers meaningful value to people and society. In practice, this means

defining clear objectives, measuring outcomes, and avoiding deployments that shift costs or risks onto individuals or communities.

Human rights and democratic values emphasize that AI should respect dignity, protect privacy, avoid discrimination, and support meaningful human oversight. Organizations put these principles into practice through impact assessments, bias testing, privacy reviews, stakeholder engagement, and escalation processes for rights-related concerns.

Transparency and explainability require more than publishing a statement about AI use. Organizations should be able to explain the role of AI, document important design decisions, and provide explanations appropriate to different audiences. Customers may need simple, plain-language explanations, while auditors may require model documentation, logs, and decision records.

Robustness, security, and safety focus on ensuring that AI performs reliably under real-world conditions, not just during testing. Organizations support this through continuous monitoring, testing for misuse and adversarial attacks, secure configuration, and incident response processes that help maintain reliable operation over time.

Accountability means that responsibility for AI decisions always remains with identifiable people and organizations. Clear ownership, documented approvals, audit trails, and governance structures ensure that accountability does not disappear behind statements such as "the model made the decision."

The OECD also encourages governments and organizations to invest in AI literacy, workforce development, international cooperation, and governance environments that support responsible innovation. For organizations, this highlights the importance of building AI capability alongside technical controls so responsible AI can be sustained over time.

The OECD AI Principles provide a practical framework for trustworthy AI built on well-being, human rights, transparency, robustness, and accountability. Applying these principles means translating them into operational controls, evidence, and oversight that support responsible AI in practice.

Identifying key ethical issues in AI

Ethical issues arise when AI systems create harm, unfairness, or reduce human control, even if they comply with legal requirements and perform well technically. Ethics is concerned with how AI affects people, how decisions are made, and whether those decisions can be justified in practice.

Legal requirements define the minimum organizations must do. Ethics goes further by asking what organizations should do to prevent harm, build trust, and maintain legitimacy. Policies

bridge the gap by translating legal obligations and ethical expectations into consistent rules, processes, and accountability.

One of the most significant ethical concerns is **fairness**. AI systems used in hiring, lending, insurance, education, or healthcare can create unequal outcomes for different groups. Organizations should therefore evaluate disparate impact alongside overall performance and consider whether a decision is appropriate for automation in the first place.

AI can also affect privacy and human dignity. Technologies introduced to improve efficiency may increase surveillance or create intrusive forms of profiling. Workplace monitoring tools, for example, can discourage normal behavior, while customer profiling may feel invasive even when technically permitted. Applying principles such as proportionality, purpose limitation, and transparency helps reduce these risks.

Transparency and explainability are equally important. People should know when they are interacting with AI and receive explanations that are appropriate to the context. Risks arise when organizations present AI outputs with unwarranted certainty or fail to disclose the system's role in decision-making. Clear disclosures, appropriate explanations, and user education help address these concerns.

Accountability requires clearly assigned responsibility throughout the AI lifecycle. When harm occurs, organizations should know who is responsible for prevention, monitoring, remediation, and decision-making. Clearly defined ownership and escalation processes help ensure that issues are addressed promptly rather than falling into organizational gaps.

Ethics also includes **safety and security**. Systems that fail unpredictably or are easily manipulated can create harm even without malicious intent. Organizations should therefore test for misuse, monitor performance over time, and maintain incident response processes that support safe operation.

Imagine a clinic using generative AI to draft patient messages. The system occasionally produces convincing but incorrect medical guidance. The ethical concern is not simply that errors occur, but that patients may trust the advice because it appears authoritative. Restricting the tool to drafting, requiring clinician review, and monitoring for unsafe outputs help reduce this risk.

Fairness, privacy, transparency, accountability, safety, and over-reliance are among the most important ethical issues in AI. Addressing them requires more than good intentions. Organizations need practical controls, documented decisions, and ongoing oversight to ensure responsible AI remains consistent over time.

Principles and frameworks define what responsible AI should look like. The next section builds on these foundations by exploring the controls, organizational culture, and operational practices that make responsible AI part of everyday governance.

Building and operationalizing responsible AI

This section focuses on putting responsible AI into practice. It explores the foundational controls that reduce ethical risk, the cultural conditions that influence whether those controls are followed, the characteristics of trustworthy AI, and the operational practices that help organizations sustain responsible AI over time. By the end of this section, you will be able to describe the controls, cultural conditions, and operational practices that make responsible AI a sustainable organizational capability.

Mitigating ethical risks with foundational controls

Foundational controls are the core governance mechanisms that reduce ethical risk across a wide range of AI use cases. Rather than one-time fixes, they are repeatable practices that help make responsible AI the default, even when teams are working under tight deadlines.

The first step is to define the system's purpose, intended users, and the decisions it will influence. From there, organizations assess the likelihood and severity of potential harm to determine the appropriate level of oversight. Higher-risk systems typically require stronger controls, including more rigorous approvals, deeper testing, and closer monitoring.

Impact assessments help teams identify who may be affected, how harm could occur, and what safeguards are needed. They also create documented evidence of the decisions made, supporting transparency and accountability throughout the AI lifecycle.

Human oversight should be meaningful rather than symbolic. People responsible for reviewing AI outputs need the authority, time, and information to question the system when necessary. Organizations should define when human review is required, what reviewers should assess, and how those decisions are documented, particularly for high-impact outcomes.

Transparency controls include informing people when AI is being used, explaining its purpose, and limitations, and maintaining internal documentation on training data, known limitations, and expected failure modes. This helps ensure that both users and teams understand what the system can and cannot do.

Testing should go beyond measuring model performance. Organizations should evaluate behavior across different groups, test for unsafe or unexpected outputs, and assess how the system performs under stress or misuse. Clear thresholds should define when deployment should pause, additional safeguards should be introduced, or retraining is required.

Even well-tested systems can become unreliable after deployment as data changes, user behavior evolves, or new attack methods emerge. Continuous monitoring, logging, and incident response procedures help ensure that ethical risk management remains active throughout the system's lifecycle.

Imagine a bank deploying an AI model to recommend credit limits. Foundational controls would include defining the limits of the model's role, testing for fairness, requiring human review of edge cases, documenting the reasons for denied applications, monitoring outcomes over time, and providing a process for handling complaints or identifying bias.

Foundational controls reduce ethical risk through clear scoping, risk classification, impact assessment, meaningful human oversight, transparency, rigorous testing, and continuous monitoring. Together, these practices create consistency, accountability, and evidence that governance is working in practice.

Building a culture of ethical AI

Policies and technical controls are only effective when people follow them. Organizational culture influences whether employees raise concerns, whether leaders support responsible decisions, and whether delivery pressure outweighs ethical considerations. A strong culture reinforces responsible behavior across the organization.

Leadership sets the tone by making responsible AI part of organizational performance, not an afterthought. This includes supporting decisions to delay or stop deployments when risks are unacceptable, investing in governance activities, and empowering reviewers to challenge high-risk decisions.

Culture is also shaped by incentives. Teams measured only on speed or cost reduction are more likely to overlook ethical concerns. Rewarding safe deployment, reduced incidents, and trustworthy outcomes encourages responsible behavior to become part of everyday work. Organizations can reinforce this by defining metrics that recognize responsible AI practices alongside business performance.

Training should prepare employees for real situations, not just definitions. People need to understand model limitations, privacy risks, bias, prompt injection, and other common failure modes so that they know how to respond when problems arise.

An ethical culture also depends on psychological safety. Employees should have clear routes for escalating concerns through ethics review boards, risk committees, or similar governance structures, without fear of retaliation. They should also know when to pause system use and how to report unexpected behavior.

Responsible AI depends on collaboration across legal, privacy, security, risk, engineering, and product teams. Governance processes such as intake reviews and stage gates work best when these groups engage early, rather than passing issues between teams late in the development process.

Imagine a product team preparing to launch a generative AI feature that summarizes customer complaints. In a strong ethical culture, someone would naturally ask whether the summaries

could expose sensitive information, misrepresent customer concerns, or require human review before being shared with executives. Culture turns these questions into routine practice rather than exceptional concerns.

Building a culture of ethical AI requires leadership commitment, aligned incentives, practical training, safe escalation, and cross-functional collaboration. These elements make responsible behavior sustainable, particularly when teams face competing priorities.

Defining and achieving trustworthy AI

Trustworthy AI performs reliably, respects people's rights, and operates within clear lines of accountability. Trust is built through evidence rather than claims, including documented controls, monitoring results, and consistent governance practices.

Trustworthy systems protect privacy, reduce discrimination, and support meaningful human oversight. Organizations achieve this through impact assessments, privacy reviews, fairness testing, and clear mechanisms for appeal and redress. The objective is not simply high performance but minimizing harm while delivering value.

Transparency is another defining characteristic. People should understand when AI is being used, what it is intended to do, and what its limitations are. Organizations also need internal traceability through documentation, decision records, and audit logs that explain how the system was developed and operated.

Trustworthy AI should perform reliably under real-world conditions and remain resilient to misuse or attack. Threat modeling, secure configuration, red teaming, monitoring for drift, and incident response planning all contribute to maintaining reliable operation, particularly in high-impact applications.

Accountability requires clearly identified owners who are responsible for deployment, monitoring, and remediation. Responsibilities should remain with the organization even when third-party vendors or external models are involved.

Trust is achieved by embedding these principles throughout the AI lifecycle. Risk classification, stage gates, documentation, testing, monitoring, and escalation processes work together to translate trustworthy AI from an aspiration into everyday practice.

Consider an organization using AI to support employee performance reviews. Building trust requires fairness testing across groups, transparency about how AI contributes to decisions, limits on employee surveillance, meaningful opportunities to challenge outcomes, and ongoing monitoring to detect drift or misuse.

Trustworthy AI combines respect for rights, transparency, robustness, and accountability. Organizations achieve these outcomes by embedding them into repeatable governance processes supported by evidence and continuous improvement.

Operationalizing responsible AI practices

Operationalizing responsible AI means embedding governance into everyday work, not treating it as a separate exercise. Teams should routinely consider risk, documentation, oversight, and monitoring throughout the development and deployment of AI systems.

Responsible AI should be integrated across the entire AI lifecycle, from intake through retirement. During intake, teams define the system's purpose and risk level. Development focuses on data governance, testing, and validation. Deployment emphasizes monitoring and incident readiness, while retirement includes decommissioning systems and applying data retention requirements. Stage gates help ensure that each phase is completed before the next begins.

Clear roles and responsibilities are equally important. Product leaders decide when automation is appropriate, engineers understand testing requirements, legal and privacy teams define approval evidence, and support teams know how to recognize and report harmful outputs. Clearly documented responsibilities reduce ambiguity and improve decision-making.

Responsible AI also depends on standard artifacts such as model documentation, data inventories, risk assessments, test reports, monitoring dashboards, and incident logs. Practical mechanisms including structured intake forms, standardized testing procedures, and monitoring dashboards make responsible behavior easier to sustain under delivery pressure.

Training is an ongoing operational requirement. Teams should understand prompt injection, data leakage, model drift, over-reliance, and other common AI failure modes. They also need practical guidance on handling sensitive information, using AI tools appropriately, and documenting incidents. AI literacy has become a core operational capability alongside cybersecurity awareness.

Monitoring extends beyond technical performance. Organizations should also monitor fairness outcomes, complaint volumes, unsafe outputs, drift indicators, and patterns of user behavior that suggest over-reliance or misuse. Clear thresholds should define when intervention is required.

Incidents are inevitable, so organizations need established reporting channels, containment procedures, communication plans, and post-incident reviews. Learning from incidents by updating guidance, prompts, safeguards, or review processes helps reduce the likelihood of similar failures in the future.

Many AI systems depend on external models, APIs, or data providers. Responsible AI therefore includes vendor due diligence, contractual safeguards, and ongoing monitoring of vendor updates, since changes outside the organization can still affect system behavior.

Operationalizing responsible AI means embedding governance into everyday work rather than treating it as a separate activity. When responsibilities, controls, monitoring, and continuous learning become part of normal delivery processes, responsible AI becomes a sustainable organizational capability, not a one-time initiative.

Summary

In this chapter, we explored the human and organizational impacts of AI systems and the principles and practices that underpin responsible AI governance.

We began by examining AI harms and how to identify them systematically. We introduced harm taxonomies as a structured approach to classifying potential harms, explored who is affected and why some groups face greater exposure, examined bias and discrimination as ongoing governance challenges, and considered the reputational, legal, operational, and strategic consequences AI harms can create for organizations.

We then looked at the risks organizations face when deploying AI systems. These included security and operational risks, privacy risks arising from over-collection, secondary use, and sensitive inference, business risks affecting revenue, reputation, and compliance, and broader AI-specific risks such as hallucinations, over-reliance, manipulation, and adversarial attacks.

Next, we introduced the ethical principles and frameworks that guide responsible AI. We examined how the Fair Information Practice Principles support responsible data use, explored the Organization for Economic cooperation and Development (OECD) AI Principles as a widely adopted framework for trustworthy AI, and discussed the ethical issues of fairness, transparency, accountability, and safety that shape governance decisions.

Finally, we focused on putting these principles into practice. We explored the foundational controls that reduce ethical risk, the organizational culture needed to support responsible AI, the characteristics of trustworthy AI, and the operational practices that help organizations embed responsible AI throughout the AI lifecycle.

In Chapter 3, Responsible AI Governance and Risk Management, we will examine how organizations establish governance structures, assign responsibilities, and conduct risk assessments to operationalize responsible AI.

3

Responsible AI Governance and Risk Management

Chapter 1, Foundations of Artificial Intelligence and AI Governance and Chapter 2, AI Impacts on People and Responsible AI Principles, established what AI is, the harm it can create, and the principles that guide responsible AI. Those principles only become effective when an organization has the structures, processes, and accountability needed to apply them consistently. This chapter focuses on how organizations operationalize AI governance by defining roles, policies, and governance structures, engaging the stakeholders who influence AI outcomes, conducting risk assessments, and applying the NIST frameworks to support consistent and auditable governance.

In this chapter, we will cover the following topics:

- Building the organizational infrastructure for AI governance
- Engaging stakeholders and aligning with organizational strategy
- Designing and conducting AI risk assessments
- Applying NIST frameworks for AI risk management

Note

Download 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*.

Building the organizational infrastructure for AI governance

This section covers the organizational work that makes AI governance operational. It explores how governance should be tailored to an organization's AI footprint and risk profile, defines the roles and responsibilities of the people involved in developing, deploying, and using AI systems, and explains how policies, leadership support, and governance structures work together to provide consistent oversight. By the end of this section, you will be able to map AI use across your organization, assign governance responsibilities using a structured approach, and explain how governance policies, role clarity, and organizational structures support sustainable AI oversight.

Tailoring AI governance to your organization

Organizations do not face the same AI risks, even when they use similar technologies. An AI system that helps prioritize patients in a hospital presents very different risks from one that ranks advertisements for a marketing campaign. Effective governance therefore cannot rely on a single, standardized approach. It must reflect both the potential impact of AI systems and the way the organization operates. Governance that ignores organizational realities quickly becomes an administrative exercise that teams work around. Governance designed to fit the organization is far more likely to be adopted and sustained.

Tailoring begins by creating an inventory of how AI is currently used or is planned to be used across the organization. Identify which teams deploy customer-facing AI, build internal models, or rely on third-party vendors. Clarify whether AI supports recommendations, automation, ranking, content generation, or decision-making. This inventory provides a clear picture of where the greatest risks and exposures exist, while also revealing unapproved or "shadow" AI usage. With this visibility, governance efforts can be prioritized instead of applying the same level of oversight to every system.

A practical way to prioritize governance is to classify AI systems according to their potential impact. High-impact systems that affect rights, opportunities, safety, or access to essential services require more rigorous review, stronger testing, and clearer accountability. Medium-impact systems typically require standard testing and monitoring with proportionate approval processes, while low-impact systems can follow a simplified governance process supported by basic documentation and security controls. Risk tiering helps organizations focus resources where they are needed most and makes the rationale behind governance requirements easier for teams to understand.

The organization's operating model also influences how governance should be implemented. A centralized approach is often effective when a dedicated data science, compliance, or

governance function can oversee most AI initiatives. A decentralized approach is better suited to organizations where business units develop and manage AI independently. Many organizations adopt a hybrid model, with centralized standards and decentralized implementation. Whatever the approach, governance should reflect how decisions are made. Excessive centralization encourages teams to bypass formal processes, while excessive decentralization can lead to inconsistent controls and make oversight difficult to audit.

Organizational maturity is another important consideration. Mature data governance supports capabilities such as data lineage, access controls, retention policies, and traceability. Organizations with less mature governance may first need to establish core practices such as maintaining data inventories and conducting basic access reviews. Engineering maturity also matters. Teams should have appropriate testing practices, monitoring capabilities, and incident response processes before deploying higher-risk AI systems. Tailoring governance means designing controls that teams can realistically implement while establishing a roadmap for progressively strengthening those controls over time.

Organizations that rely on external AI models and **application programming interfaces** (**APIs**) face additional governance considerations. Vendor updates, model changes, and modified safety controls can all alter system behavior. Governance should therefore include procurement requirements, contractual safeguards, operational monitoring, and procedures for validating vendor updates before deployment. Vendor dependence is not inherently risky, but it does require stronger change management and ongoing oversight.

Distinguishing AI developers, deployers, and users

When AI systems fail, one of the first questions organizations ask is who is responsible. Without clearly defined roles, accountability becomes unclear, delays increase, and corrective actions may focus on the wrong part of the AI lifecycle. Distinguishing between developers, deployers, and users helps organizations assign responsibilities appropriately and identify where governance controls should be applied. These are functional roles rather than job titles, so they may be performed by internal teams, vendors, contractors, or business units. In many organizations, a single organization or even a single team may perform more than one of the following roles.

Developers create or substantially modify AI systems. Their responsibilities include selecting and preparing training data, developing or fine-tuning models, implementing safety controls, evaluating performance, and documenting system capabilities and limitations. Decisions made during development influence fairness, privacy, robustness, and explainability, making developers responsible for many of the technical controls that support trustworthy AI. Their role also extends beyond initial deployment, as ongoing updates, retraining, and maintenance continue throughout the AI lifecycle.

Deployers determine how AI systems are introduced into real-world environments. They integrate models into products and business processes, define how the system will be used, determine the level of automation, and establish monitoring and incident response procedures. Many governance challenges arise during deployment because the operating environment, user population, or business objectives differ from those assumed during development. In many organizations, deployers are product owners or business leaders rather than engineering teams.

Users interact directly with AI systems or rely on their outputs when making decisions. They may include employees, customers, clinicians, reviewers, or customer support agents. Users help reduce risk by following approved workflows, verifying outputs when appropriate, and escalating unexpected behavior. Conversely, overreliance on AI, misuse of outputs, or entering sensitive information into inappropriate systems can introduce additional risk. User training and interface design therefore play an important role in ensuring that AI systems are used responsibly.

Most AI incidents involve more than one of these roles. A developer may introduce a limitation during model training, a deployer may implement the system in a context that increases risk, and a user may rely on an output beyond its intended purpose. Understanding how these roles interact allows organizations to identify the source of a problem more accurately and implement targeted improvements rather than assigning blame to a single team.

Consider an AI-assisted hiring system that consistently disadvantages applicants from a protected group. Developers may need to review training data and fairness testing, deployers may need to change how hiring scores are used or introduce mandatory human review, and users may require additional guidance to ensure that AI recommendations are not treated as final decisions. This illustrates why effective AI governance depends on assigning responsibilities to the people best positioned to influence design, deployment, and operational use.

Building policies for oversight and accountability

Policies provide a consistent basis for decision-making across teams and overtime. Without them, similar projects can be handled differently depending on who is involved. Well-defined policies establish what is permitted, what requires approval, and what evidence teams must provide. They also help new staff adopt established practices more quickly. Because AI systems evolve over time and interact with people in unpredictable ways, governance policies should support consistent, defensible decision-making rather than attempt to anticipate every possible scenario.

A policy defines the rule; for example, customer-facing AI must be disclosed. A procedure explains how that rule is implemented, such as where the disclosure appears and who

approves it. A standard establishes measurable expectations, including how unsafe outputs are tested or which monitoring metrics must be tracked. These elements work together. Policies without procedures leave teams unsure how to comply, while procedures without standards make consistent implementation difficult. Together, they create governance programs that work in practice rather than existing only on paper.

Oversight policies define when decisions require additional review. High-impact AI systems may require approval from legal, privacy, security, and business stakeholders, while medium-impact systems might only require a simplified review involving security and business owners. Low-impact systems may be managed through documentation and baseline controls. Oversight should also extend to material system updates and vendor changes so that significant modifications are reviewed before deployment rather than after problems emerge.

Accountability determines who owns an AI system after deployment. Many organizations have well-defined approval processes but less clarity around long-term ownership. Effective accountability policies assign responsibility for performance, risk management, monitoring, and incident response, while defining the actions to take if monitoring identifies harmful outcomes, such as restricting use or pausing automation. Maintaining documentation and updating controls as risks evolve are equally important. Without clear ownership, AI systems gradually become nobody's responsibility, increasing the likelihood that issues remain unresolved.

AI policies should also provide specific guidance on data handling. This includes defining which personal data may be used for training, what information employees must not enter into generative AI tools, and which retrieval sources are approved. Policies for generative AI often distinguish between systems that may draft content and those permitted to send communications automatically. They may also require human review before AI-generated outputs are used in legal, medical, or financial contexts. Clear guidance reduces the risk of data exposure, unsafe advice, and compliance failures while giving employees confidence about appropriate AI use.

Policies should be written so compliance can be demonstrated. For example, a policy requiring fairness should specify how fairness will be assessed and documented, while a policy requiring performance monitoring should define the relevant metrics and review frequency. Policies also need effective enforcement. Audits, periodic reviews, and monitoring dashboards help confirm that governance requirements are being followed and that policies continue to support day-to-day operations rather than existing solely as documentation.

Clarifying roles and responsibilities

Policies define what should happen, but clearly assigned roles and responsibilities determine whether those policies are carried out effectively. When ownership is well defined, teams

spend less time resolving uncertainty and more time delivering AI systems responsibly. Clear responsibilities also improve incident response by reducing delays, preventing duplicated effort, and ensuring that each participant understands their role throughout the AI lifecycle.

A useful distinction is between responsible and accountable roles. Responsible individuals perform the work, such as conducting fairness testing or preparing documentation. Accountable individuals own the final outcome, including decisions such as approving a high-impact AI deployment. Separating these responsibilities avoids the common situation where everyone appears responsible, but no one ultimately owns the decision. In practice, specialist teams often carry out the work, while business owners or senior leaders remain accountable for the final decision.

Most organizations require a combination of business, technical, and governance functions. Business owners define the purpose of the system and accept residual risk. Product or project leads coordinate delivery, while engineers and data scientists develop and test the solution. Privacy, security, and legal specialists review domain-specific risks, and operations teams monitor deployed systems and respond to incidents. High-impact AI systems may also require review committees. Even when one individual performs multiple roles, each governance function should still be clearly assigned.

Responsibilities should also be mapped across the AI lifecycle. During intake, teams define the business purpose, identify affected stakeholders, and determine the system's risk tier. Development focuses on data quality, model training, and testing. Before deployment, teams validate guardrails, review results, and prepare user guidance. Once deployed, monitoring shifts to model drift, user complaints, unsafe outputs, and unexpected behavior. Incident response includes root-cause analysis, corrective action, and documentation, while retirement involves removing access and enforcing retention and deletion requirements. Assigning responsibilities at each stage helps eliminate governance gaps.

Committees can strengthen governance for high-impact decisions, but requiring committee approval for every change often creates unnecessary delays. A tiered decision-making approach is generally more effective, allowing lower-impact changes to be handled by designated owners while escalating only material updates or higher-risk decisions. Clear escalation thresholds improve efficiency without weakening oversight.

Incident response also depends on clearly assigned responsibilities. Teams should know who has the authority to pause an AI system, communicate with customers, report internally, collect evidence, and lead investigations. They should also understand how to determine whether an issue is isolated or systemic. Defined responsibilities allow organizations to respond consistently and improve their governance processes over time.

Securing leadership support for AI governance

Roles, policies, and governance processes remain effective only when they are supported by leadership. Many AI governance decisions involve trade-offs between speed, cost, operational objectives, and risk. When leaders consistently reinforce governance expectations, teams are more likely to raise concerns and delay high-risk deployments when necessary. Without that support, governance processes are easily bypassed in favor of faster delivery.

Leadership discussions are most effective when AI risks are expressed in business terms. Bias can lead to employment disputes or consumer harm. Privacy failures may result in regulatory action and loss of customer trust. Model instability can affect service availability and customer retention. Connecting technical risks to business outcomes helps leaders prioritize governance alongside other organizational risks.

Practical examples are often more persuasive than abstract discussions. A simple cause-and-effect narrative that explains the AI capability, how it might fail, who could be affected, and how governance controls reduce those risks gives leaders a clear basis for decision-making and investment.

Governance teams should request specific decisions, not general endorsements for responsible AI. These might include approving a risk-tiering methodology, requiring human review for high-impact systems, funding monitoring and incident response capabilities, or establishing procurement requirements for AI vendors. Concrete decisions create governance practices that teams can rely on during implementation.

Senior leaders also need clear visibility into the organization's AI landscape rather than detailed technical reports. A concise dashboard might show the number of AI systems in production, identify those supporting critical decisions, highlight systems requiring attention, and summarize incident trends and resolution times. This level of visibility makes governance easier to monitor and reinforces executive ownership.

Leadership involvement should remain focused on decisions with significant organizational impact. Clearly defined escalation thresholds, for example, high-impact deployments, material model changes, or customer-facing incidents ensure that executive attention is directed where it adds the greatest value without slowing routine operational decisions.

Choosing governance structures and models

A governance structure defines how AI-related decisions are made across the organization. It establishes who reviews proposals, approves deployments, resolves disagreements, maintains governance policies, and manages exceptions. Well-designed governance structures create predictable decision-making processes that balance effective oversight with operational efficiency.

In a centralized governance model, a dedicated team establishes standards and often performs governance reviews for the entire organization. This approach works well when AI development is concentrated within a small number of teams or when organizations require consistent documentation and oversight. However, excessive centralization can create bottlenecks and reduce ownership within business units. Clear service expectations and efficient review processes help address these challenges.

A decentralized model gives individual business units greater responsibility for governance decisions. This often supports faster delivery and allows governance activities to reflect local business needs. The trade-off is a greater risk of inconsistent practices and unapproved AI adoption. Shared templates, training, and periodic audits help maintain consistency while preserving local ownership.

Many organizations adopt a hybrid approach that combines centralized governance standards with decentralized implementation. A central governance function establishes policies, templates, and review requirements for higher-risk AI systems, while business units remain responsible for day-to-day delivery and operations. Hybrid models work best when escalation thresholds clearly define which decisions require central review.

Governance committees can improve decision quality by bringing together expertise from multiple disciplines. Their effectiveness depends on a clearly defined charter, a focused scope, and predictable review cycles. Committees are most valuable for high-impact systems, significant model changes, and exception requests. They should support decision-making by reviewing evidence rather than replacing the responsibilities of product or engineering teams.

Some organizations also embed responsible AI reviewers within product and engineering teams, similar to security champions. This enables governance activities to occur earlier in the development lifecycle and allows reviewers to assess AI systems within their operational context. Maintaining consistency across embedded reviewers requires shared standards, regular calibration, and ongoing collaboration with the central governance function. This approach can improve both governance maturity and development speed by reducing unnecessary handoffs.

Together, roles, policies, leadership support, and governance structures provide the organizational foundation for responsible AI. With these internal governance mechanisms in place, the next section expands the focus to the broader stakeholder ecosystem and examines how governance aligns with organizational strategy.

Engaging stakeholders and aligning with organizational strategy

AI governance depends not only on policies and processes but also on the people who shape, use, and oversee AI systems. This section examines how to identify and engage stakeholders throughout the AI lifecycle, develop role-specific training that builds practical competence, and align AI initiatives with the organization's strategic objectives and risk appetite. By the end of this section, you will be able to map and engage stakeholders effectively, design role-specific training programs, and connect AI strategy to risk management in a way that supports informed leadership decisions.

Identifying and engaging key stakeholders

AI systems affect far more people than the teams that design and build them. Limiting governance discussions with engineers and product owners can overlook risks that emerge only in day-to-day operations. Effective stakeholder engagement helps organizations understand how different groups define success, identify potential harm, and surface unintended consequences before deployment. It does not require involving everyone in every decision. Instead, the goal is to engage the right stakeholders at the right points in the AI lifecycle.

Key internal stakeholders typically include legal, privacy, security, risk and compliance, customer support, and operations teams. Each brings a different perspective on AI risk. Customer support teams often identify complaints and usability issues first, operations teams detect system drift and reliability problems, privacy and security teams assess data handling risks, and legal teams identify regulatory obligations and documentation gaps. Engaging these groups early makes issues easier to address, while involving them only at the end of a project often results in delays or last-minute changes.

External stakeholders may include customers, employees, vendors, auditors, regulators, and, in some cases, community representatives. Customers can highlight whether AI-driven interactions feel transparent and fair, while employees can identify workflows that encourage overreliance on AI. Vendors provide insight into model updates, data handling, and operational dependencies, and auditors focus on evidence, consistency, and governance processes. Their perspectives are particularly valuable for high-impact AI systems, where internal teams may overlook risks that become apparent only in real-world use.

A practical way to prioritize engagement is to map stakeholders according to impact and influence. Impact reflects who is affected if the AI system fails, while influence identifies those who can shape its design, deployment, or operation. Stakeholders with both high impact and high influence should be involved early. Those with high impact but less influence should still

be represented through mechanisms such as user research, advocacy groups, or structured feedback. This approach ensures that engagement remains focused, proportionate, and well documented.

Stakeholder engagement can take many forms, including workshops, interviews, pilot programs, and structured review sessions. For high-impact AI systems, scenario-based discussions help identify potential harms and misunderstandings before deployment. Customer-facing systems can benefit from usability testing and feedback on disclosures, while internal tools may be introduced through pilot programs with defined feedback loops. The objective is to capture actionable insights, for example, that users misunderstand an output or staff cannot explain a decision rather than collecting opinions without a clear path to improvement.

Disagreement among stakeholders is inevitable. Product teams may prioritize speed, legal teams' defensibility, security teams stronger controls, and customer support team's better user outcomes. Governance should make trade-offs explicit and document the rationale behind decisions. Maintaining structured decision records that capture what was decided, why it was decided, which risks remain, and how they will be managed improves consistency and reduces repeated debate during future projects and incidents.

Designing role-specific training and awareness programs

Policies and governance processes are only effective when people understand how to apply them. Training bridges the gap between governance requirements and day-to-day practice. AI governance training should extend beyond technical teams to include product managers, executives, legal and privacy specialists, customer support teams, and end users. Its purpose is to build a shared understanding of AI risks, system limitations, and appropriate escalation procedures while reducing the likelihood of overreliance on AI outputs.

Different roles require different knowledge and skills. Engineers benefit from training on testing, monitoring, and threats such as prompt injection. Product teams need to understand high-impact use cases, user harms, and safe design practices. Legal and privacy professionals should understand how AI changes data use and decision-making, while customer support teams need guidance on handling AI-related complaints and reporting unsafe outputs. Executives require a strategic understanding of risk tiering, governance thresholds, and the questions they should ask when reviewing AI initiatives. A single generic training program cannot address these varied responsibilities.

Training should focus first on the most common and significant sources of AI risk. Hallucinations illustrate that fluent outputs can still be incorrect. Bias demonstrates how strong overall performance may conceal harm to specific groups. Privacy leakage highlights the risk of exposing sensitive information, while model drift shows how performance can change

after deployment. Security training should also address misuse scenarios, particularly for generative AI systems that accept user input. Familiarity with these patterns helps teams recognize risks before they escalate.

Effective training emphasizes actions as well as awareness. Each topic should explain what participants should do when they encounter a potential issue. For example, sensitive information in an AI output should be reported immediately, discriminatory outcomes should be escalated with supporting evidence, declining model performance should trigger a formal review and attempts to manipulate a generative AI system should be logged and investigated. Clear response guidance reduces uncertainty during real incidents.

Scenario-based exercises are one of the most effective ways to reinforce learning. Tabletop exercises allow teams to respond to simulated incidents such as unsafe outputs or data leakage, while product teams can practice reviewing high-impact use cases and selecting appropriate controls. Customer support teams can rehearse responses to complaints about unfair AI decisions. These exercises expose gaps in both processes and understanding while building confidence before real incidents occur.

Training effectiveness should be measured through changes in behavior rather than attendance alone. Useful indicators include faster reporting of near misses, fewer unsafe outputs during pilots, reduced sharing of sensitive information with AI tools, and more efficient review processes because evidence has been prepared correctly. Short assessments and role-specific certifications can also help evaluate understanding. Ultimately, effective training should contribute to fewer incidents and faster response times.

Aligning AI and risk strategy

AI strategies typically focus on innovation, efficiency, and business growth, while risk strategies emphasize preventing harm, maintaining compliance, and reducing organizational exposure. Treating these objectives separately often results in projects being delayed or redesigned late in the development process. Aligning AI and risk strategies from the outset allows organizations to pursue innovation while integrating appropriate governance controls into delivery plans.

Alignment begins with clearly defined objectives. Rather than setting broad goals such as improve customer service, organizations should define measurable outcomes, such as reducing response times while maintaining quality and safety. Similarly, initiatives to automate decisions should specify which decisions may be automated and which require human review. Clear objectives enable governance teams to design proportionate controls and reduce the risk of AI being used beyond its intended purpose.

Risk assessments become more effective when they are directly linked to business objectives. A customer support initiative may prioritize hallucination and privacy risks, while fraud

detection focuses on false positives and model drift. Hiring systems require particular attention to bias and explainability. Mapping risks to organizational goals keeps governance activities proportionate and positions risk management as an enabler of business strategy rather than an obstacle.

Risk appetite defines the level of uncertainty an organization is prepared to accept. Some organizations encourage experimentation for low-impact use cases while requiring stricter controls for systems that affect people, rights, or critical services. Clearly communicating these boundaries, for example, prohibiting fully automated employment decisions without human review or restricting sensitive data in external AI tools helps teams make decisions confidently and consistently.

Governance controls should also be connected to business outcomes rather than compliance alone. Monitoring improves service reliability and reduces customer complaints, documentation simplifies audits and regulatory reviews, human oversight reduces the likelihood of harmful decisions, and vendor controls minimize operational disruption. Demonstrating these benefits helps position governance as a contributor to business value and supports more informed investment decisions.

Viewing AI initiatives as a portfolio further strengthens strategic alignment. Some projects offer high value but also require significant governance oversight, while others provide limited value relative to their risk and may not justify further investment. A portfolio perspective helps organizations allocate governance resources efficiently, prioritize monitoring and training, and understand cumulative risk across multiple AI systems rather than assessing each project in isolation.

Aligning AI initiatives with organizational strategy ensures that governance supports both business objectives and risk management. The next section builds on this foundation by introducing structured AI assessment processes that help organizations evaluate AI systems consistently before and after deployment, using evidence-based methods, not ad hoc judgment.

Designing and conducting AI risk assessments

This section focuses on the practical processes that support AI risk management. It explains how to design assessment processes that scale according to risk, evaluate AI systems by considering their context and real-world outcomes, and prioritize identified risks using probability and severity. By the end of this section, you will be able to design a scalable AI assessment process, assess risk by examining context and outcomes, and use a scoring matrix to prioritize the risks that require immediate attention.

Building scalable AI assessment processes

With governance structures, leadership support, stakeholder engagement, training, and strategic alignment in place, the next step is translating those foundations into a consistent assessment process that supports decisions before deployment and throughout an AI system's lifecycle. An AI assessment provides a structured way to answer three practical questions: What is the system intended to do? What could go wrong? Is it ready for deployment? Assessments provide evidence that supports consistent decision-making and identifies issues before they become incidents.

An effective assessment begins with a focused intake process that captures the essential characteristics of the AI system. This includes its purpose, the people it affects, the decisions or influence it has, the data it relies on, the origin of that data, whether it is used internally or externally, and the degree of automation involved. Defining this context early determines how extensive the assessment should be and helps prevent systems from expanding into higher impact uses without appropriate governance.

Assessment depth should be proportionate to risk. High-impact systems require more extensive testing, stronger oversight, and more comprehensive documentation. Medium-impact systems generally require standard reviews and monitoring plans, while low-impact systems can often be managed through simplified documentation and baseline controls. Applying consistent risk tiers allows organizations to direct resources where they are most needed while avoiding unnecessary delays for lower-risk initiatives.

Assessments should collect evidence that supports practical governance decisions. Typical evidence includes the data used to develop the system and any associated restrictions, testing results for performance, stability, safety, and fairness, privacy and security controls, human oversight mechanisms, and post-deployment monitoring plans. Documentation should be detailed enough to support informed decisions while remaining practical to maintain throughout the system's lifecycle.

Assessment output should also be tailored to different audiences. Engineers need clear remediation actions, leaders require concise summaries of risk and deployment readiness, operations teams need monitoring thresholds and incident procedures, and customer support teams need guidance for handling complaints and escalation. Presenting assessment results in formats that support these different responsibilities increases the likelihood that assessments will be used in practice rather than treated as administrative documentation.

Assessment should not end when an AI system is deployed. Changes in data, user behavior, operating environments, or third-party models can all alter the system's risk profile. Ongoing monitoring, periodic reviews, and change-management processes should therefore trigger reassessments whenever significant changes occur. Viewing assessment as a continuous

process rather than a one-time approval helps organizations manage AI risks throughout the system's lifecycle.

Assessing risk through context and outcomes

Two AI systems built using the same technology can present very different risks depending on where and how they are used. Risk assessment therefore begins with understanding the system's context, including the domain in which it operates, the people affected, and the role its outputs play in decision-making. A recommendation engine for entertainment carries very different consequences from an AI system supporting clinical decisions. Organizational priorities, such as pressure to reduce costs or accelerate delivery, can also influence how people interact with AI and increase the likelihood of overreliance. Understanding these contextual factors provides the foundation for a meaningful risk assessment.

It is equally important to distinguish between outputs and outcomes. Outputs are what the AI system produces, such as scores, recommendations, or generated content. Outcomes are the real-world consequences of those outputs, including decisions such as approving a loan, rejecting a job applicant, or delaying medical treatment. Governance focuses on outcomes because this is where harm occurs. An AI system may achieve strong technical performance while still producing unfair or harmful outcomes for particular individuals or groups. Considering outcomes also highlights indirect effects, such as changes in behavior caused by increased surveillance or automation.

Context also includes understanding who may be affected by the AI system, both directly and indirectly. Customers, employees, applicants, patients, and children may experience very different levels of harm from the same technical error. Vulnerable groups often have fewer opportunities to challenge or appeal AI-assisted decisions, increasing the need for stronger safeguards. Systems that influence rights, opportunities, or access to essential services generally require greater scrutiny because the consequences of failure are more significant.

A structured assessment considers multiple categories of harm within the relevant context. Privacy risks may arise through inappropriate data use or inference. Fairness risks may result in discriminatory outcomes or unequal error rates. Safety risks can stem from incorrect recommendations or excessive automation, while security risks include manipulation and data leakage. Transparency risks arise when decisions cannot be explained or challenged, and operational risks include model drift and instability. Assessing these categories systematically reduces the likelihood that important risks will be overlooked.

Risk should also be evaluated across the AI lifecycle. Some risks originate during data collection, others emerge during model development or system integration, and still others appear only after deployment as user behavior or operating conditions change. An AI system that performs well at launch may become unreliable as data drifts or its scope expands.

Considering risk across the lifecycle helps organizations identify likely points of failure and establish monitoring processes that detect emerging issues early.

A risk assessment should conclude with a clear governance decision. An AI system may proceed, proceed subject to defined conditions, or require redesign before deployment. Conditions might include mandatory human oversight, stronger monitoring, restricted data access, or a phased rollout. Any remaining residual risk should also be documented so that leadership understands which risks have been accepted and why. The objective is not to eliminate all risks but to ensure that remaining risks are understood, managed, and consciously accepted.

Scoring risk using probability and severity

Once risks have been identified, they must be prioritized. Risk scoring provides a structured way to compare risks consistently, instead of relying solely on individual judgment. Separating probability from severity creates a shared framework for prioritization and helps organizations allocate resources where they will have the greatest impact.

Probability estimates how likely a harmful event is to occur. Factors such as system complexity, user exposure, automation, operational maturity, and monitoring effectiveness all influence this assessment. For example, a customer-facing generative AI assistant is generally exposed to more unpredictable interactions than an internal decision-support tool. Probability estimates do not require precise calculations, but they should be based on a consistent approach that different teams can apply reliably.

Severity measures the potential impact of an identified risk. This includes the consequences for affected individuals, the number of people who may be affected, and the potential legal, financial, operational, or reputational impact on the organization. Prioritizing severity helps ensure that high-impact AI systems receive appropriate governance attention, even when harmful events are relatively unlikely.

Combining probability and severity in a risk matrix provides a consistent basis for prioritization. Risks that are both highly probable and highly severe generally require immediate mitigation before deployment or expansion. Risks with lower probability or lower severity may be managed through monitoring or proportionate controls, while high-severity risks often justify stronger safeguards regardless of their estimated likelihood. A documented matrix also provides a clear rationale for governance decisions and helps organizations explain why particular risks receive priority.

Risk scoring is most effective when common assessment errors are avoided. Technical performance metrics alone rarely provide a complete picture of risk, and severity should not be measured solely in terms of business impact while overlooking effects on people. Similarly, the absence of previous incidents should not automatically be interpreted as evidence that a risk is

unlikely. Combining evidence from testing, pilot deployments, incident history, and stakeholder feedback produces more balanced and defensible assessments that can be refined as additional information becomes available.

Risk scores should ultimately guide action. High-risk findings may require stronger monitoring, additional testing, mandatory human review, or phased deployment. Medium-risk findings may call for targeted training or documentation updates, while lower-risk issues may be managed through routine monitoring. Linking risk scores to predefined governance actions makes the assessment process predictable, supports consistent decision-making, and reinforces confidence that governance activities are driving practical outcomes rather than simply producing documentation.

Structured assessments and risk scoring provide a consistent approach to identifying, evaluating, and prioritizing AI risks. The next section introduces the NIST AI Risk Management Framework and the ARIA program, showing how these frameworks provide a structured foundation for managing AI risk across the system lifecycle.

Applying NIST frameworks for AI risk management

Chapter 2, AI Impacts on People and Responsible AI Principles introduced the ethical principles and frameworks that define responsible AI. This section focuses on their practical application through the NIST AI Risk Management Framework (AI RMF) and the NIST ARIA program. Together, they provide structured approaches for identifying, evaluating, and managing AI risks throughout the AI lifecycle. By the end of this section, you will be able to explain how the NIST AI RMF structures AI risk management and describe how ARIA supports model evaluation before and after deployment.

Applying the NIST AI Risk Management Framework

The **NIST AI Risk Management Framework (AI RMF)** provides organizations with a structured approach to managing AI risk throughout the design, development, deployment, and operation of AI systems. The framework treats AI risk as something that can be planned for, tested, and improved over time while continuously improving governance practices. Its principles are applicable across industries and AI technologies, making it a practical foundation for operational AI governance.

One of the challenges of AI governance is that responsibility is distributed across multiple teams. Data scientists, product owners, engineers, compliance specialists, and operational teams each manage different aspects of AI systems and may not have complete visibility into the entire lifecycle. The AI RMF addresses this by encouraging organizations to establish clear governance structures, define roles and responsibilities, and equip teams with the tools,

processes, and evidence needed to support consistent decision-making. This shifts governance from informal judgment to documented, repeatable practice.

The framework also defines a set of trustworthy AI characteristics that provide a shared understanding of what responsible AI should achieve. Instead of relying on broad statements about trustworthiness, these characteristics establish concrete qualities that technical and business teams can use to evaluate AI systems and justify governance controls.

Valid and reliable systems perform as intended under expected operating conditions and continue to produce consistent results beyond controlled testing environments. Reliability depends on testing with representative data, evaluating performance under realistic conditions, and confirming that models generalize appropriately beyond their training environment.

Safe systems minimize the likelihood of causing harm, particularly in high-impact applications. Safety includes preventing unsafe recommendations, limiting harmful automation, designing systems that fail safely when uncertainty is high, and continuously monitoring deployed systems for emerging risks. For generative AI, safety also encompasses reducing toxic content, misleading outputs, and unsafe instructions.

Secure and resilient systems protect models, data, prompts, and supporting infrastructure against misuse while maintaining reliable operation under changing conditions. Practical measures include access controls, logging, threat modeling, adversarial testing, protection against prompt injections, and recovery procedures that allow organizations to respond effectively when failures occur.

Accountable and transparent AI requires clearly defined ownership together with appropriate visibility into how AI systems are used. Accountability establishes who is responsible for governance decisions, while transparency ensures that users, reviewers, and regulators receive information appropriate to their needs about system capabilities, limitations, and decision processes.

Explainable and interpretable AI provides explanations that are appropriate for the intended audience. Customers may require plain-language explanations, while auditors and technical reviewers often need more detailed evidence. Explainability also supports bias detection, appeals processes, and ongoing governance by helping stakeholders understand how AI systems influence decisions.

Privacy-enhanced AI protects personal information throughout the AI lifecycle through practices such as data minimization, access controls, retention management, and safeguards against information leakage. Because AI systems may also infer sensitive information from seemingly non-sensitive data, privacy protections should address both direct and indirect disclosure risks.

Fair with harmful bias managed means organizations actively identify, assess, and mitigate discriminatory outcomes instead of assuming fairness will emerge automatically. This includes evaluating performance across different groups, identifying proxy variables, validating outcomes within their operational context, and monitoring fairness after deployment as data and user populations change.

The AI RMF organizes these principles through four interconnected functions: **Govern, Map, Measure, and Manage**. Together, these functions provide a repeatable structure for integrating AI governance into organizational practice. These functions reinforce one another throughout the AI lifecycle.

Govern establishes the organizational foundation for AI risk management. It defines governance structures, assigns responsibilities, develops policies, provides training, and creates the accountability mechanisms that enable the remaining functions. Strong governance ensures that AI risk management is applied consistently across projects and business units.

Map focuses on understanding how AI systems will be used and identifying the risks associated with those uses. This includes defining the operational context, identifying affected stakeholders, recognizing dependencies on external models or data sources, and determining where risks related to privacy, fairness, safety, security, transparency, and operational performance may arise.

Measure evaluates identified risks using evidence. Activities include testing model performance, fairness, safety, privacy, robustness, and resilience while monitoring these characteristics over time. Continuous measurement ensures that governance reflects changing operating conditions, not only pre-deployment testing.

Manage translates assessment results into governance decisions. Organizations determine appropriate mitigation measures, adjust deployment conditions, establish monitoring requirements, respond to incidents, and continuously improve governance processes based on new evidence. Together, the Measure and Manage functions transform AI governance from periodic review into an ongoing operational discipline.

Using the NIST ARIA program for model evaluation

While the AI RMF provides an overall structure for AI risk management, the NIST ARIA program focuses on strengthening AI evaluation, particularly for large language models. ARIA complements the AI RMF by promoting practical, evidence-based evaluation methods that help organizations determine whether AI systems perform as expected under realistic operating conditions.

ARIA emphasizes scenario-based evaluation that reflects how AI systems are used, instead of relying primarily on benchmark scores. Test scenarios may involve policy compliance,

customer support interactions, safety boundary testing, or situations involving sensitive information. These realistic evaluations expose weaknesses that standardized benchmarks often fail to detect while producing evidence that is easier for technical teams and leadership to interpret.

Large language models are particularly sensitive to changes in prompts, context, and user behavior. As a result, traditional static evaluation methods may overlook failures such as hallucinations, unsafe advice, policy violations, or prompt injection attacks. Scenario-based evaluation provides a more representative picture of operational performance by testing models under realistic conditions, including ambiguous or adversarial inputs.

ARIA also promotes improvements in the tools, methods, and metrics used to evaluate AI systems. Organizations benefit from repeatable testing frameworks, structured prompt libraries, standardized scoring methods, and evaluation techniques that improve consistency across teams. Expanding evaluation beyond fluency and benchmark performance allows organizations to assess broader trustworthiness characteristics that are directly relevant to deployment decisions.

Another important objective is verifying claims about model capabilities. Demonstrations often highlight ideal use cases, but operational deployment requires evidence that models perform consistently across different users, contexts, languages, and edge cases. Structured evaluation helps organizations determine whether a model is appropriate for a particular use case instead of relying solely on individual judgment.

ARIA also encourages red teaming, where organizations deliberately test how models respond to attempts to bypass safety controls, manipulate outputs, or extract restricted information. For generative AI systems, prompt injection, data leakage, and other adversarial techniques represent significant operational risks. Red teaming helps identify weaknesses before deployment so that safeguards can be strengthened proactively.

In addition to controlled testing, ARIA emphasizes field evaluation under real operating conditions. Users interact with AI systems differently from test engineers, often asking ambiguous questions, working under time pressure, or using AI in unexpected ways. Observing system behavior in realistic workflows reveals issues such as overreliance, inconsistent performance, latency, or model drift that may not appear during laboratory testing.

Structured evaluation also supports technology selection and deployment decisions. Before adopting an external model or releasing an internally developed one, organizations need evidence that it satisfies their operational requirements and risk tolerance. ARIA provides a repeatable basis for comparing models, defining deployment conditions, and supporting go/no-go decisions with documented evidence, not subjective judgment.

The initial ARIA activities focus on the distinctive risks associated with generative AI, including hallucinations, prompt injection, unsafe content generation, and information leakage. As AI technologies continue to evolve, the same evaluation principles are expected to expand to additional AI system types and emerging risk categories, ensuring that AI evaluation evolves alongside the technologies it is designed to assess.

Summary

In this chapter, we covered how organizations translate AI governance principles into day-to-day practice. We began by establishing the governance infrastructure, exploring how governance should be tailored to an organization's AI footprint and risk profile, how roles and responsibilities are assigned across the AI lifecycle, and how policies, leadership support, and governance structures work together to provide consistent oversight.

We then addressed the human and strategic layer that influence successful AI governance. This included identifying and engaging internal and external stakeholders, developing role-specific training that promotes responsible AI practices, and aligning AI initiatives with organizational objectives and risk appetite.

Building on these governance foundations, we examined the practical mechanics of risk assessment. We learned how to design scalable assessment processes, evaluate AI systems by considering their context and potential outcomes, and prioritize risks using probability and severity to support consistent, evidence-based decision-making.

Finally, we introduced two important NIST resources for operational AI governance. The NIST AI Risk Management Framework (AI RMF) provides a structured approach to governing, mapping, measuring, and managing AI risks while incorporating the characteristics of trustworthy AI. We also introduced the NIST ARIA program, which strengthens model evaluation through scenario-based testing, capability validation, red teaming, and real-world evaluation of large language models.

In *Chapter 4, Governing AI in Production,* we will apply these governance foundations to the practical work of building AI systems, examining how to oversee each stage of the development lifecycle from problem framing and impact assessment to testing, data governance, and the documentation that makes AI systems accountable.

4

Governing AI Development

The previous chapters established what AI is, how it affects people, and the governance and risk management principles that support its responsible use. This chapter shifts from those foundations to the practical work of developing AI systems. Governance becomes part of every stage of the development lifecycle, from defining a system's purpose and scope to testing, deployment, monitoring, and eventually retirement. Decisions made throughout that lifecycle, particularly those involving data, system boundaries, testing, and documentation, shape whether an AI system remains trustworthy in practice.

Using the AI development lifecycle as its organizing framework, this chapter follows the governance decisions that influence each stage of development. It explores how to define AI use cases responsibly, assess potential impacts before deployment, govern data throughout the pipeline, and produce the documentation needed to support accountability. Together, these practices help organizations build AI systems that are both technically effective and governed throughout their lifecycle.

In this chapter, we will cover the following topics:

- Governing the AI development lifecycle
- Testing, monitoring, and operationalizing AI systems
- Governing data across the AI pipeline
- Governing model features and conformity documentation

Governing the AI development lifecycle

This section introduces the AI development lifecycle as the framework for understanding how governance is applied throughout the process of building AI systems. It focuses on the early decisions that shape a system's purpose, scope, and potential impact, where governance has

the greatest influence. You will also explore how impact assessments and stakeholder engagement help identify risks early and guide responsible development.

By the end of this section, you will be able to explain the AI development lifecycle from problem framing through retirement, conduct an impact assessment that connects harms to controls, and apply a practical method for stakeholder engagement.

Navigating the AI system development lifecycle

The AI development lifecycle provides the framework for governing AI systems from conception to retirement. Governance is embedded throughout the lifecycle, ensuring that decisions about design, data, testing, deployment, and ongoing operation remain transparent and accountable. Viewing development through this lifecycle helps organizations build AI systems that are not only technically effective but also safe, lawful, and aligned with their intended purpose.

The lifecycle begins by defining why the system is needed, the problem it is intended to solve, and the outcomes it is expected to achieve. It then progresses through data collection and preparation, model selection and development, testing and validation, deployment, operational monitoring, and retirement. Each stage introduces different governance responsibilities, and the decisions made at one stage often influence the risks that emerge later.

Governance is particularly important because AI systems continue to evolve after deployment. Changes in data, user behavior, operating environments, or third-party models can alter system performance over time. Without ongoing monitoring and change management, systems may drift from their original purpose or behave differently from when they were initially approved. Maintaining documentation enables organizations to explain how decisions were made and demonstrate accountability to regulators, auditors, customers, and other stakeholders.

The sections that follow focus on the planning activities that precede development, beginning with how to define the business problem, determine system scope, conduct an impact assessment, and engage stakeholders.

Defining the business problem and use cases

Defining the business problem lays the foundation for responsible AI development. Before selecting a model or designing a solution, organizations need a clear understanding of the decision or workflow the system is intended to support and how success will be measured. A use case defines that purpose by describing a specific, bounded application of AI within a real business process. Well-defined use cases focus on specific tasks with clear objectives and measurable outcomes, such as drafting customer responses for human review.

A practical starting point is to identify the decision being made today, who makes it, and the consequences of an incorrect decision. These questions help determine whether AI should support human decision-making or automate the process entirely. For high-impact decisions, such as those affecting employment, healthcare, or financial services, organizations should carefully evaluate whether full automation is appropriate. A clearly defined decision also helps identify the information required, the appropriate controls, and the level of human oversight needed.

The same AI model may require very different governance controls depending on how it is used. For example, using AI to summarize résumés for a recruiter who makes the final hiring decision presents a different risk profile from automatically ranking or rejecting applicants without transparency or an opportunity for review. Similarly, drafting a patient message for clinician approval carries lower risk than providing medical advice directly to patients without oversight. These examples demonstrate that governance requirements are shaped by the intended use case and the potential impact of the decisions the system supports.

A complete use case definition also assigns ownership and accountability. Organizations should identify who is responsible for approving deployment, monitoring ongoing performance, responding to complaints, investigating incidents, and suspending the system when necessary. Responsibilities may extend across product, engineering, legal, privacy, security, and operations teams, and should also account for third-party vendors where external models are used. Clearly defined ownership keeps governance responsibilities visible, coordinated, and consistently assigned throughout the AI lifecycle.

Determining the scope of an AI system

Defining the scope of an AI system means establishing the boundaries of what is being built, how it operates, and what the organization is responsible for governing. Although this may appear straightforward, unclear system boundaries are a common source of governance failures. When scope is poorly defined, important elements such as data pipelines, retrieval systems, user interfaces, and human workflows may be overlooked, resulting in incomplete testing, monitoring, and documentation. Scope also determines whether governance applies to a single model or to a broader system of interconnected components that work together.

The system boundary extends beyond the model itself. It includes the training or fine-tuning approach, prompts, and prompt templates, external tools and plug-ins, retrieval mechanisms, user interfaces, logging and monitoring capabilities, and any human review processes that form part of the control strategy. For example, an interface that presents AI-generated responses with excessive confidence may encourage users to rely on outputs without appropriate verification. Defining these elements within the system boundary provides a more accurate basis for risk assessment and governance.

Governance also requires organizations to consider dependencies that fall outside their direct control. Vendor model updates, third-party data sources, upstream data quality, regulatory changes, and downstream user behavior can all influence system performance. Identifying these dependencies during scoping allows organizations to establish appropriate monitoring, change management, and contractual controls before deployment.

Scope should also reflect how people interact with the system. Developers, deployers, end users, supervisors, and customer support teams each influence how AI is used in practice. Mapping workflows reveals where human judgment is applied, how AI outputs influence decisions, and what happens when those outputs are incorrect. For example, an analyst investigating fraud may prioritize cases based on AI-generated risk scores, while a customer support representative may send AI-generated responses without review during periods of high workload. Considering these interactions helps identify socio-technical risks that emerge through the combination of technology and human behavior.

A practical way to define scope is to classify system use into three categories:

- **Intended use:** the approved workflow and expected use of the AI system.
- **Foreseeable misuse:** ways the system might reasonably be used outside its intended purpose, particularly under operational pressure.
- **Prohibited use**: activities the system must never support because they introduce unacceptable risk.

This approach helps organizations design safeguards, monitoring activities, and user guidance that reflect both expected and unintended patterns of use. It is particularly valuable for general-purpose AI systems and large language models, which can be adapted to many different tasks.

Once the system scope has been clearly defined, organizations can determine the testing, documentation, and governance activities the system requires. For example, retrieval-augmented systems require testing for information leakage, human review processes should be validated under realistic operating conditions, and vendor-managed components require monitoring for external changes. Clear ownership of prompts, data, models, and operational monitoring also ensures that governance responsibilities remain assigned after deployment.

Conducting impact assessments

An impact assessment is a structured process for evaluating how an AI system may affect people, organizations, and society before it is deployed. More than a compliance document, it supports informed decision-making by identifying potential harms, assessing their likelihood and impact, and determining whether appropriate safeguards are in place. Impact assessments are particularly important for AI systems that influence high-impact decisions, such as employment, healthcare, housing, education, or financial services, where errors or unintended

outcomes can have serious consequences. As AI systems evolve, the assessment also provides a baseline for reviewing whether changes to the system introduce new risks. An effective impact assessment addresses a series of practical questions. What is the intended purpose of the system, and what decisions does it influence? Who may be affected, including indirect users and vulnerable groups? What data is used, where does it originate, and does it include sensitive information? What potential harms could arise, including risks to privacy, fairness, safety, security, or operational reliability? Which controls reduce those risks, and what residual risks remain? Finally, what evidence supports the decision to proceed, including testing results, governance reviews, and oversight plans? Answering these questions enables decision-makers to evaluate AI systems using documented evidence.

Impact assessments are most effective when they connect identified harms to specific components of the AI system. Privacy risks may arise from training data retention or retrieval mechanisms that expose sensitive information. Fairness concerns may result from historical bias in training data or uneven error rates across different groups. Safety risks may stem from harmful outputs or excessive reliance on AI-generated recommendations, while operational risks may emerge from system outages or model drift. Mapping harm to the components and workflows that create them enables organizations to implement targeted controls that address the underlying causes of risk.

The assessment should be supported by evidence gathered throughout development. This may include fairness evaluations, privacy leakage testing, security assessments, usability studies, pilot results, stakeholder feedback, and, legal or policy reviews. Evidence helps demonstrate that governance decisions are based on systematic evaluation and makes those decisions easier to defend if the system is later questioned or audited. It also helps teams make informed trade-offs, such as introducing additional human review to improve safety while recognizing the operational impact on efficiency.

Consider a bank developing a large language model assistant to support customer service agents. An impact assessment would evaluate whether the assistant could disclose personal information, generate incorrect financial advice, or lead customers to believe they are interacting with a human representative. Appropriate controls might include limiting retrieval to the current customer session, redacting sensitive information, requiring human approval for high-risk actions, and clearly disclosing when customers are interacting with AI. The assessment would also define prohibited uses, such as approving transactions or modifying customer accounts, and establish monitoring indicators such as complaint volumes, unusual prompt activity, or increases in inaccurate responses. As the assistant expands to support additional tasks, the assessment should be revisited to evaluate whether new risks require additional controls.

To remain useful, impact assessments should produce clear, actionable outcomes. They should specify the conditions for deployment, identify escalation triggers that require reassessment, assign ownership for implementing controls, and provide both executive summaries and detailed technical records for different audiences. When integrated into the development process, impact assessments support consistent decision-making and provide a reusable governance framework as organizations scale AI across multiple projects.

Engaging stakeholders in development

Effective stakeholder engagement involves identifying and involving the right people early enough to influence system design. AI systems affect privacy, security, legal compliance, business processes, and user experience, making it unlikely that a single team will identify every potential risk. Stakeholders who understand how systems operate in practice can often identify operational challenges and failure modes that are not apparent during design. Early engagement helps organizations address these issues before they become costly redesigns or deployment delays.

A practical approach is to map stakeholders according to two factors: **impact** and **influence**. Impact reflects how a person or group may be affected by the AI system, while influence reflects their ability to shape requirements, approvals, or operational decisions. Stakeholders with both high impact and high influence should be involved early because they can identify key risks and practical constraints. Groups with high impact but limited influence may be particularly vulnerable and should not be overlooked, while stakeholders with high influence but lower direct impact, such as senior leadership or procurement teams, remain essential because they can enable or delay deployment. This approach helps organizations prioritize engagement where it will have the greatest value.

Different stakeholders contribute distinct perspectives throughout development. Product and business teams define objectives and business constraints, while engineering teams understand system architecture, failure modes, and testing requirements. Privacy specialists assess data governance and regulatory obligations, security teams identify threats and defensive controls, and risk and compliance teams align AI initiatives with organizational governance requirements. Operations and customer support teams contribute practical knowledge about day-to-day workflows and incident management, while human resources or communications teams may play a key role in employee-facing systems or initiatives with broad public visibility. Bringing these perspectives together helps reduce governance blind spots before deployment.

Where AI systems affect external groups, stakeholder engagement should extend beyond the organization. External stakeholders may include customers, employees, suppliers, partners, vendors, or, in some situations, independent reviewers. Although direct engagement is not always necessary, organizations should consider how these groups may be affected by the AI

system and incorporate those considerations into impact assessments. For example, an AI system used in recruitment may require transparency and mechanisms for redress even if applicants are not involved during development. Organizations that rely on external AI providers should also engage vendors to understand model updates, data handling practices, and incident response responsibilities. The level of external engagement should remain proportionate to the potential impact of the AI system.

Stakeholder engagement should produce clear, actionable outputs. Workshops can validate use cases, system boundaries, and likely failure modes, while interviews with operational teams often reveal workarounds that bypass intended controls. Tabletop exercises help evaluate incident response procedures before deployment, and pilot projects provide valuable feedback on usability and the potential for overreliance on AI outputs. Recording assumptions, decisions, and identified risks creates a shared record that supports future reviews and change management. Scheduling engagement activities at key stages such as project intake, before testing, and before deployment helps maintain continuity without creating unnecessary delays.

Disagreement between stakeholders is a normal part of AI governance because different groups balance competing priorities such as safety, cost, speed, and regulatory compliance. Organizations should document the trade-offs that inform key decisions. Decision records should capture what was decided, why the decision was made, the risks that remain, and the controls required to manage them. When unresolved concerns cannot be resolved for a high-impact system, organizations should consider limiting the system's scope, introducing additional human oversight, or conducting a pilot before wider deployment. Documenting these decisions supports transparency, provides valuable context for future updates, and ensures that stakeholder engagement produces clear decisions, defined responsibilities, and documented governance outcomes.

Testing, monitoring, and operationalizing AI systems

This section focuses on the governance activities that support AI systems from pre-deployment testing through ongoing operation. Testing and validation establish whether a system is ready for deployment, while monitoring, documentation, and communication help ensure it continues to operate safely, reliably, and transparently over time. These practices enable organizations to identify emerging risks, maintain accountability, and support the responsible operation of AI systems throughout their lifecycle. By the end of this section, you will be able to design a testing approach that covers performance, safety, fairness, and operational readiness, establish monitoring and change control practices, maintain documentation that supports accountability, and develop a communication plan that reduces misuse and supports transparency.

Testing and validating AI system behavior

Testing evaluates whether an AI system behaves as intended, while validation determines whether it is ready for deployment in a real-world environment. Strong model performance alone is not sufficient. AI systems operate within broader socio-technical environments, so testing should extend beyond the model itself to include the data pipeline, user interface, supporting workflows, and the human oversight processes that govern system use. Validation confirms that appropriate controls, monitoring mechanisms, and escalation procedures are in place before deployment. For AI systems that influence high-impact decisions, validation should be proportionate to the level of risk and supported by clear documentation.

A common mistake is to focus testing solely on model accuracy while overlooking potential harms. Effective testing considers how the system performs under realistic operating conditions, including whether it could expose sensitive information, produce unfair outcomes, generate unsafe outputs, or encourage excessive reliance by users. It should also evaluate how the system responds to uncertain, incomplete, or unexpected inputs, since these conditions often reveal weaknesses that are not apparent during controlled testing. Designing tests that reflect real-world operating conditions helps organizations identify risks before deployment and build confidence in the system's reliability.

Assessing baseline performance, robustness, and stability

Baseline performance testing evaluates whether an AI system meets its defined success metrics using representative data. Robustness testing assesses how the system performs when operating conditions change, such as differences in phrasing, user populations, or incomplete records. Stability testing determines whether results remain consistent across repeated runs and normal variations in input. For large language model (LLM) systems, this also includes evaluating prompt sensitivity to determine whether small changes in prompts produce materially different or potentially harmful outputs. Consistent performance across these dimensions is essential because variability can lead to inconsistent decisions, particularly in high-impact applications where reliability directly affects safety, fairness, and trustworthiness.

Testing for fairness and safety

Fairness testing begins by defining what fairness means for the specific decision context. Organizations should select measures that reflect the intended use of the AI system and the potential harms associated with its decisions. Testing should evaluate whether error rates or outcomes differ across relevant groups, whether those differences result in unfair treatment, and whether the system reinforces historical bias present in training or operational data. When fairness issues are identified, organizations should revisit the underlying data, features, decision thresholds, and human oversight processes. Documenting these evaluations also helps ensure that governance decisions remain transparent and explainable.

Safety testing evaluates whether the AI system could produce harmful recommendations, unsafe instructions, or misleading outputs under realistic operating conditions. It should also evaluate foreseeable misuse, particularly for prompt-based systems, by simulating adversarial inputs, attempts to bypass safeguards, and challenging edge cases. An equally important consideration is how the system behaves when it cannot respond safely, since refusing or limiting responses is often an essential safety control. Testing should be repeatable and based on representative scenarios, allowing organizations to evaluate whether safety controls perform consistently under real-world conditions.

Validating operational readiness

Validation extends beyond confirming that an AI system performs accurately. Before deployment, organizations should verify that ownership is clearly assigned, monitoring processes are in place, escalation routes are defined, and incident response procedures are operational. Validation should also confirm that personnel understand when human intervention is required, how exceptions should be documented, and whether logging provides sufficient information to investigate issues after deployment.

Operational readiness also depends on ensuring that governance policies are reflected in day-to-day workflows. Staff responsible for using or overseeing the system should receive appropriate training, and human oversight procedures should be practical, clearly defined, and consistently applied. A technically accurate system may still be unsuitable for deployment if monitoring responsibilities are unclear or users are likely to rely on its outputs without appropriate review. Readiness is best demonstrated through activities such as pilots, operational drills, and formal sign-offs that verify governance controls under realistic conditions.

Managing and monitoring AI systems post-launch

Deployment marks the beginning of an AI system's operational lifecycle. Once a system is in production, continuous monitoring is essential to identify model drift, misuse, emerging harms, and declining performance before they affect users or business outcomes. Without effective monitoring, organizations often become aware of problems only after customer complaints, operational failures, or regulatory scrutiny.

Monitoring also helps identify where workflows, controls, or system behavior require refinement. Because AI systems operate in dynamic environments, changes in data, user behavior, business processes, or external dependencies can alter performance over time. This applies to both predictive models and generative AI systems, making ongoing monitoring a fundamental governance responsibility throughout the operational lifecycle.

A mature monitoring program extends beyond model performance to include safety, fairness, privacy, and operational reliability. Depending on the use case, organizations may monitor

complaint trends, human override rates, escalation frequencies, latency, availability, or unusual spikes in high-risk prompts. Monitoring should also identify whether outputs expose sensitive information, generate prohibited content, or produce materially different outcomes across relevant groups. The metrics selected should reflect the specific risks associated with the AI system and provide meaningful insight into its operational performance.

One of the primary objectives of monitoring is to detect drift. Drift occurs when real-world conditions diverge from those observed during development, causing model behavior to change over time. It may result from changes in input data, business processes, user behavior, or, in the case of LLM systems, vendor updates, prompt modifications, or new retrieval content. Organizations should establish thresholds that trigger investigation or retraining, such as material declines in performance or increases in unsafe outputs. Because drift is an expected characteristic of operational AI systems, governance processes should treat it as a routine operational concern rather than an exception.

Effective monitoring also depends on clear operational ownership. Organizations should define who is responsible for reviewing monitoring dashboards, investigating anomalies, authorizing system pauses, and escalating incidents when predefined thresholds are exceeded. Change control procedures should ensure that updates including data revisions, prompt modifications, model replacements, and vendor version changes are reviewed, tested, and documented before deployment. Treating major updates as controlled releases helps maintain system stability while reducing operational risk.

Effective incident management spans monitoring, documentation, and communication, each covering a distinct responsibility. AI incidents may involve security breaches, harmful outputs, unfair outcomes, unsafe recommendations, or other unexpected behaviours. An incident response plan should define how harm will be contained, how stakeholders will be informed, and how evidence will be collected through appropriate logging and documentation. Following an incident, organizations should conduct a root cause analysis and use the findings to improve testing, safeguards, documentation, and operational procedures.

Continuous improvement should follow a structured, deliberate process. Regular reviews of monitoring data, scheduled model evaluations, stakeholder feedback, and periodic updates to documentation help ensure that AI systems continue to operate safely as conditions evolve. For LLM-based systems, improvements may include refining prompts, improving retrieval mechanisms, or strengthening refusal behaviors. A structured improvement process enables organizations to reduce risk while maintaining system performance and avoiding emergency changes driven by preventable issues.

Maintaining documentation for AI operations

Documentation provides the record of how an AI system was designed, evaluated, deployed, and governed throughout its lifecycle. As AI systems evolve, team members change, and operational requirements shift, documented decisions help preserve continuity, make decisions traceable, and enable effective troubleshooting. Documentation also allows organizations to explain a system's purpose, the data it uses, the safeguards that have been implemented, and the reasoning behind key governance decisions. Without these records, responding to incidents, audits, or stakeholder questions becomes more difficult and less reliable.

At a minimum, documentation should describe the system's purpose, intended users, and operational boundaries. It should identify the data used by the system, its sources, and any restrictions on its use. Documentation should also record design decisions, including why a particular model approach was selected, the trade-offs that were considered, testing outcomes, known limitations, human oversight arrangements, monitoring plans, incident response procedures, and any approvals or exceptions granted during development. Maintaining these records supports consistent governance as the system evolves.

Documentation is most valuable when it is designed for the people who rely on it in day-to-day operations. Engineers require information about system dependencies, update procedures, and monitoring thresholds, while operations teams need runbooks that define incident response procedures and escalation paths. Product teams benefit from clear statements of intended use and system boundaries, and support teams need guidance on responding to customer concerns and escalating issues appropriately. Legal and compliance teams rely on documentation that explains data handling practices, transparency commitments, and mechanisms for redress. Documentation that reflects the needs of its intended audience is more likely to remain accurate, relevant, and actively used.

One of the most valuable governance artefacts is the decision record. A decision record captures what was decided, why the decision was made, the risks that remained, and the safeguards required to manage those risks. It provides continuity as teams evolve, reduces the need to revisit previously resolved decisions, and supports incident investigations by documenting the assumptions that informed earlier choices. Decision records also strengthen audit readiness by demonstrating that governance decisions are supported by a structured and well-documented process.

Documentation should be maintained as a living resource not a one-time project deliverable. Changes such as new data sources, prompt revisions, model updates, additional user groups, or changes to retrieval mechanisms should trigger a review and update of the relevant documentation. Incidents should also result in documentation updates so that lessons learned

are incorporated into future governance practices. Standardized templates and documentation requirements integrated into release processes help organizations maintain consistency while reducing administrative effort.

Documentation also supports organizational transparency by providing a clear record of governance decisions, system limitations, approvals, and accountability from development through operation. It enables leaders to understand the organization's risk posture, helps operational teams explain system behavior, and provides evidence that governance activities have been carried out as intended. Well-maintained documentation therefore serves as a practical control that makes governance decisions traceable, operations consistent, and trust in AI systems defensible.

Developing a communication plan

A communication plan explains what an AI system is designed to do, its limitations, and how it should be used safely. Communication is an important governance control because misuse often results from misunderstanding. If users overestimate a system's capabilities or are unaware of its limitations, they may rely on it in inappropriate or unsafe ways. Clear communication helps establish realistic expectations and promotes responsible use.

Communication should address both internal audiences, such as employees and leadership, and external audiences, including customers or other directly affected parties. It should also define how information will be communicated during incidents to maintain transparency and confidence. Providing timely, accurate information reduces confusion and helps organizations respond consistently when issues arise.

Different audiences require different types of information. End users need clear guidance on appropriate use, system limitations, and when to seek human review. Operators require information about monitoring, escalation procedures, and operational controls, while leadership needs a concise understanding of system readiness, governance measures, and overall risk. Legal and compliance teams require information about disclosures, data handling practices, and mechanisms for redress, and customer support teams need clear procedures for responding to questions and complaints. Tailoring communication to each audience helps ensure that information is relevant and actionable.

System limitations should be communicated through clear, practical guidance. For example, users should understand when AI-generated content requires human review, when sensitive information should not be entered into the system, and when the AI system should not be relied upon for professional advice in areas such as healthcare or financial services. Communicating limitations in terms of expected user actions makes them easier to understand and more likely to be followed.

Organizations should also be transparent about when users are interacting with AI and what that means for their experience. Disclosures help users interpret AI-generated outputs appropriately, reduce the risk of deception, and clarify when human oversight remains part of the decision-making process. Where AI systems rely on retrieval or external knowledge sources, users should also understand that responses may reflect limitations or inaccuracies in those sources. Consistent and visible disclosures support transparency and are often expected in higher-risk applications.

Communication plans should also define how incidents and system changes will be communicated. This includes identifying who is responsible for communicating with different audiences, what information should be shared, and how updates such as new limitations, revised workflows, or temporary service interruptions will be announced. Effective communication during incidents should focus on explaining what occurred, the actions taken to reduce harm, and the next steps, supported by evidence from monitoring and documentation.

Finally, communication should incorporate mechanisms for ongoing feedback. User reports, support requests, and structured reviews help identify usability issues, safety concerns, fairness risks, and privacy issues that may not have been identified during development. Reviewing this feedback enables organizations to prioritize corrective actions, improve documentation and training, and refine governance practices over time. Feedback loops not only strengthen AI systems but also demonstrate that user concerns are taken seriously and contribute to continual improvement. In the next section, we will turn to data governance across the AI pipeline, from the foundational questions teams should ask about data sources through to privacy-enhancing technologies that reduce personal data exposure without sacrificing utility.

Governing data across the AI pipeline

This section focuses on the data governance as a continuous practice throughout the AI development lifecycle, from sourcing and collection through preparation, transformation, and delivery to the model. Decisions made at each stage influence the quality, fairness, privacy, and reliability of AI systems, and many of these risks cannot be corrected through model selection alone. The section explores how to evaluate data throughout the pipeline, maintain data quality and traceability, and apply governance practices that support responsible AI development. By the end of this section, you will be able to apply data governance questions to any AI project, trace data lineage and provenance, identify quality and format risks during gathering, manage data preparation using the five Vs, and select appropriate privacy-enhancing technologies for a given development context.

Applying data governance questions

Data is often where AI governance risks first emerge. AI systems learn from data, and in many cases expose information through their outputs, making data governance fundamental to building trustworthy systems. Asking structured governance questions helps organizations understand what data is being used, why it is needed, and what constraints apply. These questions also help prevent scope creep, since introducing new data sources frequently expands how an AI system can be used. For LLM systems, data governance extends beyond training data to include retrieval content, prompts, and logging practices.

The first questions should focus on purpose and necessity. Organizations should establish what decision the data supports, whether the data is required for that purpose, and whether additional data is being collected simply because it is available. Collecting unnecessary information increases privacy and security risks without necessarily improving system performance. Applying the principles of **data minimization**, which limits data collection to what is necessary, and **purpose limitation**, which prevents data from being reused for unrelated purposes without review, helps reduce risk while keeping development aligned with its intended objectives. These considerations are particularly important for high-impact AI systems, where the rationale for using specific data should be clearly justified.

Organizations should also understand where data originates and what conditions govern its use. Data may be collected internally, obtained from third parties, purchased from external providers, or generated by users, and each source may carry different contractual, legal, or ethical obligations. Governance should also consider whether the data contains sensitive information, whether its use is appropriate for the intended purpose, and whether the organization can demonstrate its provenance. Data that is quietly repurposed for model training or lacks a clear origin introduces legal, quality, and trust risks that should be resolved before development continues.

Data governance also requires evaluating whether data is fit for its intended purpose. This includes assessing its accuracy, completeness, timeliness, and representativeness of the population affected by the AI system. Missing or poorly labeled data may reduce model performance for particular groups, creating fairness concerns, while historical data may embed existing biases that influence future decisions. Data quality issues can also create safety risks, such as inaccurate medical records or outdated customer information. Profiling datasets and evaluating potential bias before training help identify these issues early, when they are generally easier to address than after deployment.

Access and retention should be governed as carefully as data quality. Organizations should define who can access data, who can export it, how access is monitored, and how long information is retained. For LLM-based systems, governance should also address what user

inputs are logged and whether those logs contain sensitive information. Logging supports monitoring, auditing, and incident investigation, but it also introduces privacy considerations that require appropriate safeguards. Applying the principle of least privilege and maintaining clear separation between development, testing, and production environments help reduce the likelihood and impact of unauthorized access.

Data governance does not end once model training is complete. It also applies to evaluation datasets, monitoring data, retrieval content, periodic model updates, and the processes used to correct data quality issues over time. Governance should consider how changes to data sources affect downstream systems and establish mechanisms to detect unintended consequences when those changes occur. Integrating data governance with change management and monitoring helps maintain data quality, system reliability, and accountability as AI systems evolve.

Tracing data lineage and provenance

Data lineage and data provenance provide the traceability needed to understand how data moves through an AI system. **Data lineage** describes how data moves through a system and the transformations applied to it, while **data provenance** explains where the data originated, how it was collected, and why it is appropriate for its intended use. Together, they help organizations understand the data that influences AI systems, make AI outcomes defensible, and reduce the risks associated with using unknown or inappropriate data. They also make it easier to investigate incidents, explain model behavior, and demonstrate that data has been handled responsibly.

Traceability is particularly important because AI systems can amplify relatively small data issues into operational or governance risks. Mislabeled records, biased annotations, incomplete datasets, or unnecessary sensitive attributes can all influence model behavior and affect decisions at scale. Maintaining lineage and provenance allows organizations to identify where these issues originated, determine whether the data was appropriate for the intended purpose, and explain how it influenced system outcomes.

In enterprise environments, data typically passes through multiple teams, platforms, and processing stages before it reaches an AI model. Each handoff introduces opportunities for errors, undocumented changes, or inappropriate reuse. Lineage provides visibility into these transitions, while provenance demonstrates that data was collected, shared, and used in accordance with applicable permissions and governance requirements. Together, they improve both operational resilience and auditability.

Data provenance records information about the origin and suitability of data. This includes where the data came from, how it was collected, the permissions or contractual restrictions that apply, and the purpose for which it was originally gathered. Provenance should also

capture whether data was generated internally, provided by customers, obtained from third parties, purchased, or collected through partnerships, together with any known quality limitations, missing populations, or uncertainty in labels. For training datasets, provenance should document the presence of sensitive information and any data minimization decisions. For retrieval-based systems, it should also record the authorship and currency of the source content. Well-documented provenance helps organizations justify why a dataset is appropriate for a particular AI application.

Data lineage documents how data changes throughout the development pipeline. This includes activities such as cleaning, deduplication, filtering, normalization, feature engineering, joins, sampling, and movement between storage environments. Lineage should also record dataset versions used for training, evaluation, and monitoring, together with the relevant time periods, since changes over time may affect model performance. For large LLM systems, lineage also extends to prompt templates, retrieval indexes, and updates to knowledge bases. Maintaining this history enables teams to reconstruct the sequence of events that influenced model behavior and identify the source of unexpected changes after deployment.

Consider a retail organization developing a demand forecasting model using sales and promotional data. After deployment, the model begins overpredicting demand in one region, resulting in excess inventory and operational disruption. Lineage analysis reveals that a new promotion table was joined incorrectly, creating duplicate records, while provenance shows that an external partner changed the format of its data feed without notice, introducing missing values that were automatically filled during processing. Without lineage and provenance, the organization might retrain the model repeatedly without addressing the underlying data issue. With appropriate traceability, the root cause can be identified and corrected quickly, reducing both operational disruption and governance risk.

Maintaining lineage and provenance should become part of routine engineering practice not a manual documentation exercise. Organizations can support this by using dataset versioning, standardized naming conventions, automated transformation logs, and data inventories that link AI models to the datasets used during development. Governance processes should also define who approves new data sources, who reviews changes, and what events such as schema changes, new data sources, or expanded use cases trigger reassessment. Integrating these practices into normal development workflows helps ensure that traceability remains consistent as AI systems evolve.

Gathering data with quality and format controls

Data gathering is the stage where many downstream AI risks are either prevented or introduced. Decisions about what data to collect, how it is collected, and how consistently it is managed have a lasting impact on model performance and governance. No model can

compensate for data that is inaccurate, incomplete, or unsuitable for its intended purpose. Effective data governance therefore begins before model development, with careful attention to data quality, representativeness, and the formats in which information is collected and managed.

Data quality extends beyond accuracy. Organizations should also evaluate completeness, timeliness, consistency, relevance, representativeness, and the quality of any labels used for supervised learning. Missing values concentrated within particular groups, outdated information, inconsistent field definitions, or biased labels can all influence model behavior and create fairness or safety risks. Assessing these characteristics early helps identify issues before they become embedded in AI systems, where they are often more difficult to correct.

Even structured data requires careful evaluation. Although tables and databases appear well organized, identical fields may have different meanings across business systems, and operational data may lack the context required for AI development. Structured datasets can also contain proxy variables for sensitive attributes or reflect sampling decisions that underrepresent important populations. Another common issue is data leakage, where information unavailable during real-world decision-making is inadvertently included during training, resulting in overly optimistic performance. Reviewing data definitions, collection methods, and sampling strategies helps reduce these risks before model development begins.

Unstructured data introduces additional governance challenges, particularly for LLM systems. Text may contain personal information, confidential business data, or sensitive contextual details that were never intended for AI development. It can also include ambiguity, cultural nuance, sarcasm, or inconsistent language that affects interpretation. Operational logs may reveal behavioral patterns that raise privacy concerns, while duplicated documents or template-based responses can bias model outputs. For retrieval-augmented systems, the quality of source documents is equally important because generated responses depend directly on the information retrieved. Filtering, redaction, and governance controls are therefore essential before unstructured data enters the AI pipeline.

Data formats also influence governance by determining how easily information can be validated, traced, and reused. Standardized schemas support automated quality checks, consistent access controls, retention policies, and audit logging, while inconsistent formats increase the likelihood of manual errors and integration problems. Metadata including timestamps, authorship, document sources, and access permissions, is particularly important for text-based systems because it supports traceability and governance. For image and audio data, format choices can affect both data quality and model performance. These decisions should therefore be treated as governance considerations as well as technical implementation choices.

A structured data gathering process begins by identifying the minimum data required to achieve the intended purpose. Candidate data sources should then be evaluated for permissions, provenance, quality, and long-term stability before profiling is performed to identify missing values, duplication, outliers, and population coverage. Organizations should document data definitions, labeling methods, and known limitations before deciding whether to proceed, narrow the scope, or collect additional data. For LLM-based systems, this process should also define what information must not be included in prompts, training datasets, or retrieval indexes. Establishing these controls before development helps reduce downstream risk, supports governance reviews, and provides a stronger foundation for building AI systems that are fit for purpose.

Wrangling and preparing data using the five V's

The five Vs provide a practical framework for understanding the challenges of preparing data for AI systems. The framework encourages teams to evaluate the characteristics of their data and the governance controls needed to manage associated risks. The five dimensions volume, variety, velocity, veracity, and value help guide decisions about data collection, storage, processing, access, and quality throughout the AI pipeline.

Volume refers to the amount of data being processed and the scale at which systems operate. As data volumes increase, so do operational, privacy, and governance risks. Large datasets place greater demands on processing pipelines, increase storage costs, and expand the amount of information that must be protected. Applying sampling strategies, dataset versioning, and automated quality checks alongside the data minimization principles covered earlier helps organizations manage scale while maintaining visibility into data quality. Large-scale pipelines should also be tested under realistic operating conditions that reflect production workloads and data volumes.

Variety reflects the diversity of data formats and sources, including structured tables, text, images, logs, and other unstructured content. Combining these sources increases the complexity of governance because each may have different meanings, permissions, and quality characteristics. Clear schemas, documented transformation rules, and validation of joins and merges help preserve data quality and reduce the risk of introducing bias or inconsistencies. For LLM systems, variety also includes differences in language, writing style, and document structure that may influence model behavior.

Velocity describes how frequently data is generated, updated, or replaced. High-velocity environments require governance processes that can respond to changing data without compromising quality. Organizations should define update schedules, establish criteria for retraining or reassessment, and continuously monitor changes in data distributions and model performance. Seasonal trends, evolving user behavior, and product changes can all alter the characteristics of incoming data, making clear ownership and review thresholds essential.

Veracity concerns the reliability and integrity of data. Errors introduced through manual data entry, inconsistent definitions, inaccurate labels, duplicated records, or data leakage can affect AI system performance. Data profiling, validation checks, and regular reviews of label quality help identify these issues before development progresses. For retrieval-based LLM systems, veracity also depends on whether source documents remain accurate, authoritative, and current. Reliable data is fundamental to producing trustworthy AI outputs.

Value focuses on whether a dataset meaningfully improves the AI system while remaining consistent with its intended purpose. Additional data does not always produce better outcomes and may increase privacy, security, or governance risks without delivering corresponding benefits. Organizations should evaluate the incremental value of new data sources, balancing improvements in performance against additional complexity or risk. Documenting these decisions also supports principles such as data minimization and purpose limitation by demonstrating why particular datasets were included in the system.

Applying privacy-enhancing technologies in data preparation

Data preparation is the stage where organizations determine what information an AI system will { XE "AI pipeline data:**privacy-enhancing technologies (PETs)**, applying" }use and how that information will be protected. Decisions made during preparation often have a greater impact on privacy risk than later technical controls. Personal data that is not required for the intended purpose should be removed or transformed before it enters training datasets or retrieval pipelines, while necessary sensitive information should be subject to appropriate access controls and documented governance decisions. Data preparation should also address retention, ensuring that information is not stored longer than necessary. For LLM systems, these decisions extend to defining what users may submit as input and what information the system is permitted to retrieve. Integrating privacy into data preparation ensures that privacy protections are embedded throughout AI development.

An effective preparation process begins by removing data that is unnecessary for the intended purpose. Where possible, direct identifiers such as names or account numbers should be removed or transformed, and data should be standardized carefully to avoid creating unintended proxy variables for sensitive attributes. Organizations should also check for duplication, data leakage, and contamination between training, validation, and test datasets. For unstructured text, preparation should include redaction and automated scanning for sensitive information. Governance should also consider whether outputs or operational logs may contain personal data and apply appropriate safeguards before the system enters production.

Privacy-enhancing technologies (PETs) reduce the exposure of personal information while preserving { XE "**privacy-enhancing technologies (PETs)**" }the usefulness of data for analysis

or AI development. Although these techniques can reduce privacy risk, they complement broader governance practices. Organizations should evaluate whether privacy-enhancing methods remain appropriate for the intended use case, whether they introduce unintended effects on fairness or model performance, and how their application should be documented as part of the AI development process.

Several privacy-enhancing approaches can support AI development. **Pseudonymization** replaces direct identifiers with alternative values that reduce the likelihood of identifying individuals. **Aggregation** summarizes information so that analysis can be performed without relying on individual-level records. **Differential privacy** techniques help reduce the risk of re-identification when statistical information or model outputs are shared. **Secure access controls** restrict access to raw datasets and support controlled processing environments, while **synthetic data** can sometimes enable development with reduced exposure to personal information, provided that it is evaluated for realism, bias, and suitability. For LLM-based systems, privacy controls may also include restricting retrieval scope and applying redaction before retrieved content is passed to the model. The choice of technique should reflect the purpose of the AI system and the level of privacy risk being addressed.

Organizations { XE "AI pipeline data:**privacy-enhancing technologies (PETs)**, applying" } should also recognize the limitations of privacy-enhancing technologies. Removing direct identifiers does not necessarily eliminate the possibility of re-identification if sufficient contextual information remains. Likewise, transformed or synthetic datasets should be evaluated to ensure they do not introduce new biases or disproportionately affect model performance for particular groups. Privacy-enhancing methods should not become a justification for collecting more data than necessary, and their use does not eliminate the need for ongoing monitoring and governance. Selecting an appropriate technique therefore requires balancing privacy, utility, fairness, explainability, and operational complexity while documenting the rationale for those decisions.

Consider an employee support assistant designed to answer questions about workplace benefits and organizational policies. Historical support tickets could improve the system's responses, but they also contain personal information and sensitive employee details. A more privacy-preserving approach is to collect only the information required for the intended purpose, redact sensitive content, and protect access to any retained records. The organization could build its retrieval system using curated policy documents, supplementing them with de-identified or summarized ticket data only where necessary to improve performance. Monitoring should also identify outputs that inadvertently reveal personal information { XE "AI pipeline data:**privacy-enhancing technologies (PETs)**, applying" }so that corrective action can be taken promptly. This example illustrates how privacy-enhancing techniques, combined with appropriate retrieval design and governance controls, reduce privacy risks while preserving the usefulness of the AI system.

Governing model features and conformity documentation

This section focuses on two governance activities that directly influence how AI systems are built and assessed: feature design and development documentation. Decisions about which features are included in a model shape its behavior and can introduce or reduce risks related to fairness, privacy, and security. Documentation, including model cards and conformity records, provides the evidence needed to explain how the system was developed, evaluated, and governed. Both activities make AI systems more transparent, auditable, and governable in practice. By the end of this section, you will be able to evaluate features for governance risk, identify inappropriate proxies, and produce model cards and conformity documentation that support safe deployment and sustained oversight.

Designing features for responsible model development

Features are the inputs an AI model uses to generate predictions or outputs. Feature engineering involves selecting, transforming, and combining those inputs so they accurately represent the problem being solved. Because features determine the information available to a model, they also influence the risks a system may introduce. Poor feature selection can embed historical bias, expose sensitive information, or rely on variables that act as proxies for protected characteristics. Responsible feature engineering therefore involves governance decisions as well as technical design.

Feature selection should begin with the intended purpose of the AI system. Features should have a clear relationship to the decision being supported and be appropriate for that context. For example, a fraud detection model may rely on transaction patterns and unusual behavioural signals, while a customer support prioritization model may use indicators of issue severity and service history. Features should not be included solely because they improve predictive performance if they introduce disproportionate privacy, fairness, or governance risks. Organizations should also confirm that features are available at the point when decisions are made to avoid data leakage and document the rationale for including each category of features.

Bias can emerge when features reflect historical inequalities or capture social patterns that disadvantage particular groups. It may also result from inconsistent data collection, differences in measurement across populations, or biased labels inherited from earlier decision-making processes. Feature engineering can unintentionally strengthen these biases by creating variables that closely approximate protected characteristics. Responsible development therefore includes evaluating how features influence model outcomes, identifying inappropriate proxy variables, and assessing whether observed differences are

justified within the decision context. Where unacceptable disparities are identified, organizations may revise feature selection, introduce fairness constraints, or incorporate additional human oversight.

Interpretability also depends on thoughtful feature design. Regardless of the model architecture, organizations should be able to explain the purpose of features and how they contribute to system behavior. Complex engineered features should be documented clearly, including the transformations used to create them and the reasons they improve the model. Features that change over time should also be monitored, since evolving values may alter model behavior and require updated explanations. Clear feature documentation supports transparency during governance reviews and provides valuable context during incident investigations.

Feature design should also address privacy and security risks. Sensitive information may be exposed through direct identifiers, rare combinations of attributes, or highly detailed location and time data that make individuals easier to identify. Features may also become targets for adversarial manipulation if attackers can influence the values used by the model. Appropriate safeguards include minimizing unnecessary detail, validating feature values, restricting access to sensitive information, and applying appropriate access controls throughout development. For LLM systems, governance should also extend to metadata used for retrieval, routing, or policy enforcement, since these elements can influence system behavior if they are altered or exposed.

Consider an insurance company developing a claims triage model to identify claims requiring further investigation. Using neighborhood-level indicators may improve predictive performance but could also act as a proxy for protected characteristics, increasing the risk of unfair outcomes. A more appropriate approach is to rely on documented behavioral indicators, claim consistency signals, and validated risk factors that are directly relevant to the decision being made. The organization should evaluate whether investigation rates differ across relevant groups, determine whether any disparities can be justified, and ensure that investigators understand the model's limitations and apply appropriate human judgment when interpreting model outputs. This example illustrates how feature selection influences fairness, transparency, and accountability throughout the AI development process.

Creating model cards and conformity documentation

A **model card** is a structured summary that describes an AI model, its intended purpose, appropriate use cases, performance characteristics, limitations, and known risks. It provides a consistent way to communicate essential information to developers, operators, auditors, and decision-makers without requiring them to understand the underlying implementation. **Conformity documentation** complements the model card by providing the broader evidence that the AI system was developed, tested, and approved in accordance with established

governance requirements. Both documents are essential governance tools that help teams, auditors, and decision-makers evaluate and operate AI systems responsibly.

A model card should clearly describe the model's intended purpose, target users, and the decisions it is designed to support. It should summarize the data used during development, including relevant provenance information and any important restrictions. Evaluation results, performance characteristics, fairness assessments, known limitations, safety considerations, security controls, monitoring expectations, and prohibited uses should also be documented. A model card serves as a practical guide that helps teams deploy and operate the model safely.

One of the most valuable aspects of a model card is its explicit documentation of known limitations and failure modes. These may include reduced performance for particular groups, sensitivity to certain inputs, known sources of error, or situations in which the model should not be relied upon. For LLM systems, limitations may include variability across prompts, hallucinations, or dependence on retrieved content, while predictive models may require warnings about concept drift or changing operating conditions. Clearly documenting these limitations enables users to apply appropriate safeguards and helps reviewers determine whether governance controls remain proportionate to the risks.

Conformity documentation provides the evidence that an AI system has been evaluated and approved according to the organization's governance process. It may include project intake records, impact assessments, testing reports, security and privacy reviews, approval records, monitoring plans, incident response procedures, user training records, and change histories. The objective is not to create unnecessary documentation but to maintain a defensible record of the decisions, evaluations, and approvals that supported deployment. Consistent conformity documentation also helps organizations apply governance practices uniformly across multiple AI projects.

Model cards and conformity documentation continue to provide value after deployment. During incidents, they allow teams to review the intended use of the system, known limitations, monitoring thresholds, and previously identified risks. They also support onboarding by helping new team members understand the system, assist leadership in evaluating governance readiness, and provide auditors with evidence of testing and oversight. When organizations adopt third-party AI models, similar documentation can support procurement decisions and vendor governance by enabling meaningful comparisons between available models.

Because AI systems evolve over time, model cards and conformity documentation should be maintained as living documents. Changes including new training data, updated prompts, revised retrieval content, model replacements, or expanded use cases should trigger a review and update of the relevant documentation. Incidents should also result in revised documentation so that lessons learned are incorporated into future governance practices.

Integrating documentation updates into release processes and scheduling periodic reviews for high-impact systems helps ensure that governance records remain accurate, current, and aligned with the operational reality of the AI system.

Summary

In this chapter, we explored AI development governance as a lifecycle practice, tracing how responsible decisions at each stage create the conditions for safe and accountable AI systems.

We began by mapping the full AI development lifecycle from problem framing through retirement, establishing that AI failures often arise from decisions made across the lifecycle rather than from individual model components. We then turned to the planning and design decisions that shape responsible development from the outset, including defining business problems and use cases with clear decision boundaries and measurable outcomes, determining system scope by identifying both owned components and external dependencies, conducting impact assessments that connect harms to specific controls, and engaging stakeholders early to identify risks and inform system design.

From planning, we moved into the build and operational phases, examining testing that covers robustness, fairness, safety, and misuse alongside model performance, and validation that confirms governance controls, human oversight, and operational readiness before deployment. We also explored how monitoring, change control, and incident response support the safe operation of AI systems over time, and how documentation and communication promote transparency, accountability, and effective governance throughout the system lifecycle.

We then turned to data governance across the full pipeline, considering foundational questions about purpose, necessity, provenance, and data access, and tracing data through lineage records that make AI outcomes defensible and incidents easier to investigate. We also looked at how quality dimensions and format choices during data gathering influence downstream fairness and reliability, applied the five Vs framework to data wrangling and preparation, and discussed privacy-enhancing technologies that reduce personal data exposure without sacrificing modelling utility. Finally, we considered feature engineering as a governance decision, showing how proxy risks and interpretability requirements influence responsible model development, before concluding with model cards and conformity documentation that make development decisions auditable and defensible. In *Chapter 5, Governing AI in Production*, we shift from building AI systems to operating them safely in production.

5

Governing AI in Production

Chapter 4, Governing AI Development, focused on the decisions, controls, and documentation needed before an AI system reaches users. This chapter follows the system into production, where governance becomes an ongoing operational responsibility. Real users interact with the system in unpredictable ways, production data differs from testing data, and models operate in environments that continually change. Controls that appeared sufficient during development must now withstand real-world conditions. Organizations often discover gaps in governance only after an incident exposes them.

The chapter begins with the policies and readiness criteria that support a safe production launch. It then explores how to govern the proprietary models, vendors, and third-party services that many AI systems depend on. From there, it covers the practices needed to operate AI systems safely, including monitoring, assessment, and incident response. The chapter then looks at the accountability and transparency measures that support oversight after deployment. It concludes with the deployment architectures and model patterns, including **retrieval-augmented generation** (**RAG**) and agentic AI, that influence how AI systems are governed throughout their operational life.

In this chapter, we will cover the following topics:

- Establishing deployment readiness
- Governing vendor and third-party dependencies
- Monitoring and responding to operational risk
- Ensuring accountability and transparency
- Selecting deployment architectures and model patterns

Establishing deployment readiness

Before an AI system enters production, organizations need more than a working model. They need policies, controls, and operational processes that reflect how the system will be used by real users and data. This section focuses on privacy, security, technical, operational, human, and compliance requirements that should be in place before launch. By the end of this section, you will be able to identify policy gaps that increase deployment risk and define the readiness criteria for a production system.

Evaluating and updating privacy and security policies

Moving an AI system into production introduces new risks that existing privacy and security policies may not address. Prompt inputs, model outputs, retrieval mechanisms, and vendor-hosted services create data flows and operational behaviors that differ from those of traditional software. If policies do not account for these patterns, teams often fill the gaps with inconsistent practices. Updating policies is less about writing new documentation and more about ensuring that common failure points are addressed before the system goes live.

Begin by reviewing the data the system will process in production. Confirm that users understand what information is collected, why it is collected, and that data collection is limited to what the use case requires. Review what is logged, how long logs are retained, who can access them, and whether model outputs could expose personal information. If the system uses retrieval, verify that access is restricted to the appropriate permissions. Also, establish a clear process for handling complaints or requests to correct or challenge how personal data has been used. These checks translate privacy principles into day-to-day operational controls.

Security policies should define who can access the system, how identities are verified, and how misuse is detected. Review whether role-based access controls are in place and whether sensitive actions require stronger authentication. Confirm that API keys, credentials, and connections to internal systems are protected throughout their lifecycle. Input validation should reduce the risk of prompt injection and other malicious inputs, while output controls should prevent sensitive information from being exposed through logs, support tickets, or external communications. Policies should also address how vulnerabilities in third-party and open-source components are identified and managed. Together, these controls reduce the system's attack surface and support secure operation.

Acceptable use policies are equally important because they define how employees and other users are expected to interact with AI systems. They should identify the types of information that must never be entered into external AI services, the approval process for exceptions, and the situations that require human review or escalation. Customer-facing systems may require human approval before responses are shared, while internal systems may require additional

safeguards for high-impact tasks. These policies should also make it clear that responsibility for decisions remains with people, even when AI supports the process. Clear expectations reduce unsafe shortcuts, particularly when teams are working under time pressure.

Policies are only effective if they reflect how work is performed. A requirement to review every AI-generated response has little value if there are no resources to carry out those reviews. Similarly, strict logging requirements are difficult to enforce if the underlying platform records data by default. Review operational workflows alongside policy requirements to confirm that teams can comply without disrupting essential business processes. Where gaps exist, organizations may need to adjust workflows, narrow the system's scope, or invest in additional resources to support compliance.

Policies should also map directly to measurable controls. Access controls, log retention periods, review thresholds, and deployment gates can all be verified through testing or audit. Supporting policies with checklists, templates, monitoring, and periodic reviews helps teams apply them consistently across projects. Measurable controls also provide evidence that governance decisions have been implemented and maintained. Policies without enforcement offer little protection when incidents occur.

Privacy and security policies play a central role in production readiness. They should address the realities of AI systems, align with operational workflows, and define controls that can be verified in practice.

Understanding deployment requirements

A successful deployment depends on more than a model that performs well in testing. The organization must also be prepared to operate, monitor, and support the system in production. Deployment readiness spans technical, operational, human, and compliance considerations, each of which contributes to safe and reliable use. A useful way to assess readiness is to ask what could fail on the first day of production and whether the people, processes, and controls are in place to respond.

Technical readiness begins with the infrastructure that supports the system. Teams should understand where the model runs, how it scales under increased demand, and how it connects to data sources. Those connections should follow the principle of least privilege, granting only the access required for the system to function. Input validation and output controls help prevent malicious or unsafe content from entering or leaving the system. Logging should support troubleshooting without retaining sensitive information unnecessarily, while version control for prompts, models, and retrieval indexes ensures that changes remain traceable. Plans for service interruptions and fallback behavior also help prevent outages from becoming operational failures.

Operational readiness focuses on running the system safely once it is live. Monitoring dashboards, alert thresholds, and clearly defined ownership allow teams to detect and respond to issues quickly. Organizations should define what constitutes an incident, how incidents are escalated, and who is responsible for containment and recovery. Where human oversight is required, staffing and review processes should already be in place. Support teams also need training to identify problematic outputs and respond consistently. Regular reviews, system updates, and clear communication with users and stakeholders help keep the system reliable as models, data, and dependencies evolve.

Human readiness ensures that everyone interacting with the system understands their responsibilities. Users should know the system's capabilities and limitations, when to rely on its outputs, and when to escalate concerns. Reviewers need clear criteria for evaluating outputs, managers should understand when intervention is necessary, and developers must be able to deploy updates without weakening existing controls. Training should reflect the responsibilities of each role and encourage people to report errors, unexpected behavior, and near misses. Strong governance depends as much on informed users as it does on technical safeguards.

Compliance readiness brings together the legal, regulatory, and governance requirements that support production use. Organizations should confirm that disclosures are appropriate for the intended audience, personal data is processed on a valid legal basis where required, and contractual obligations support responsible operation. Deployment decisions should be documented, together with the evidence used to justify them. Processes for complaints, appeals, and redress should already be established, particularly for high-impact use cases. Records should also be maintained so that regulators, auditors, or internal reviewers can verify how decisions were made and how the system has been governed over time. Vendor obligations should form part of this review, particularly where external providers manage models or supporting services.

Many organizations reduce deployment risk through a staged rollout. A pilot or limited release provides an opportunity to observe how the system performs under real operating conditions before expanding its use. Early deployments can be restricted by user group, geography, business function, or data sensitivity, with additional oversight applied until the system demonstrates consistent performance. Evidence gathered during each stage informs decisions about broader deployment and helps identify issues while their impact remains limited.

Deployment readiness extends beyond technical performance. It requires operational processes, trained people, governance controls, and compliance measures that allow the system to operate safely in production. With these foundations in place, the next section explores how to govern the vendors and third-party services that many production AI systems depend on.

Governing vendor and third-party dependencies

Most AI systems depend on external models, cloud platforms, data providers, and other third-party services that organizations do not fully control. Changes to any of these dependencies can affect system behavior, security, compliance, and reliability in production. This section explains how to govern proprietary models, assess and manage third-party risk, evaluate vendor and open-source agreements, and apply these principles in a practical deployment scenario. By the end of this section, you will be able to assess third-party risk, evaluate vendor agreements against operational and governance requirements, and apply appropriate controls to external dependencies.

Governing proprietary AI models

A proprietary model operates outside your direct control and offers limited visibility into how it is built and maintained. It may run on a vendor's platform, evolve over time, and provide little information about its training data or internal design. These characteristics do not make the model inherently unsafe, but they do change how risk is managed. Organizations must assume that uncertainty is part of operating proprietary AI systems and prepare for vendor updates that can alter model behavior with little notice. This is particularly relevant for large language models, where even small changes can affect outputs. Although the vendor develops and maintains the model, responsibility for its use and impact remains with the deploying organization.

Limited visibility affects explainability because organizations cannot always provide detailed technical reasons for individual outputs. Instead, explanations should focus on the available information, such as the data retrieved to answer a request, the rules governing system behavior, and the circumstances in which human review is required. Known limitations should be documented, particularly where the model supports decisions that require explanation. Consistent testing, monitoring, and analysis of failure patterns become more important when the model's internal behavior cannot be inspected. Governance relies on evidence gathered through evaluation and operational monitoring.

Vendor-hosted models continue to evolve through safety updates, new training data, bug fixes, and feature improvements. While these changes may improve performance, they can also introduce unexpected behavior or disrupt established workflows. Organizations should monitor vendor release notes, define the types of changes that trigger reassessment, and perform regression testing before or immediately after updates. Rollback procedures and contingency plans help reduce disruption if an update affects system performance or compliance. Users should also be informed when changes materially affect how the system behaves, particularly in regulated or high-impact environments.

Proprietary models often process prompts outside the organization's environment, increasing the risk of exposing sensitive information. Controls should limit the information included in prompts through redaction, masking, and clear acceptable use policies. Retrieval systems should provide access only to approved content, and vendor logging and retention practices should be understood before deployment. Access to sensitive workflows should be restricted, and staff should be trained to avoid entering confidential information into prompts. These safeguards combine policy with technical controls to reduce the risk of unintended disclosure.

Because proprietary models cannot be inspected directly, governance depends on strong guardrails and continuous evaluation. Guardrails may define approved tasks, restrict retrieval sources, enforce refusal behavior, or require human review for high-impact outputs. Evaluation should include red teaming, stress testing, and a library of test prompts that reflects both expected use and potential misuse. Monitoring should track output quality, stability over time, and emerging failure patterns so issues can be identified and addressed quickly. Regular testing confirms that safeguards remain effective as the model and its operating environment change.

Managing proprietary models also requires careful procurement and contract review. Agreements should clearly define responsibilities for data handling, retention, breach notification, vulnerability management, and incident support. They should also describe service availability, rate limits, change notification processes, and available transparency documentation. Exit planning is equally important so the organization can migrate to another provider or reduce its dependence on a single vendor if circumstances change. Procurement decisions influence technical and operational risk, making contract review an important part of AI governance.

Proprietary models introduce uncertainty through limited visibility, ongoing vendor changes, and external data processing. Managing those risks requires continuous testing, effective guardrails, operational monitoring, and well-defined contractual obligations. Throughout the model's lifecycle, accountability remains with the organization deploying the system.

Managing third-party risk

Third-party risk arises whenever an AI system depends on external vendors, partners, contractors, or data providers. In production environments, those dependencies often include foundation models, cloud platforms, monitoring services, retrieval tools, and external datasets. Changes to any of these components can affect system performance, security, compliance, or availability. Managing third-party risk therefore extends beyond procurement. It requires ongoing governance throughout the system's operational life.

Begin by identifying every external dependency and understanding the role it plays. This includes model providers, cloud platforms, data sources, and supporting tools. Each

dependency should then be classified according to its potential impact on the system. A vendor supporting a customer-facing AI assistant presents a different level of risk from one providing an internal prototype environment. Documenting the systems, data, and business processes each vendor can access helps determine the controls, monitoring, and contractual protections required. A dependency inventory also makes it easier to assess the impact of service outages or vendor changes before they disrupt operations.

Vendor due diligence should reflect the level of risk associated with each dependency. Review the vendor's security controls, access management, logging practices, vulnerability management, and incident response capabilities. Assess privacy practices, data retention policies, and whether customer data is used for model training. Understand how model updates are communicated, what rollback options exist, and how service reliability is maintained. Where subcontractors are involved, review the additional risks they introduce. The depth of this assessment should be proportionate to the sensitivity of the data and the importance of the service.

Contracts establish important governance controls but should be supported by technical and operational safeguards. Agreements should define responsibilities for data handling, confidentiality, retention, breach notification, incident support, and model updates. Operational controls such as least-privilege access, network segmentation, monitored integrations, and periodic vendor reviews help enforce those commitments. Data minimization remains important even when a vendor agrees not to retain customer information. Governance is strongest when contractual obligations and technical controls reinforce one another.

Third-party risk does not remain static. Vendors release software updates, change service terms, publish security advisories, and introduce new features throughout the life of a service. Continuous monitoring helps detect reliability issues, unexpected changes in model behavior, security incidents, and declining service quality before they affect production systems. Periodic contract reviews, security assessments, and monitoring of approved tool usage help ensure that external dependencies continue to meet organizational requirements as risks evolve.

Exit planning reduces dependence on any single provider and supports business continuity. Organizations should understand how documentation, prompts, retrieval content, and data can be transferred to another platform if necessary. They should also assess how changes to pricing, service availability, licensing terms, or vendor strategy would affect operations. Critical systems benefit from tested fallback options or alternative providers that can be adopted with minimal disruption. Planning for vendor replacement strengthens resilience and reduces the operational and governance risks associated with vendor lock-in.

Third-party risk extends across the entire AI system because external services influence how the system operates over time. Managing that risk requires a clear understanding of

dependencies, proportionate due diligence, effective contractual and technical controls, continuous monitoring, and practical exit strategies.

Evaluating vendor and open-source agreements

Vendor agreements define how AI services can be used, how data is handled, and where responsibilities lie when issues arise. They influence operational risk as much as legal risk by determining how incidents are investigated, how information is protected, and how changes are managed. Open-source software introduces additional considerations, including licensing obligations, security maintenance, and long-term support. Reviewing these agreements helps organizations identify constraints and obligations before systems enter production.

One of the first areas to review is how the vendor handles data. Agreements should specify whether prompts, logs, and outputs are retained, how long they are stored, whether they are used for model training, and whether customers can opt out of that use. They should also identify where data is stored, who can access it, including subcontractors, and what happens to the data when the agreement ends. For high-risk use cases, organizations may require additional contractual protections and independent verification. Where terms are unclear, data sharing should be limited until the risks are understood.

Security provisions should define expectations for breach notification, vulnerability management, incident response, and ongoing cooperation during investigations. Agreements should also describe logging capabilities, audit evidence, access controls, encryption practices, and the protection of credentials and secrets. High-impact deployments often require stronger assurance through security certifications, independent audits, or regular assessments. Where vendor controls provide limited visibility, organizations should strengthen their own monitoring and technical safeguards.

Vendor updates present another contractual consideration. Agreements should explain how changes are communicated, whether specific model versions can be retained, and how customers are notified before updates are introduced. Service availability, rate limits, support arrangements, and continuity plans should also be defined. For critical systems, organizations should understand the vendor's uptime commitments, escalation procedures, and approach to service discontinuation. Where version control cannot be guaranteed, additional regression testing and operational monitoring become essential.

Most vendor agreements limit the provider's liability, but responsibility for the AI system remains with the deploying organization. Contracts should clearly define ownership of outputs, responsibilities during investigations, regulatory support, and the handling of misuse or harmful outcomes. Open-source software typically offers no contractual support, leaving organizations responsible for security updates, maintenance, and operational oversight.

Contractual protections are most effective when supported by internal governance processes, technical controls, and clearly assigned accountability.

Open-source software offers flexibility but requires active management. Organizations should understand licensing requirements before redistributing software or incorporating it into commercial products. Dependency tracking, vulnerability scanning, patch management, and release testing help reduce security risks throughout the software supply chain. Community activity, documentation quality, and long-term project maintenance should also be considered when evaluating whether an open-source component is suitable for production use. Treating open-source software as a managed dependency helps reduce operational and security risk over time.

Vendor and open-source agreements establish the conditions under which AI systems operate. Reviewing data handling practices, security obligations, update policies, liability provisions, and licensing terms helps organizations identify risks early and define the controls needed to manage them.

Applying deployment governance in practice

Jason Home Furnishings plans to deploy a generative AI assistant to support customer service and sales. The assistant will answer product questions, reduce customer wait times, and help staff draft responses. It also introduces new risks. Incorrect product information, inappropriate commitments, exposure of customer data, or unnecessary access to internal systems could all affect customers and the business. The scenario demonstrates how governance decisions shape a production deployment when business objectives and operational risks must be balanced.

The first step is to define clear boundaries for the assistant. It can draft responses using approved product information, delivery guidance, and return policies, but it should not process refunds, modify customer accounts, access payment information, or make commitments that depend on live business data unless those details are verified through authorized systems. Customer-facing interactions should remain separate from internal staff workflows because they require different controls. Access restrictions, connector permissions, and tool limitations should enforce these boundaries, keeping the assistant in a decision-support role.

Customer conversations contain names, addresses, order numbers, and other personal information. The assistant should receive only the data required to complete each task and retain it only for as long as necessary. Where retrieval is used, it should access approved product and policy documentation. Order status requests should rely on controlled integrations that return only the information needed to answer the query. Logging should support troubleshooting without capturing unnecessary personal data, and staff should be

trained not to include payment details or other confidential information in prompts. Testing should also verify that the model does not memorize or disclose sensitive information.

Product specifications, pricing, delivery commitments, and return policies are all areas where inaccurate responses can have direct business consequences. The assistant should generate answers from curated content and acknowledge when reliable information is unavailable without speculating. Human review may be appropriate for high-impact communications, particularly during the early stages of deployment. Testing should cover discontinued products, regional delivery variations, and policy exceptions. Monitoring customer complaints about pricing disputes, delivery errors, and incorrect product information can help identify issues that require corrective action.

Security controls should address both malicious attacks and everyday misuse. Prompt injection attempts, unauthorized document retrieval, and inappropriate use of confidential business information should all be included in testing. Role-based access controls, restricted connectors, and monitoring for unusual activity reduce the risk of unauthorized access. Incident response procedures should define how suspected data leakage or harmful outputs are reported, contained, and investigated. Where the assistant relies on an external model provider, governance processes should also account for vendor updates that could affect system behavior.

A phased rollout allows the organization to validate the assistant before expanding its use. An initial pilot with internal staff provides an opportunity to measure operational improvements alongside safety indicators such as override rates, complaint volumes, and incorrect product information. Predetermined thresholds can define when deployment should pause for investigation or remediation. Feedback from the pilot can be used to refine prompts, retrieval content, guardrails, documentation, and training before broader deployment. Expansion should be based on evidence that the system performs reliably and that governance controls operate as intended.

This scenario illustrates how deployment governance combines technical controls, operational processes, and organizational oversight. Defining boundaries, protecting customer data, validating outputs, strengthening security, and introducing the system gradually all contribute to a safer production deployment. The next section focuses on maintaining that level of assurance through ongoing assessment, monitoring, and incident response.

Monitoring and responding to operational risk

Once an AI system is in production, governance shifts from deployment readiness to ongoing oversight. Models change, user behavior evolves, and external dependencies introduce new risks over time. Maintaining trust requires regular assessment, continuous monitoring, and a structured approach to incident response. This section explains how to review deployed

systems, monitor their performance and safety, and respond to incidents in a way that limits harm and supports continuous improvement. By the end of this section, you will be able to design a post-deployment review process, identify the monitoring signals that matter for safety and reliability, and apply a structured incident response process when issues arise.

Conducting periodic performance and safety assessments

Periodic assessment is a scheduled review that confirms an AI system continues to meet its operational, performance, and governance requirements after deployment. Models evolve, user behavior changes, and external dependencies are updated, making regular review an essential part of production governance. Assessments evaluate whether the system remains accurate, reliable, and safe for its intended use, while confirming that controls such as human oversight, monitoring, and incident response continue to operate as expected. Without regular review, issues may remain undetected until they affect users or business operations.

Performance assessments should focus on metrics that reflect the system's purpose. For a customer service assistant, relevant measures might include resolution rates, escalation rates, and customer satisfaction. Decision-support systems may require calibration analysis and closer examination of errors in borderline cases. Where fairness is a concern, performance should be assessed across different groups because overall averages can mask uneven outcomes. Generative AI systems also require evaluation of output quality, hallucination rates, refusal behavior , and whether retrieval responses remain grounded in current source material. Combining quantitative metrics with qualitative review provides a more complete picture of system performance and helps identify emerging issues before they become operational problems.

Reliability assessments evaluate whether the system remains available and performs consistently under normal operating conditions. Service availability, response times, error handling, and performance during peak demand all influence how people use the system. Persistent outages or slow response times may encourage users to bypass established workflows or seek unsupported alternatives. Reviews should therefore include incident records, downtime trends, infrastructure performance, and the stability of external dependencies such as cloud services, models, and data connectors. Reliable operation supports effective oversight by allowing governance processes to function consistently.

Safety assessments confirm that the system continues to operate within its intended boundaries. Reviews should look for harmful or misleading outputs, policy bypass, prompt injection attempts, and emerging patterns of misuse. Periodic red team exercises help evaluate new attack techniques and evolving prompt strategies, while reviews of escalated interactions and near misses reveal issues that routine testing may overlook. Assessments should also confirm that required human review is taking place and that operational safeguards continue

to function as intended. Findings should lead to practical improvements, such as updating guardrails, strengthening retrieval constraints, or improving user guidance.

Although periodic assessments follow a planned schedule, some events should trigger an immediate review. Examples include model updates, new data sources, expanded use cases, additional user groups, policy changes, declining performance, complaint spikes, security alerts, or operational incidents. Governance processes should define what constitutes a material change and the evidence required before the system continues operating under its existing approval. Trigger-based reassessments help control scope changes and ensure that documentation reflects the current state of the system.

Assessments create value only when they lead to action. Outcomes may include continuing operations without change, narrowing the system's scope, strengthening guardrails, introducing additional human oversight, updating training, or temporarily suspending deployment while issues are addressed. Each action should have a defined owner, realistic timeline, and supporting documentation so that future reviews can evaluate progress. Repeated findings may also indicate that the use case itself requires reconsideration. Regular assessment supports continuous improvement by turning operational evidence into governance decisions.

Periodic assessment provides the evidence needed to maintain confidence in an AI system throughout its operational life. Reviewing performance, reliability, safety, and material changes allows organizations to identify emerging risks, verify that governance controls remain effective, and make informed decisions about continued deployment.

Monitoring for risks, mitigation, and model drift

Once an AI system enters production, gradual change often poses a greater risk than sudden failure. User behavior evolves, business processes change, models are updated, and new data enters the system over time. Monitoring provides continuous visibility into how the system performs under real operating conditions, helping organizations identify emerging issues before they become incidents. It also confirms that governance controls continue to operate as intended.

Monitoring should focus on indicators that reflect user impact, including complaint trends, escalation rates, and reversed decisions. Quality metrics may include unsafe outputs, policy violations, or incorrect refusals. Where fairness is relevant, organizations should monitor differences in outcomes or error rates across affected groups. Privacy and security monitoring should identify prompt injection attempts, unusual access patterns, and outputs that expose sensitive information. Reliability indicators, such as latency, availability, and service interruptions, are equally important because operational issues often encourage users to

bypass established controls. Every monitoring signal should have defined thresholds, clear ownership, and a documented response process.

Model drift occurs when an AI system operates under conditions that differ from those for which it was originally designed or evaluated. Data drift reflects changes in the characteristics of incoming data, while concept drift occurs when the relationship between inputs and expected outcomes changes. Generative AI systems may also drift following vendor updates, changes to retrieval content, or modifications to prompt templates. The effects may appear as declining accuracy, changes in response style, inconsistent refusal behavior, or emerging bias. Early detection allows organizations to investigate and address these issues before they affect users or business processes.

Monitoring should also verify that mitigation measures remain effective. Human oversight can be measured through review and override rates. Retrieval controls can be evaluated by monitoring the sources accessed and identifying unauthorized retrieval attempts. Output filtering should be assessed by tracking how frequently safeguards are triggered and whether issues cluster around specific workflows. User training can be evaluated by monitoring prohibited prompt patterns or other indicators of unsafe behavior. Connecting monitoring activities to governance controls provides evidence that safeguards remain effective and highlights areas that require improvement.

When monitoring identifies drift or another issue, organizations should follow a structured response process. The first priority is containment to reduce exposure while the issue is investigated. This may involve narrowing the approved use case, increasing human oversight, restricting specific prompts, or reverting to a previously validated configuration. The next step is diagnosis to determine whether the issue stems from changing data, workflow modifications, vendor updates, or misuse. Corrective actions may include retraining, adjusting decision thresholds, updating retrieval content, or strengthening guardrails. Finally, lessons learned should be reflected in documentation, test cases, and monitoring practices so similar issues can be detected and addressed more quickly in the future.

An effective monitoring program combines continuous oversight with scheduled review. Automated alerts identify immediate concerns, while regular trend analysis and periodic governance reviews provide a broader view of system performance over time. Responsibilities for reviewing alerts and taking corrective action should be clearly assigned so important signals are not overlooked. High-impact systems often require cross-functional oversight, bringing together technical, operational, legal, and governance perspectives. Monitoring processes should also be tested periodically to confirm that alerts trigger correctly and escalation procedures remain effective.

Monitoring provides the operational evidence needed to identify drift, verify governance controls, and respond before issues become incidents. Consistent monitoring, supported by

clear response processes, helps organizations maintain confidence in AI systems throughout their operational life.

Managing incident response and downstream consequences

An AI incident is any event in which a system causes, or has the potential to cause, harm. Incidents extend beyond security breaches and include unsafe outputs, privacy violations, discriminatory outcomes, model drift, and misleading information provided to users. An effective incident response process limits further harm, preserves evidence, and restores safe operation. AI systems often influence many users and decisions before problems become visible, making a timely and consistent response essential.

Effective incident response begins with early recognition. Organizations should define what constitutes an incident and ensure that employees can identify and report harmful outputs, suspected data exposure, performance degradation, or unusual patterns of user complaints. Monitoring should support this process by detecting prompt injection attempts, repeated policy bypasses, unexpected changes in model behavior, and other indicators of elevated risk. Frontline teams are often the first to observe emerging issues, so reporting processes should be simple, well understood, and supported by a culture that encourages escalation. Near misses should also be documented and reviewed because they often expose weaknesses before they result in harm.

Once an incident has been identified, the immediate priority is containment. For a customer-facing assistant, this may involve limiting the system to draft responses or temporarily handing interactions back to human agents. Decision-support systems may require automated recommendations to be suspended while additional human oversight is introduced. Retrieval systems may need temporary restrictions on access to sensitive information, while prompt injection attacks may require tighter input validation, connector restrictions, or additional filtering. Containment measures should be planned in advance for high-impact systems so they can be implemented quickly without creating unnecessary disruption.

After the immediate risk has been contained, the investigation should establish what happened, why it occurred, and how similar incidents can be prevented. This requires reviewing logs, version histories, recent changes to models, prompts, retrieval content, and supporting infrastructure. Investigations should also consider how people interacted with the system, since many incidents result from a combination of technical and operational factors. Root causes may include data drift, vendor updates, ineffective controls, or changes in business workflows. The investigation should identify affected users, assess the impact of AI-generated outputs, and determine the corrective actions required.

Remediation addresses both the immediate effects of an incident and the underlying causes. Harmful decisions may need to be reviewed or reversed, affected individuals may require

notification or redress, and governance controls may need to be strengthened through updated prompts, guardrails, retrieval rules, or human review requirements. Test cases should also be expanded to detect similar failures earlier in the future. Clear communication throughout the remediation process helps maintain confidence among users, regulators, and other stakeholders.

The final stage is organizational learning. Post-incident reviews should identify missed warning signs, evaluate the effectiveness of the response, and determine which controls require improvement. Training, monitoring thresholds, documentation, and decision records should all be updated to reflect the lessons learned. The objective is to strengthen governance processes and reduce the likelihood of similar incidents recurring. Organizations that treat incident reviews as opportunities for continuous improvement are better prepared to respond to future risks.

Incident response is a core element of AI governance because it provides a structured process for identifying, containing, investigating, and resolving events that affect deployed systems. The next section explores the accountability and transparency measures that help organizations communicate AI use, demonstrate responsible governance, and support ongoing oversight.

Ensuring accountability and transparency

Effective AI governance extends beyond operating a system safely. Organizations must also demonstrate how the system is used, how risks are managed, and how governance controls are evaluated over time. This section explains how to communicate AI use through appropriate disclosures, manage risks and incidents as part of ongoing operations, and use audits to verify that governance controls work in practice. By the end of this section, you will be able to design deployment disclosures for different audiences, establish an ongoing risk management process for deployed AI, and identify the evidence required during an AI audit.

Meeting disclosure and transparency obligations

Transparency helps people understand when they are interacting with AI, what the system is designed to do, and where its limitations lie. Disclosures provide that information at the appropriate point in the user journey, helping people make informed decisions about how to use the system. Clear communication reduces the risk of users overestimating the system's capabilities and supports organizational accountability by explaining how AI is being used in practice.

Disclosure requirements vary by audience. Customers should know when they are interacting with AI and how to reach a human if necessary. Employees need guidance on approved tools, acceptable data inputs, and situations that require human review. Individuals affected by AI-

assisted decisions, such as applicants or borrowers, may require meaningful explanations and clear routes for requesting review or redress. Leadership benefits from visibility into system limitations, operational risks, and incident readiness, while regulators and auditors may expect evidence that disclosures have been communicated and maintained. The level of detail should reflect the potential impact of the system.

Where AI influences decisions, organizations should explain the role it plays within the overall process. Decision-support systems should make it clear that AI provides recommendations that are reviewed by people, while automated decisions should explain the level of automation, the oversight in place, and the options available for review or appeal. High-level information about the categories of data used and the factors considered can improve understanding without revealing proprietary information. Clear explanations help people understand how decisions are made and how they can challenge outcomes when appropriate.

User interfaces should avoid creating unrealistic expectations about AI capabilities. Systems should be clearly identified as AI, communicate uncertainty where appropriate, and avoid presenting speculative responses with unwarranted confidence. Human review remains essential in domains where accuracy has consequences, such as legal, medical, or contractual advice. Clear communication about system capabilities and limitations helps users rely on AI appropriately and reduces the risk of misuse or over-reliance.

Disclosures should be integrated into operational processes as well as user interfaces. Staff should understand how to reinforce them during customer interactions, and organizations should maintain records showing what information has been communicated and when it was last updated. User feedback and usability testing can help determine whether disclosures are understood or require improvement. Disclosure practices should also be reviewed following system changes or incidents so users continue to receive accurate and relevant information. Maintaining this evidence supports governance activities, including internal reviews and external audits.

Clear disclosures help users understand how AI systems operate, when human oversight is available, and how concerns can be raised. Embedding transparency into interfaces, operational processes, and governance records supports informed use and provides evidence that organizations are meeting their accountability obligations.

Managing incidents, issues, and risks

Managing AI in production involves more than responding to major incidents. Organizations also need processes for tracking smaller issues, monitoring known risks, and improving controls over time. Incidents require immediate containment and remediation, while issues may expose weaknesses that warrant investigation before they become more serious. Risks represent potential sources of harm that have not yet materialized but require ongoing

attention. Managing all three within a single governance process helps organizations prioritize work, assign ownership, and strengthen the system as it evolves.

The first step is triage. Every event should be classified according to its severity, potential impact, and likelihood of causing harm. A single incorrect response may be recorded as an issue, while repeated harmful outputs or widespread failures may require a formal incident response. Classification should also identify the affected domain, such as privacy, fairness, safety, security, or reliability, so the appropriate teams can investigate and respond. Defined escalation paths, response times, and evidence requirements help ensure that similar events are handled consistently.

A risk register provides a structured record of known risks, the controls designed to manage them, the individuals responsible for those controls, and the remaining level of residual risk. For AI systems, the register may include risks such as model drift, bias, prompt injection, privacy breaches, unsafe outputs, or excessive reliance on AI-generated content. It should also document how controls are monitored and what actions will be taken if a risk materializes. Keeping the register up to date helps organizations maintain continuity as systems, teams, and business priorities change, while giving leadership a consolidated view of governance risks across multiple AI deployments.

Corrective actions should address the underlying cause of an issue and define measurable improvements. Actions may include refining prompts, updating retrieval constraints, retraining models, adjusting human review thresholds, improving staff guidance, or strengthening technical safeguards. Repeated exposure of sensitive information, for example, may require a combination of input filtering, user training, and tighter access controls. Every action should have a designated owner, an agreed timeline, and a method for verifying that the change has achieved the intended outcome.

Change management plays an important role because AI systems continue to evolve after deployment. Model updates, new data sources, expanded use cases, and changes to business processes can all introduce new risks. Organizations should define what constitutes a material change, the evidence required before implementation, and the approvals needed for higher-risk changes. Rollback procedures should also be established so systems can be restored quickly if problems emerge. For externally hosted models, change management should include monitoring vendor updates and performing regression testing where appropriate.

Effective governance also depends on visibility into how risks evolve over time. Dashboards can summarize incidents, open issues, risk trends, response times, and the effectiveness of key controls. Metrics such as human review rates, drift alerts, or policy violations help identify areas requiring attention, while trend reporting allows leadership to prioritize resources and evaluate whether governance controls are improving over time. Reporting should focus on the information needed to support decisions and direct improvement efforts.

Managing incidents, issues, and risks as part of a single governance process helps organizations respond consistently, strengthen controls, and adapt to change throughout the system's operational life.

Auditing AI systems and demonstrating accountability

AI audits evaluate whether a deployed system operates as intended and whether governance controls remain effective in practice. Accountability ensures that organizations can explain who approved a system, why it was approved, and what evidence supported those decisions. Responsibility for a deployed AI system always remains with the organization, regardless of whether the model is developed internally or provided by a third party. Audits provide independent assurance by evaluating how governance operates under real conditions and identifying control gaps before they lead to harm. The depth and frequency of an audit should reflect the potential impact of the system.

Audits review whether a system is being used within its approved scope and whether controls prevent unauthorized or inappropriate use. They examine data handling practices, including access controls, retention, logging, and compliance with organizational policies. Evidence of testing, including fairness, safety, security, and robustness evaluations, should be available and appropriate for the system's intended use. Audits also assess monitoring processes, incident management, documentation, change records, and the effectiveness of human oversight. The focus is on objective evidence that governance controls are operating as intended.

Clear accountability depends on well-defined roles and responsibilities. Product teams typically own the business purpose and approve the use case, while engineering teams are responsible for implementation, monitoring, and change management. Privacy teams oversee data handling practices and privacy risk, security teams manage technical safeguards and incident preparedness, and risk or compliance functions provide governance oversight and escalation processes. Operations and support teams manage day-to-day use, user guidance, and the identification of operational issues. Clear ownership allows organizations to make decisions efficiently and respond quickly when problems arise.

Audit planning should begin before deployment. Readiness reviews confirm that governance controls are in place before launch, while post-deployment audits verify that those controls continue to operate effectively. Additional audits may be required following model updates, new data sources, expanded use cases, incidents, or other material changes that could affect risk. Higher-impact systems generally require more frequent or more detailed reviews because changes can invalidate earlier assumptions and testing.

Effective audits focus on the controls that matter most. Standardized assessment templates, monitoring dashboards, decision records, and other reusable evidence reduce unnecessary administrative effort and improve consistency across projects. Audits should include

interviews, observation of operational workflows, and verification that documented controls are functioning in practice. Findings should identify clear actions, assign ownership, and establish timelines for remediation. Tracking findings over time helps organizations measure improvements in governance maturity and identify recurring weaknesses.

Accountability becomes most visible when systems fail or produce unexpected outcomes. Organizations should be able to demonstrate that incidents were managed appropriately, affected users received appropriate support or redress, and governance controls were updated in response to lessons learned. Decision records, investigation findings, and implementation of corrective actions provide evidence that the organization has fulfilled its governance responsibilities. Regular audits help verify that these practices are consistently applied across deployed AI systems.

AI audits provide independent evidence that governance controls remain effective throughout a system's operational life. By evaluating governance processes, assigning clear accountability, and verifying that corrective actions are implemented, audits strengthen organizational oversight and support continuous improvement.

Selecting deployment architectures and model patterns

Technical architecture shapes how AI systems are governed in production. Decisions about deployment environments, model customization, retrieval mechanisms, and autonomous capabilities influence privacy, security, operational complexity, and the controls organizations need to maintain over time. This section explains how deployment environments affect governance, when to use foundation models or fine-tuned models, how retrieval-augmented generation (RAG) changes knowledge governance, and why agentic AI systems require additional oversight. By the end of this section, you will be able to evaluate deployment environment trade-offs, choose an appropriate model deployment strategy for different risk levels, and apply governance controls to retrieval-augmented generation (RAG) and agentic AI systems.

Choosing a deployment environment

Where an AI system is deployed affects what an organization can control, what it must trust, and the risks it must manage. Cloud deployments offer scalability and managed services but increase dependence on external platforms. On-premises deployments provide greater control over data and infrastructure while requiring more internal expertise and maintenance. Edge deployments process data closer to where it is generated, reducing latency and data transfer but introducing challenges for consistency and updates. The appropriate deployment environment depends on the use case and its level of risk.

Deploying in the cloud

Cloud deployments can speed implementation because infrastructure and scaling are easier to manage. They can also provide strong security capabilities when configured correctly, including identity management, access controls, and monitoring. Cloud environments support rapid updates and experimentation, making them well suited to systems that evolve frequently. At the same time, data may leave the organization's environment and be processed on shared infrastructure. Dependence on cloud providers also introduces risks such as service outages and changes outside the organization's control. Misconfiguration remains one of the most common causes of cloud security incidents. Organizations should also consider how logging, data retention, and access are managed across cloud services. For sensitive workloads, cloud deployments require strong data minimization practices and robust access controls. Whether the cloud is appropriate depends on the use case and the sensitivity of the data being processed.

Deploying on-premises

On-premises deployments keep sensitive data within the organization's controlled environment. This can support stricter access controls, custom monitoring, and clearer data handling boundaries while reducing dependence on external vendors for core processing. The trade-off is greater operational responsibility. Organizations must maintain hardware, apply security patches, and ensure system reliability. They also need the expertise to manage scaling, performance tuning, and incident response. Slower update cycles can leave models outdated and delay critical security fixes. An on-premises deployment is not inherently secure because internal environments can still be misconfigured, and weak access controls or retention practices can create privacy risks. On-premises deployments are often well suited to high-sensitivity workloads, provided the organization has the operational capability to maintain them effectively.

Deploying at the edge

Edge deployments run AI systems close to the data source, such as on devices, local servers, or embedded systems. A key advantage is low latency, making edge deployment well suited to real-time applications such as safety monitoring and industrial control. Keeping data close to its source can also reduce data transfer and limit privacy exposure. The trade-off is greater operational complexity. Limited computing resources can restrict model size and performance, while distributed endpoints make updates, patching, and security more difficult. Physical access to devices also increases the risk of tampering. Monitoring can be more challenging because logs are distributed across multiple devices, and differences between local environments may lead to inconsistent model behavior. Edge deployments can be effective for safety-critical applications but require disciplined update, monitoring, and maintenance processes.

Most enterprises use hybrid deployment models that combine cloud, on-premises, and edge environments. Training may occur in a controlled environment while inference runs in the cloud, with sensitive processing remaining on-premises. Retrieval systems may also operate on-premises while the model runs in the cloud. Hybrid deployments can reduce risk when they are designed deliberately, but they also increase operational complexity if governance is not planned from the outset. Organizations should document architectural boundaries, define how data moves between environments, and apply monitoring, access controls, and change management across the entire system.

Selecting a model use pattern

Model use patterns describe whether an organization uses a model as-is or adapts it through fine-tuning or other techniques. This choice affects performance, control, risk, and accountability. Using a model as-is can speed deployment but limits customization and often requires stronger guardrails to manage behavior. Fine-tuning can improve performance for specific tasks but introduces additional risks related to training data and changes in model behavior. Both approaches require testing, monitoring, and documentation, although the focus differs depending on how the model is used.

Using a model as-is can accelerate deployment by avoiding retraining and complex model development pipelines. It also avoids introducing new bias through additional training data. However, the model may not align well with a specific domain, increasing the likelihood of confident but incorrect responses. Organizations often rely more heavily on prompting, guardrails, and human oversight to keep behavior within acceptable boundaries. Vendor updates can also change model behavior unexpectedly, increasing the need for monitoring and regression testing. Limiting the system's scope and grounding responses in curated retrieval content can help reduce these risks. Using a model as-is is often appropriate for low- to moderate-risk tasks with well-defined boundaries.

Fine-tuning adapts a model using additional training data to improve performance for specific tasks or domains. It can improve response quality, consistency, and use of domain-specific terminology. Fine-tuning may also reduce hallucinations in narrowly defined use cases when supported by high-quality training data. Because it changes model behavior, previous evaluations may no longer apply. Organizations must also manage the training dataset carefully, as poor data governance can introduce bias, increase memorization risks, or weaken existing safety behaviors. Fine-tuned models should be versioned, documented, and evaluated as distinct model variants.

Training data is one of the most important governance considerations for fine-tuning. Personal information, confidential documents, sensitive communications, and data collected for unrelated purposes should not be used without appropriate authority and safeguards. Even when identifiers have been removed, contextual information may still reveal an individual's

identity, making data minimization and de-identification essential. Organizations should also evaluate training data for bias, information leakage, and memorization risks. Documentation should record the types of data used, the purpose of the training, and the governance controls applied throughout the process.

The governance emphasis differs between these approaches. Using a model as-is relies heavily on guardrails, retrieval constraints, and human oversight, while fine-tuning places greater emphasis on training data quality, model evaluation, and version management. Both approaches require testing, but fine-tuned models should also undergo comparative evaluations to confirm that task-specific improvements have not introduced new safety, fairness, or privacy issues. Monitoring should reflect the chosen approach, whether tracking vendor updates for foundation models or behavioral changes in fine-tuned models. Documentation remains important in both cases, with fine-tuned models requiring more detailed records of training data, evaluations, and model versions.

Choosing between these approaches starts with the intended use case and its level of risk. Fine-tuning may be appropriate for narrow, repeatable tasks that benefit from domain-specific knowledge, while using a model as-is is often sufficient for exploratory or lower-risk applications with well-defined boundaries. Where training data is sensitive or permissions are uncertain, retrieval over approved content may be more appropriate than fine-tuning. Organizations should also consider operational capability, ongoing maintenance, vendor dependence, and lifecycle costs before making a decision. The selected approach should align with the system's risk profile and be supported by clear documentation of the reasons for that choice.

Deploying retrieval-augmented generation safely

Retrieval-augmented generation (RAG) is a deployment pattern in which the system retrieves information from a curated knowledge source before using a model to generate a response. This approach reduces the need to retrain the model whenever underlying content changes. It can improve accuracy when the knowledge base is well maintained and access is appropriately controlled. At the same time, the system may confidently repeat incorrect, outdated, or unauthorized information if retrieval is poorly governed. In production deployments, failures often stem from retrieval scope, document quality, and access permissions.

A retrieval system consists of three components: the knowledge source, the retrieval process, and the generation step. The knowledge source may include policies, manuals, contracts, support tickets, or product documentation, making content quality a key factor in response quality. The retrieval process determines which documents are indexed, ranked, and returned while enforcing permission checks. The generation step determines how closely the model grounds its response in the retrieved content. Retrieval does not guarantee accurate answers. Outdated, conflicting, or poorly scoped information can still produce incorrect responses.

Knowledge sources should therefore be curated, governed, and maintained with clear ownership.

Access control is one of the most important governance controls in a retrieval system. Users should only be able to retrieve documents they are authorized to access. Weak permission controls can turn retrieval into a pathway for unauthorized disclosure, even when the underlying model is secure. Sensitive information should be included in the knowledge base only when required for the intended use case, and retrieval queries, retrieved passages, and logs should be handled according to the organization's privacy requirements. Employee-facing systems should also enforce confidentiality boundaries between teams and business functions.

Retrieval depends on the quality of the underlying knowledge base. Organizations should assign ownership for maintaining documents, retiring obsolete content, and correcting errors. Metadata such as effective dates, authorship, and applicability help ensure that the correct information is retrieved in the appropriate context. Conflicting or outdated documents can produce inconsistent responses even when retrieval functions correctly. Curating authoritative content and introducing review processes for high-impact documents help maintain the reliability of the knowledge base.

Even with retrieval, models can generate unsupported or inaccurate information. Grounding techniques help keep responses aligned with retrieved content, while refusal behavior prevents the system from inventing answers when sufficient information is unavailable. Testing should include prompts that encourage unsupported conclusions or request information that the knowledge base does not contain. High-impact use cases may require human review for edge cases or sensitive responses. Monitoring should identify situations where responses are generated without sufficient retrieval support, helping distinguish knowledge gaps from model behavior.

Monitoring should include retrieval activity as well as model outputs. Organizations should monitor frequently accessed sources, failed queries, irrelevant retrieval results, and unauthorized retrieval attempts. Repeated questions that cannot be answered may indicate gaps in the knowledge base, while output quality issues should be traced back to the retrieved content wherever possible. Governance processes should also cover updates to indexed content, ranking logic, and permission rules because these changes can alter system behavior. Treating retrieval as a managed operational capability helps maintain the accuracy and reliability of deployed AI systems as knowledge sources evolve.

Governing agentic AI architectures in deployment

Agentic AI extends beyond generating text by planning tasks and interacting with external tools to complete them. An agent might search internal knowledge, draft a response, create a support ticket, and update a customer record as part of a single workflow. This can improve

efficiency but also increases risk because the system performs actions that affect business operations. Incorrect actions, such as sending information to the wrong recipient or updating the wrong record, can have greater consequences than inaccurate responses alone. Tool access also expands the attack surface, making prompt injection and unauthorized actions more risks. Agentic systems should therefore be governed as operational automation with clearly defined controls.

Agentic systems differ from conversational AI because they can change systems as well as generate information. Their impact can be broader, and errors can propagate quickly across connected tools and workflows. Agents often combine retrieval, tool integrations, and action execution, increasing both technical complexity and governance requirements. Broader capabilities also increase the need for clear accountability, detailed logging, and well-defined operational boundaries. As autonomous capabilities expand, governance controls should increase accordingly.

Controlling tool access and permissions

Tool access is the primary governance control for agentic systems. Agents should receive only the permissions required to perform their assigned tasks, following the principle of least privilege. A customer support agent, for example, may need permission to view order status without the ability to modify billing information. High-impact actions such as refunds, cancellations, or account changes should require additional approval. Testing environments should remain isolated from production systems, and controls should restrict the information that agents can pass to external tools. Identity management, role-based access controls, and comprehensive logging of tool use help prevent unauthorized actions and support investigation when incidents occur.

Implementing human oversight and confirmation gates

Confirmation gates require human approval before an agent performs high-impact or irreversible actions. An agent can prepare a plan or draft an action, while a person reviews the context and confirms that it is appropriate. Effective oversight requires reviewers to understand what they are approving, including the privacy, security, and policy implications of the action. Lower-risk tasks, such as drafting or routing information, may be automated with fewer restrictions, while broader autonomy should be introduced gradually as operational evidence demonstrates that controls remain effective. Human oversight also provides feedback that can improve prompts, tool configurations, and governance processes over time.

Defending against prompt injection

Prompt injection is a risk because attackers may attempt to manipulate an agent into ignoring instructions or misusing connected tools. Effective controls include input filtering, clear policies governing tool use, and separation of system instructions from user inputs. Tool

requests can also be validated using structured formats before execution, while tool outputs should be checked before they influence subsequent actions. Limits on the number or scope of actions an agent can perform during a session reduce the impact of successful attacks. Monitoring should identify unusual tool usage, repeated access attempts, and other suspicious patterns, while testing should include prompt injection and jailbreak scenarios as part of routine validation. Protecting agentic systems requires multiple layers of technical and operational controls.

Building audit trails and recovery plans

Agentic systems should maintain detailed audit trails showing the plans generated, the tools invoked, the data accessed, and the actions performed. These records support troubleshooting, compliance activities, and incident investigations while providing evidence of how decisions were made. Recovery plans should define how incorrect actions are reversed, including rollback procedures, manual correction processes, and conditions that require automated actions to be suspended. Thresholds for pausing the system should be established in advance, such as repeated incorrect actions or suspicious tool activity. A phased rollout can further reduce operational risk by allowing agents to recommend actions before they are permitted to execute them automatically. Expanding autonomy should be based on evidence that the system performs reliably under production conditions.

Agentic AI systems introduce governance challenges because they perform actions that extend beyond generating information. Managing those risks requires controlled tool access, meaningful human oversight, layered security controls, comprehensive audit trails, and well-defined recovery procedures. These measures allow organizations to deploy agentic systems while maintaining accountability, operational control, and the evidence needed to investigate and improve system behavior over time.

Summary

In this chapter, we learned about the governance practices required to deploy AI systems safely and sustain that safety throughout production.

We began by establishing the foundation for responsible deployment. We identified the privacy and security policy gaps introduced by AI-specific patterns such as prompt inputs, model outputs, and vendor-hosted processing, and defined the technical, operational, human, and compliance readiness needed before a system goes live.

We then explored the vendor and third-party dependencies that shape most production AI systems. We discussed how to govern proprietary models with limited visibility, assess, and manage third-party risk, evaluate vendor, and open-source agreements, and apply these principles in a practical deployment scenario.

Next, we focused on sustaining safe operations after deployment. We covered periodic assessment, continuous monitoring, and incident response, showing how organizations detect model drift, manage operational risks, respond to incidents, and use those experiences to strengthen governance over time.

We then looked at the organizational practices that support accountability in production. We discussed how disclosures and transparency help users understand AI systems, how incidents, issues, and risks are managed through ongoing governance processes, and how audits verify that controls are operating effectively through documented evidence.

Finally, we covered the technical decisions that influence governance in production. We compared cloud, on-premises, and edge deployment environments, explored the governance implications of using foundation models directly or fine-tuning them, examined retrieval-augmented generation (RAG) and the controls needed to govern it, and discussed how agentic AI systems require stronger oversight through controlled tool access, human confirmation, layered security controls, and comprehensive audit trails.

In *Chapter 6, The EU AI Act,* we will learn how these deployment governance practices connect to regulatory obligations, how the Act classifies AI systems, assigns responsibilities to providers and deployers, and establishes obligations based on risk.

6

The EU AI Act

The **EU AI** is the world's first comprehensive regulatory framework for artificial intelligence, establishing a risk-based approach that scales compliance obligations according to the potential for harm. For AI governance professionals, the Act is not primarily a legal text to memorize. It is a decision-making framework for determining the appropriate response to a given situation. The ability to classify a scenario quickly and identify the correct next steps is the practical skill the Act requires.

This chapter works through the Act systematically. It begins with scope, exemptions, and role definitions, providing the foundation for determining whether the Act applies and who is responsible. It then introduces the classification toolkit, covering AI literacy requirements and the four-tier risk framework. The chapter applies those tools to each risk tier in sequence before turning to general-purpose AI as a distinct regulatory track. It concludes with the operational compliance controls and enforcement consequences that give the framework its practical effect.

In this chapter, we will cover the following topics:

- Establishing the scope and actors of the EU AI Act
- Applying AI literacy and risk classification principles
- Navigating obligations across the four risk tiers
- Governing general-purpose AI and building operational compliance

Establishing the scope and actors of the EU AI Act

This section introduces the foundational elements of the EU AI Act, including what the framework aims to achieve, when and where it applies, and who bears responsibility under it. It covers the Act's purpose and territorial reach, the exemptions that may narrow or remove regulatory obligations, and the roles of providers, deployers, importers, and distributors,

whose responsibilities shape compliance across the AI supply chain. By the end of this section, you will be able to determine whether the Act applies to a given scenario, assess applicable exemptions, and assign role-based obligations to the relevant actors.

Applying EU AI Act essentials to governance practice

The European Union Artificial Intelligence (EU AI) Act is the world's first comprehensive regulatory framework for AI. It adopts a risk-based approach that scales regulatory obligations according to the potential for harm. For AI governance professionals, the Act provides a framework for evaluating AI systems, identifying applicable obligations, and determining the governance measures required in different contexts.

The Act seeks to reduce risks associated with AI while supporting innovation and economic growth. It addresses concerns such as unsafe systems, discriminatory outcomes, manipulation, and misleading interactions, while creating a consistent regulatory framework across the European Union. To achieve these objectives, it emphasizes governance practices such as documentation, testing, monitoring, and human oversight, with obligations that increase as the potential impact of an AI system increases.

Applicability depends on where an AI system is placed on the market, put into service, or affects individuals within the European Union. Organizations outside the European Union may also fall within the Act's scope if their AI systems are used in the European market or produce outputs that affect people there. The Act assigns different responsibilities to providers and deployers, while supply chain participants may also have obligations depending on their role. Many organizations adopt governance practices that satisfy the Act across all operations to avoid maintaining different compliance programs for different jurisdictions.

The Act organizes obligations into four risk categories: unacceptable risk, high risk, limited risk, and minimal risk. This framework determines the level of governance expected for a particular AI system. Unacceptable-risk systems are prohibited. High-risk systems are subject to extensive governance and compliance requirements because they may affect safety or fundamental rights. Limited-risk systems primarily introduce transparency obligations, while minimal-risk systems are subject to relatively few additional regulatory requirements beyond general governance practices.

Article 4 introduces AI literacy as a foundational requirement for organizations that develop or deploy AI systems. AI literacy is based on role-appropriate competence, ensuring that individuals understand the capabilities, limitations, and risks of the AI systems they develop, deploy, oversee, or use. Effective training supports transparency, human oversight, appropriate escalation, and the consistent application of governance controls across the organization.

Determining when exemptions remove regulatory obligations

Exemptions define the situations in which the EU AI Act does not apply or applies differently. They should be considered at the beginning of the analysis because they determine whether regulatory obligations are triggered. An exemption may change the applicable obligations, but it does not replace the need to assess the system's risks or governance requirements. Once scope has been established, the system should be evaluated using the Act's risk-based framework.

The first step is to determine whether the AI system is placed on the market, put into service, or produces outputs that affect individuals within the European Union. If it falls outside the Act's territorial scope, the Act may not apply. If it falls within scope, the analysis continues with risk classification and the obligations associated with the relevant risk category. Territorial scope is the starting point, not the conclusion of the assessment.

An exemption should not be interpreted as permission to ignore governance controls. Even where formal regulatory obligations do not apply, organizations continue to benefit from practices such as testing, monitoring, documentation, and human oversight because the underlying risks remain. Internal systems, research activities, or pilot deployments may still affect individuals or business decisions and should be governed accordingly.

A practical way to assess exemptions is to consider the system's purpose, the people or organizations it affects, and where it is deployed. These factors determine whether the Act applies and whether an exemption remains valid. The analysis should focus on the system's actual use.

Scope may also change over time. A system developed for research or internal use can become subject to the Act if it is later deployed in regulated contexts or made available in the European Union. Change management and system inventories help organizations identify these transitions and reassess their regulatory obligations before deployment expands.

Exemptions should therefore be treated as context-dependent decisions. They narrow regulatory obligations where appropriate but do not remove the need for proportionate governance.

Identifying provider, deployer, importer, and distributor roles

The EU AI Act assigns responsibilities according to an organization's role, not its job title. Two organizations may use the same AI system but have different obligations depending on whether they develop, place, distribute, import, or deploy it. Understanding these roles is

essential because they determine which compliance obligations apply throughout the AI lifecycle.

A **provider** develops an AI system and places it on the market under its own name or trademark. Providers are responsible for demonstrating that regulated systems meet the Act's requirements through appropriate documentation, technical evidence, and lifecycle controls. Their responsibilities include managing updates, maintaining compliance documentation, and ensuring that required governance measures remain in place throughout the system's lifecycle. Organizations that develop and market AI systems typically act as providers.

A **deployer** uses an AI system in an operational setting, such as human resources, lending, education, or customer service. Deployers are responsible for using the system as intended, maintaining appropriate human oversight, and implementing operational controls such as monitoring, incident management, and staff training. Deploying a third-party AI system does not transfer responsibility to the vendor. Organizations remain accountable for how the system is used and the decisions it supports.

An **importer** places an AI system from outside the European Union onto the EU market, while a **distributor** makes that system available within the supply chain. These roles help ensure that compliance responsibilities extend beyond the original developer. Importers and distributors should verify that required documentation accompanies regulated systems and should not ignore obvious compliance deficiencies. Their responsibilities strengthen governance throughout the AI supply chain.

Some organizations perform multiple roles at the same time. A company may develop an AI system, market it under its own brand, and deploy it internally, creating obligations associated with both provider and deployer roles. Others may import, distribute, and operate the same system in different business contexts. Responsibility therefore depends on how an organization interacts with the AI system, and obligations accumulate when multiple roles apply.

Consider a software company that develops an AI screening tool and sells it to employers in the European Union. The software company acts as the provider, while each employer acts as the deployer. If the tool is imported into the European Union, an importer also becomes part of the supply chain, and a reseller may assume the role of distributor. If the employer uses the system without the required oversight, responsibility rests with the deployer. If the provider fails to supply the required technical documentation, responsibility remains with the provider. Assigning the correct role is the first step in determining the obligations that follow.

In the next section, we will build the analytical foundation for applying the EU AI Act by covering the AI literacy requirements under Article 4 and the **four-tier risk classification framework** that determines the obligations applied to AI systems.

Applying AI literacy and risk classification principles

This section builds the analytical framework needed to understand how the EU AI Act applies across different levels of risk. It covers the AI literacy requirements introduced in Article 4, which define the competence expected of providers and deployers, and the four-tier risk classification framework that determines the obligations associated with an AI system. By the end of this section, you will be able to describe role-appropriate AI literacy expectations and apply the four-tier framework to classify AI systems and identify the appropriate governance response.

Developing role-based AI literacy for safe deployment

AI literacy is the role-appropriate knowledge needed to develop, deploy, and use AI safely and responsibly. It ensures that individuals understand the capabilities, limitations, risks, and appropriate use of AI systems. Many AI failures result from misuse, overreliance, or poor judgment, not model defects. Effective training turns governance policies into consistent operational practice and enables staff to communicate the use and limitations of AI accurately, supporting transparency and responsible oversight.

Training requirements vary across roles. End users need to understand what a system can and cannot do, when to rely on its outputs, and when to escalate to a human reviewer. Reviewers should be able to evaluate outputs, identify bias or unsafe content, and recognize situations that require intervention. Engineers need to manage model changes, monitor performance, and respond to incidents, while leaders should understand how to evaluate evidence and make informed governance decisions. Support teams also require clear guidance and escalation procedures for handling complaints and user concerns. Effective training builds confidence without encouraging unquestioning reliance on AI.

AI literacy is reflected in day-to-day behavior. Staff validate outputs before acting on them, document exceptions, report issues promptly, and recognize when human intervention is required. Managers introduce automation gradually and support deployment decisions with evidence, while technical teams recognize prompt injection attempts, data leakage risks, and other indicators of misuse. These practices embed governance into routine operations and strengthen incident detection and response.

Training is most effective when it is continuous and aligned with operational responsibilities. One-time awareness sessions, generic content, or training that focuses only on product features rarely prepare staff for real-world use. Programmes should include role-based examples, periodic refreshers, incident response exercises, and clear procedures for documenting decisions and escalating concerns. Reinforcement over time helps establish consistent governance practices across the organization.

AI literacy also supports transparency and accountability. Trained staff communicate the role of AI appropriately, explain its limitations, maintain accurate records, and apply the level of human oversight required by the use case. These behaviors strengthen compliance while reducing the risk of misuse, overconfidence, and inappropriate reliance on AI-generated outputs.

The effectiveness of AI literacy should be assessed through observable behaviors as well as formal training records. Useful indicators include fewer policy violations, earlier escalation of issues, improved documentation quality, consistent use of approved tools, and appropriate human review where required. Monitoring these behaviors provides evidence that training is influencing operational practice and strengthening governance.

Classifying AI systems using the four-tier risk framework

The EU AI Act classifies AI systems into four risk categories: unacceptable risk, high risk, limited risk, and minimal risk. This framework determines the obligations that apply to an AI system based on its intended use and the potential harm it may cause. Classification begins with the impact of the system, not the technology it uses.

Unacceptable-risk AI systems are prohibited because the risks they present are considered incompatible with the Act. These systems involve practices that manipulate behavior, exploit vulnerable individuals, or undermine human autonomy in ways that create unacceptable harm. Where a proposed use falls within this category, the appropriate governance response is to halt deployment or redesign the system so that the prohibited capability is removed.

High-risk AI systems remain permissible but are subject to the Act's most extensive governance requirements because they can affect fundamental rights, safety, or access to important opportunities. Typical examples include systems used in employment, education, healthcare, or other high-impact decisions. Organizations deploying high-risk systems should implement risk management, testing, technical documentation, human oversight, monitoring, and change management throughout the system's lifecycle.

Limited-risk AI systems primarily introduce transparency obligations. Individuals should be informed when they are interacting with AI or consuming AI-generated content where required by the Act. Appropriate disclosures help reduce confusion, deception, and overreliance while supporting informed use. Although these systems carry fewer regulatory obligations than high-risk systems, transparency remains an important governance control.

Minimal-risk AI systems are subject to relatively few specific obligations under the Act. Even so, organizations should continue to apply proportionate governance measures, including appropriate security controls, monitoring, and acceptable use policies. Risk classification should also be reviewed whenever the system's purpose or deployment context changes, as a low-risk system may become subject to stronger obligations when used in a different setting.

Risk classification is guided by three questions: What decision does the system influence? Who is affected? What harm could result? These questions help determine the appropriate risk category and the governance measures that follow. Focusing on the system's intended use and potential impact provides a more reliable basis for classification than considering the underlying technology alone.

In the next section, we will examine the compliance obligations associated with each risk tier, covering prohibited AI, high-risk systems, limited-risk AI, and minimal-risk AI.

Navigating obligations across the four risk tiers

This section focuses on the compliance obligations associated with each tier of the EU AI Act's risk classification framework. It covers prohibited AI systems, where deployment is not permitted; high-risk AI systems, where providers, deployers, importers, and distributors each have distinct responsibilities; limited-risk AI systems, where transparency requirements are the primary governance control; and minimal-risk AI systems, where governance remains proportionate to the level of risk. By the end of this section, you will be able to identify the compliance obligations associated with each risk tier and explain how the responsibilities of providers, deployers, importers, and distributors differ for high-risk AI systems.

Recognizing and responding to prohibited AI practices

Prohibited AI systems represent the highest level of risk under the EU AI Act. These uses are considered inherently incompatible with fundamental rights because their purpose undermines human autonomy, fairness, or human dignity. The appropriate governance response is to prevent deployment by stopping the activity, redesigning the use case, or removing the prohibited capability. The defining concern is the intended use of the system. Even a technically well-designed implementation does not make a prohibited use case acceptable.

Common indicators include systems that manipulate behavior, coerce individuals, exploit vulnerable groups, support social scoring, or enable forms of broad behavioural surveillance. Although these systems differ in their implementation, they share a common characteristic: the intended purpose creates unacceptable harm that cannot be addressed through ordinary governance controls. The assessment therefore focuses on the nature of the use case as well as its potential impact on fundamental rights.

Measures such as transparency notices, consent mechanisms, or additional human oversight do not generally make a prohibited use acceptable. Those controls reduce risk in other contexts, but they do not change a use case whose purpose is fundamentally incompatible with the Act. Distinguishing prohibited AI from high-risk AI is therefore essential. High-risk

systems remain permissible when appropriate controls are implemented, whereas prohibited AI systems should not be deployed.

When a prohibited use is identified, organizations should pause development or deployment, escalate the issue through the appropriate governance channels, document the reasons for the decision, and redesign the use case to remove the prohibited purpose. Where third-party products are involved, procurement and vendor management teams should also be informed so that inappropriate technologies are not introduced through the supply chain. Lessons learned should feed back into governance processes so similar proposals are identified earlier.

For example, an AI system designed to exploit psychological vulnerabilities to influence purchasing decisions remains problematic even if it performs accurately. Similarly, a system that assigns broad trustworthiness scores to individuals across unrelated aspects of life raises concerns because of its purpose, not technical defects. In both cases, governance focuses on preventing deployment by stopping the activity, redesigning the use case, or removing the prohibited capability.

The defining distinction between prohibited and high-risk AI lies in the purpose of the system. High-risk AI supports legitimate activities that require extensive governance controls to reduce risk. Prohibited AI pursues objectives that the Act considers unacceptable regardless of the safeguards applied. Evaluating the intended purpose alongside the potential harm provides the basis for determining the appropriate governance response.

Governing high-risk AI as a provider or deployer

High-risk AI systems remain permissible under the EU AI Act but are subject to the Act's most extensive governance requirements because they can affect fundamental rights, safety, and access to important opportunities. Common examples include systems used in employment, education, healthcare, and other high-impact contexts. Compliance requires governance controls that operate throughout the system's lifecycle and evidence demonstrating that those controls function as intended.

Providers are responsible for designing high-risk AI systems that satisfy the Act's governance requirements. They should establish a risk management process, document the system's intended purpose, data sources, limitations, and performance characteristics, and evaluate the system against the requirements of its intended use. Clear instructions for use help deployers understand the system's operating boundaries, required oversight, and known limitations. Providers should also maintain technical documentation and lifecycle records that demonstrate continued compliance as the system evolves.

Deployers are responsible for operating high-risk AI systems within their intended scope. Their responsibilities include training staff, maintaining meaningful human oversight, monitoring

system performance, responding to incidents, and managing changes that could affect risk. Deploying a third-party AI system does not transfer accountability to the provider. Organizations remain responsible for how the system is used and the decisions it supports.

Meaningful human oversight requires more than reviewing AI outputs. Individuals responsible for oversight should understand the system's capabilities and limitations, recognize situations that require intervention, and have the authority to challenge or override AI-assisted decisions. Oversight should be embedded within operational workflows and supported by clear escalation procedures, training, and decision criteria so that it remains effective under routine operating conditions.

Documentation and traceability support accountability throughout the lifecycle of a high-risk AI system. Technical documentation, risk assessments, testing results, monitoring plans, and change records provide evidence that governance controls have been implemented and maintained. These records also support incident investigation, auditing, and continuous improvement as systems evolve.

High-risk AI governance extends beyond deployment. Model updates, new data, operational changes, and vendor modifications can all affect system behavior and risk. Continuous monitoring, change management, and periodic reassessment help identify emerging issues before they affect individuals or business operations. Regression testing, defined reassessment criteria, and rollback procedures support safe operation as the system changes over time.

Fulfilling importer and distributor obligations for high-risk AI

High-risk AI obligations extend beyond providers and deployers. Importers and distributors play an important role in preventing non-compliant AI systems from entering or spreading through the European Union market. Their responsibilities help ensure that compliance information, technical documentation, and governance evidence remain available throughout the AI supply chain, supporting accountability and traceability at every stage.

An **importer** places an AI system from outside the European Union onto the EU market. Before making a high-risk system available, importers should confirm that the required documentation accompanies the product and that the provider has supplied the information needed to demonstrate compliance. Where documentation is incomplete or obvious compliance deficiencies exist, importers should pause distribution, seek clarification from the provider, and escalate concerns through the appropriate governance channels. These responsibilities make importers an important checkpoint within the supply chain.

A **distributor** makes an AI system available within the supply chain, including through resale or regional distribution. Distributors should ensure that products are accompanied by the

information and instructions required for safe deployment and should avoid presenting systems in ways that encourage inappropriate or unintended use. Where compliance concerns are identified, distribution should pause until those issues have been addressed. These responsibilities help preserve the integrity of compliance information as AI systems move through the market.

Effective supply chain governance depends on the continuous flow of evidence. Technical documentation, instructions for use, update information, and compliance records should accompany the system throughout its lifecycle so that importers, distributors, deployers, and regulators have access to the information needed to support safe operation, auditing, and incident response. Procurement and distribution teams should understand what documentation to request, retain, and communicate as systems evolve.

Supply chain participants should also respond consistently when compliance concerns arise. Missing documentation, unclear intended use, inconsistent capability claims, or limited transparency about updates are all indicators that require further investigation. Pausing distribution, requesting additional evidence, documenting the issue, and escalating concerns where appropriate help prevent non-compliant systems from entering operational use while preserving accountability across the supply chain.

Importers and distributors strengthen the governance framework by ensuring that compliance responsibilities extend beyond the original developer. Evidence, documentation, and accountability remain connected as AI systems move from providers to deployers, reducing the risk that governance controls are lost as products enter wider use.

Meeting Article 50 transparency obligations for limited-risk AI

Limited-risk AI systems are generally permitted under the EU AI Act but require transparency measures to reduce the risk of deception or manipulation. The primary safeguard is ensuring that people understand when they are interacting with AI or consuming AI-generated content. Unlike high-risk systems, the emphasis is not on extensive conformity requirements but on enabling informed decisions through clear and timely communication.

Transparency requires users to know when they are interacting with an AI system unless that fact is already obvious from the context. It also requires appropriate disclosure when content has been generated or manipulated by AI so that users can interpret it appropriately. These measures reduce the risk of deception, overreliance, and misplaced confidence in AI-generated outputs while supporting trust in AI-enabled services.

Responsibility for transparency is shared across the AI lifecycle. Providers should design systems that support appropriate disclosure and provide guidance on communicating AI use and synthetic content. Deployers are responsible for implementing those measures where the

system is used, ensuring that users receive clear and timely notice. Effective transparency depends on both roles working together to provide consistent information throughout the user experience.

Transparency is most effective when it is provided at the point of interaction. Users should know they are communicating with an AI assistant before relying on its responses, and AI-generated or manipulated content should be identified before it influences decisions or actions. Where AI supports sensitive activities, users should also understand the system's limitations and know how to seek human assistance when required. Clear communication strengthens trust while reducing the likelihood of misunderstanding or inappropriate reliance on AI.

Organizations should also address common transparency failures. Users may assume they are interacting with a person, synthetic content may be distributed without appropriate labeling, or disclosures may be inconsistent across different channels. Staff training, clear disclosure practices, and operational guidance help ensure that transparency requirements are applied consistently and remain effective over time.

A practical way to assess limited-risk systems is to consider whether users could reasonably be unaware that AI is involved. If so, appropriate notice should be provided. Where AI generates or manipulates content, clear labeling supports informed interpretation. Governance should also ensure that staff understand when transparency measures apply and how they should be implemented in day-to-day operations.

Maintaining proportionate governance for minimal-risk AI

Minimal- or no-risk AI systems are generally not subject to the extensive obligations that apply to high-risk AI systems. The European Commission notes that the vast majority of AI systems used in the European Union fall into this category. Although the regulatory burden is lighter, organizations should continue to apply basic governance practices because operational risks can still arise through misuse, data leakage, or changes in how a system is used. A system classified as minimal risk today may require reassessment if its use changes over time.

Minimal-risk systems are typically used for productivity or convenience and do not influence decisions about individuals. Public guidance often cites spam filters and similar low-impact automation as examples. Classification depends on how a system is used. A generative AI tool used for internal drafting may remain minimal risk when it supports routine activities without determining outcomes or processing sensitive information. The same tool may require a different classification if it is later used in hiring, lending, or other high-impact contexts.

Assessing the intended use, the people affected, and the potential consequences reflects the Act's risk-based approach.

Even where the Act imposes few additional requirements, organizations benefit from maintaining baseline governance controls. These include security and access controls, appropriate logging, incident response processes, and acceptable use policies that prevent sensitive information from being entered into unsuitable systems. Monitoring helps identify scope changes that could alter a system's risk classification, while role-appropriate AI literacy supports the safe and responsible use of these tools. A consistent governance baseline reduces operational risk and supports responsible AI use across the organization.

Scope creep is one of the most common governance challenges for minimal-risk AI systems. A tool developed for one purpose may gradually be used in another without reassessment. A summarization tool may become part of disciplinary decision-making, or a chatbot may begin supporting eligibility or policy decisions. As the context changes, so can the system's regulatory obligations. Maintaining accurate AI inventories, applying change management, and reassessing risk before expanding a system's use help ensure that governance remains aligned with the Act. Procurement processes should also consider new integrations that introduce sensitive data or extend systems into higher-impact workflows.

Organizations often adopt voluntary governance practices that extend beyond the Act's minimum requirements. These may include documenting intended use, defining prohibited uses, performing basic quality checks for user-facing outputs, and promoting fairness where systems interact with diverse groups. Proportionate governance applies controls that reflect the level of risk, reducing avoidable harm without imposing unnecessary requirements on low-impact AI systems.

A practical assessment begins by considering the decision the system influences and the potential consequences of its use. Additional considerations include whether sensitive information is processed, whether users could be misled about AI involvement, and whether future changes could increase the system's impact. These questions help determine whether a system remains appropriately classified as minimal risk or whether reassessment is required.

In the next section, we will examine general-purpose AI models as a distinct regulatory track before turning to the operational controls that sustain compliance across regulated AI systems, including documentation, human oversight, quality management, and enforcement.

Governing general-purpose AI and building operational compliance

This section focuses on **general-purpose AI** (**GPAI**) models as a distinct governance challenge before introducing the operational compliance mechanisms required for regulated AI systems.

It explains how responsibilities differ between GPAI model providers and AI system deployers, how documentation and conformity assessment provide evidence of compliance, how human oversight and quality management support ongoing governance, and how enforcement and penalties influence organizational compliance priorities. By the end of this section, you will be able to explain GPAI obligations across the model-to-system chain, describe a compliance approach based on documentation, oversight, and quality management, and connect enforcement outcomes to governance investment decisions.

Managing obligations and systemic risk for general-purpose AI models

GPAI models are designed for use across a wide range of tasks and contexts, making their risks more difficult to predict than those of purpose-built AI systems. The same model may support routine drafting in one application and high-impact decision-making in another, depending on how it is integrated. This flexibility creates distinct governance challenges because model providers do not control every downstream use. The EU AI Act therefore introduces specific obligations for GPAI model providers and additional requirements for models that present systemic risk.

A clear distinction exists between a **model** and an **AI system**. A model is a reusable component, while an AI system applies that model within a specific operational context. A GPAI model provider may supply the underlying model, while another organization develops and deploys an AI system for activities such as recruitment, education, or customer service. The Act assigns different responsibilities across this chain. Model providers supply documentation, transparency information, and other supporting materials, while system providers and deployers remain responsible for managing risks associated with their specific use cases. Compliance depends on both parties fulfilling their respective responsibilities.

GPAI model providers support downstream governance by supplying technical documentation, integration guidance, and information needed to manage risk. Public guidance also highlights responsibilities relating to transparency, copyright compliance, and information about training data, while recognizing the need to protect intellectual property. These materials help downstream providers and deployers understand how the model should be used and what limitations should be considered during integration.

The Act also introduces additional obligations for GPAI models that present **systemic risk**. As models become more capable and widely deployed, their potential impact extends beyond individual applications. Risks such as large-scale misinformation, cybersecurity threats, and widespread misuse require stronger governance measures. Managing systemic risk may involve enhanced testing, incident preparedness, monitoring, and change management to address the broader effects that highly capable models can have across multiple systems and sectors.

Deployers remain responsible for integrating GPAI models safely within their own AI systems. This includes defining an appropriate intended use, preventing scope expansion into higher-impact activities without reassessment, implementing transparency where required, and testing for risks such as hallucinations, prompt injection, and unsafe content generation. Ongoing monitoring, incident response planning, and careful data handling remain essential because updates to the underlying model can change system behavior over time. Vendor management also plays an important role in maintaining awareness of model updates and their governance implications.

When assessing a GPAI deployment, governance should focus on the intended use of the AI system. The model alone does not determine the applicable obligations. The responsibilities of the model provider and the deployer should be considered together, with obligations determined by the system's purpose, impact, and level of risk. A widely used GPAI model does not automatically create a high-risk AI system, but its capabilities and scale may require additional governance measures depending on how it is deployed.

Producing technical documentation and conducting assessments

For regulated AI systems, the EU AI Act requires evidence of compliance, not simply statements of intent. Technical documentation describes the system, its intended purpose, how it was developed, and the conditions for its safe use. Record-keeping provides the information needed to reconstruct events, investigate incidents, and demonstrate accountability. Conformity assessment establishes whether regulated systems, particularly high-risk AI systems, satisfy the Act's requirements before deployment and throughout their lifecycle. Together, these activities support consistent governance, effective monitoring, and informed incident response while providing evidence that regulatory obligations have been met. Documentation, records, and conformity assessments should be maintained as living governance practices that evolve with the system throughout its lifecycle.

Preparing and maintaining technical documentation

Technical documentation must be prepared before a high-risk AI system is placed on the market or put into service and maintained throughout its lifecycle. It describes the system's design, intended purpose, safeguards, and operating conditions in a form that can be reviewed by regulators and other stakeholders. Documentation typically includes a high-level description of the model, the categories of data used, testing, and validation activities, known limitations, and instructions that support the system's safe and intended use. Maintaining current documentation supports transparency, meaningful human oversight, and effective governance by providing the context needed to evaluate the system throughout its lifecycle. Outdated or incomplete documentation can undermine compliance, particularly when changes have been made to the system without corresponding updates to its technical records.

Technical documentation should be maintained as an integral part of ongoing change management throughout the system's lifecycle.

Implementing record-keeping for traceability and auditing

Record-keeping captures operational evidence, including system behavior, version changes, configuration updates, and other events throughout an AI system's lifecycle. It provides the traceability needed to understand how a system evolved and to reconstruct events when incidents occur. Records support incident investigations by linking system inputs, configurations, and outputs, while also providing evidence that governance controls were applied as intended. They enable auditing, fairness assessments, safety monitoring, and ongoing compliance by showing how the system has performed over time. Effective record-keeping also strengthens operational resilience by helping organizations identify issues, respond to incidents, and demonstrate compliance with regulatory requirements. Logging practices should provide sufficient operational detail while protecting sensitive information and respecting applicable privacy obligations.

Conducting conformity and impact assessments

Conformity assessment is the process through which providers demonstrate that high-risk AI systems satisfy the EU AI Act's requirements before they are placed on the market or put into service. It provides structured evidence that risk management measures, technical documentation, and governance controls meet the required standard. Conformity assessment supports stage-gate governance by confirming that regulatory requirements have been satisfied before deployment proceeds. It also establishes the foundation for ongoing quality management and monitoring throughout the system's lifecycle. Where applicable, the EU declaration of conformity provides formal confirmation that the system complies with the relevant regulatory requirements.

Impact assessment complements conformity assessment by focusing on the potential effects of an AI system on individuals and society. It considers who may be affected, the types of harm that could arise, and the governance controls needed to reduce those risks. This process helps organizations justify measures such as human oversight, transparency requirements, and other safeguards based on the intended use of the system. It also supports decisions about whether an AI system is appropriate for a particular context, reflecting the Act's emphasis on protecting fundamental rights alongside technical performance.

Documentation, conformity assessment, and impact assessment should be maintained throughout the AI system's lifecycle. As systems evolve through updates, new data, or changing operational contexts, technical documentation, records, and supporting evidence should be reviewed and updated to reflect those changes. Material modifications may require reassessment to confirm that the system continues to satisfy regulatory requirements.

Maintaining current evidence supports effective auditing, incident response, and ongoing governance while reducing the risk of compliance gaps as systems change over time.

Designing human oversight, transparency, and quality management

Human oversight, transparency, and quality management form the operational foundation of the EU AI Act's requirements for high-risk AI systems. Together, these controls ensure that governance extends beyond documented compliance into the day-to-day operation of AI systems. Human oversight enables intervention when systems produce unsafe or inappropriate outcomes. Transparency provides the information needed to operate AI systems safely and appropriately. Quality management establishes the processes that enable providers to maintain compliance throughout the system lifecycle.

Human oversight should enable individuals to understand AI-assisted decisions, intervene when necessary, and override system outputs where appropriate. Oversight should be integrated into operational workflows as an active governance control. Effective oversight requires trained personnel, clearly defined review thresholds, escalation procedures, and the authority to suspend or correct automated decisions when risks to health, safety, or fundamental rights arise. Monitoring also provides the information needed to identify situations that require human intervention.

Transparency supports safe and appropriate system use by providing deployers with the information needed to understand an AI system's intended purpose, limitations, and operating conditions. It should also communicate circumstances in which the system should not be used, helping to prevent misuse and scope expansion beyond the provider's intended design. Documentation, operational guidance, and role-appropriate training allow deployers to interpret system outputs correctly and exercise meaningful oversight throughout deployment.

Transparency also extends to interactions with end users. Where individuals engage directly with AI systems or consume AI-generated content, they should receive clear and timely notice that AI is involved. Appropriate disclosure, including the identification of synthetic or manipulated content where required, helps reduce deception, supports informed decision-making, and encourages appropriate reliance on AI outputs. Providing users with clear routes for seeking human assistance or reporting concerns further strengthens accountability and trust.

A quality management system provides the organizational framework that enables providers to build, maintain, and continually improve compliance for high-risk AI systems. It includes processes for risk management, documentation, testing, change management, corrective actions, and evidence retention, together with clearly defined roles and responsibilities. By embedding these activities into routine operations, quality management helps ensure that

compliance remains consistent as systems evolve and supports auditing, continuous improvement, and effective governance at scale.

These controls operate together as an integrated governance framework. Human oversight depends on meaningful transparency, while quality management provides the processes that sustain both throughout the AI system lifecycle. Monitoring, documentation, and change management reinforce these controls by ensuring that governance remains effective as systems are updated, expanded, and deployed in new contexts.

Preparing for enforcement and understanding penalty structures

Enforcement gives the EU AI Act practical effect by turning regulatory obligations into organizational risk. Penalties encourage organizations to establish governance controls before incidents occur and reinforce the importance of maintaining compliance throughout the AI system lifecycle. The Act links enforcement to the seriousness of the violation, with prohibited practices attracting the most severe penalties, followed by failures to meet obligations for high-risk AI systems and the provision of incorrect or misleading information to regulators. This risk-based approach encourages organizations to prioritize governance activities according to the potential impact of non-compliance.

The Act provides for administrative fines where prohibited AI practices are identified. Article 99 establishes penalties of up to €35 million or 7% of worldwide annual turnover, whichever is higher, for certain prohibited practice violations. These penalties reinforce that prohibited AI systems are not intended to be mitigated through additional controls but prevented from being developed, placed on the market, or deployed. Effective governance therefore begins with identifying prohibited use cases early and ensuring that they are escalated, documented, and removed before implementation.

Non-compliance with obligations for high-risk AI systems may also result in substantial penalties. Providing incorrect, incomplete, or misleading information to competent authorities can likewise lead to enforcement action, highlighting the importance of accurate documentation, record-keeping, and reporting throughout the system lifecycle. These requirements reinforce the need for organizations to maintain reliable governance processes, demonstrate compliance through evidence, and ensure that AI inventories, risk classifications, and supporting documentation remain current.

Enforcement also influences how organizations prioritize governance activities. Maintaining an inventory of AI systems, identifying prohibited uses, implementing documentation and testing for high-risk AI, providing appropriate transparency measures, and maintaining monitoring and change management processes all help reduce regulatory exposure while strengthening operational governance. These activities also improve organizational resilience

by reducing the likelihood of incidents and supporting more effective responses when issues arise.

Enforcement readiness depends on an organization's ability to demonstrate how its AI systems are governed. This includes maintaining accurate inventories, documenting system classifications, retaining technical documentation, monitoring records, incident logs, AI literacy records, transparency information, and evidence of change management activities. Governance responsibilities also extend across procurement and vendor relationships, requiring organizations to obtain and maintain appropriate compliance information from third-party providers. Effective enforcement readiness therefore depends on coordinated governance across legal, risk, engineering, security, operations, procurement, and product teams.

Enforcement priorities follow the Act's risk-based approach. Prohibited AI systems require immediate cessation and redesign. High-risk AI systems require documented evidence of compliance before deployment. Transparency failures should be addressed through appropriate user notices and operational controls, while inaccurate or misleading compliance information requires correction and stronger governance processes. The severity of enforcement measures reflects the seriousness of the underlying risks, helping organizations prioritize the controls that require the greatest attention.

Summary

In this chapter, we worked through the EU AI Act as a practical governance framework for AI professionals. We began by establishing the Act's scope, examining its purpose, territorial reach, and the circumstances in which exemptions narrow or remove regulatory obligations. We also mapped the four roles defined by the Act (provider, deployer, importer, and distributor) and explained how each carries distinct responsibilities based on market placement and operational use.

We then built the analytical foundation for classification. We covered Article 4's AI literacy requirements, showing that role-appropriate competence is a governance control because effective oversight depends on people being able to recognize risks, make informed decisions, and escalate concerns appropriately. We also introduced the Act's four-tier risk framework and the reasoning approach it supports. Classification begins with an assessment of potential harm and impact. Technology labels alone do not determine the appropriate risk category.

With the classification framework in place, we worked through the compliance obligations associated with each risk tier. We examined prohibited AI and established that the correct governance response is to stop and redesign harmful use cases. We then covered high-risk AI, working through the distinct obligations of providers and deployers before extending the analysis to importers and distributors, whose supply chain duties prevent non-compliant

systems from spreading without accountability. We also addressed limited-risk transparency obligations under Article 50 and minimal-risk governance, noting that a lighter regulatory burden still requires proportionate controls and reassessment when use cases expand.

We then explored general-purpose AI models as a distinct regulatory track, distinguishing the responsibilities of model providers from those of system deployers and examining the additional governance expectations associated with systemic risk. We also covered the operational controls that support ongoing compliance, including technical documentation, record-keeping, conformity and impact assessments, human oversight, quality management, and the enforcement framework that drives governance investment and accountability.

In *Chapter 7, Other Laws and Standards Related to AI*, we will broaden the regulatory perspective by examining other laws and standards that influence AI governance. We will compare the EU AI Act's risk-based approach with regulatory and standards-based frameworks in the United States, South Korea, and international standards, highlighting where they align and where their requirements diverge.

7

Other Laws and Standards Related to AI

The previous chapter covered the European Union's AI Act as a comprehensive risk-based framework for AI regulation. AI governance extends beyond a single framework. As jurisdictions introduce their own legal requirements, some aligned with the EU model and others following different approaches, organizations need governance programs that can operate across multiple legal environments.

This chapter broadens that perspective by exploring the wider AI regulatory landscape. It introduces regulatory alignment and dissonance across jurisdictions, follows the U.S. approach from federal executive guidance through state and city legislation, explains the privacy principles that underpin AI governance, and concludes with international governance frameworks and the **International Organization for Standardization (ISO)** and the **International Electrotechnical Commission (IEC)** standards that support consistent terminology, management systems, and AI impact assessments.

In this chapter, we will cover the following topics:

- Navigating the global AI regulatory landscape
- Implementing city and state employment and consumer AI controls
- Embedding privacy principles and governance best practices
- Applying international AI frameworks and ISO standards

Navigating the global AI regulatory landscape

This section explores how shared concerns about safety, bias, privacy, and transparency are addressed across different jurisdictions. It explains regulatory alignment and dissonance,

outlines how **U.S. federal** guidance shapes AI governance through executive action, use-case inventories, and procurement expectations, and shows how state privacy rules translate into product controls and reusable compliance programs. By the end of this section, you will be able to explain why global AI regulation is both aligned and dissonant, describe the core operational requirements of U.S. federal AI governance, and map state privacy obligations to practical governance controls.

Understanding alignment and dissonance across jurisdictions

Many countries share concerns about AI safety, bias, privacy, and transparency, even though they regulate them differently. These shared concerns allow organizations to build a core AI governance program that applies across multiple jurisdictions. AI regulation is becoming more closely aligned because many jurisdictions address the same underlying risks.

Countries may agree on the risks AI presents without regulating them in the same way. Some rely on comprehensive legislation, while others use guidance and sector-specific regulators. Definitions, implementation timelines, and enforcement mechanisms also differ. This regulatory dissonance means compliance with one framework does not automatically satisfy another. Organizations should establish a baseline governance program and adapt it to local legal requirements.

Organizations deploying AI in multiple jurisdictions may need additional controls for specific markets. One jurisdiction may require transparency notices. Another may impose additional privacy obligations. Maintaining an inventory of AI systems, where they are deployed, and how they are used allows organizations to identify where those controls apply. Vendor management also remains part of compliance because third-party AI systems become part of an organization's governance program.

Regulators expect organizations to demonstrate governance through evidence. Training records, monitoring results, incident logs, and change management documentation show that governance controls are operating as intended. AI governance has become an operational function supported by documented processes and continuous oversight.

A baseline governance program includes an AI inventory, risk classification, documented decisions, testing before deployment, and lifecycle monitoring. Local requirements can then be incorporated where they apply. This avoids creating separate governance programs for every jurisdiction and keeps governance activities consistent across the organization.

Organizations assessing regulatory obligations should identify the decisions an AI system influences, the people affected, and the potential harm that could result. They should also determine whether personal data is processed, whether transparency obligations apply, and whether local requirements introduce additional controls.

Consider a customer support chatbot deployed in several countries. Core governance controls include approved knowledge sources, clear disclosure that users are interacting with AI, logging, monitoring, and a path to human assistance. Local requirements may add specific transparency notices or privacy protections without changing the underlying governance program.

Governance applies to all AI systems according to their impact. Responsibility for governance does not transfer to a third-party vendor. Model updates, data changes, vendor modifications, and new deployment contexts can change a system's risk profile. Ongoing monitoring, documentation, and periodic reassessment keep governance aligned with those changes.

Building AI governance under U.S. federal guidance

In the United States, AI governance develops through executive actions, agency guidance, and procurement expectations because there is no single national AI law. Organizations establish repeatable governance processes for approving AI use, reviewing risk, and monitoring systems after deployment. Teams document intended use, define operating boundaries, and assign ownership before deployment at scale. Consistent governance practices also reduce differences in how AI systems are managed across business units. AI governance becomes an ongoing management function that extends beyond a one-time policy exercise.

Executive orders establish policy priorities, which agencies translate into operational requirements through programs, oversight, and procurement. *Executive order 14110, "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence,"* strengthens expectations for safe and secure AI development, accountability, and continuous oversight. Organizations respond by testing systems, documenting decisions, monitoring performance, and applying change management, incident response, and escalation procedures. These expectations also require organizations to identify system ownership, maintain governance evidence, and define how incidents are handled.

A use-case inventory identifies where AI is used across the organization. It records the system's purpose, business owner, deployment location, and the data it processes. The inventory also identifies high-impact use cases, including systems that influence hiring, eligibility, or access to services. Without an inventory, AI systems can spread across business functions and vendors before governance controls are applied. Maintaining an inventory allows organizations to prioritize reviews and apply controls where they are needed.

Many organizations adopt vendor solutions, cloud services, and AI capabilities embedded in existing platforms. Procurement therefore becomes part of AI governance. Organizations should obtain evidence describing a system's intended use, known limitations, update practices, and incident reporting procedures before deployment. This reduces the risk of deploying systems that perform well during demonstrations but fail under operational

conditions or change after vendor updates. Clear procurement requirements also strengthen vendor accountability.

Governance follows a repeatable review cycle that includes assessment before deployment, a stabilization review after deployment, and ongoing monitoring throughout operation. Data changes, user behavior, and vendor updates can alter system performance over time. Governance reviews define responsibility for monitoring, rollback decisions, and incident communication. Documented review activities also provide evidence of oversight.

Risk discussions often differ across technical, legal, and business teams. A common risk framework establishes shared terminology for discussing potential harms, likelihood, severity, and governance controls. **NIST AI Risk Management Framework (AI RMF 1.0)** provides a structure for managing AI risk across roles and lifecycle stages. Organizations use it to document intended purpose, testing activities, monitoring plans, and improvement actions, allowing governance activities to remain coordinated across teams.

Translating state privacy rules into product controls

State privacy rules often translate directly into user-facing controls, including notices, opt-out mechanisms, limits on data use, and risk assessments for higher-risk processing. Product teams incorporate these requirements into operational workflows. When AI influences decisions, privacy obligations focus on transparency and opportunities to challenge or avoid harmful outcomes. Organizations document a system's purpose, the data it processes, and the safeguards applied. Consistent implementation across teams reduces compliance risk and makes privacy governance part of product design.

Organizations should treat AI as a form of data processing with associated privacy risks. Systems that profile, score, rank, or recommend actions about individuals require additional scrutiny because they process personal information that may affect outcomes. Classifying these systems within an AI inventory allows organizations to apply governance controls according to their level of impact. A chatbot that provides general information, for example, may require additional controls if it later influences approvals or eligibility decisions.

Risk assessments support governance by documenting the purpose of an AI system, the data it uses, potential harms, and the controls applied to reduce those harms. They should also record design decisions, including the use of sensitive data or the decision to retain meaningful human involvement. This documentation provides traceability and demonstrates how governance decisions were made.

Transparency measures should be incorporated into the user experience through clear notices, plain-language explanations of AI use, and opt-out mechanisms where required. Building these controls into products from the outset promotes consistent implementation across channels and supports customer-facing teams with clear guidance on how AI systems operate.

Where audits are required by law or organizational policy, organizations should document what was tested, the issues identified, the corrective actions taken, and the results of those actions. Audit evidence should remain current because vendor updates, model changes, and retraining can affect previous findings. Governance activities also require coordination across technical, legal, business, and operational teams.

Organizations operating across multiple jurisdictions can maintain a common control library that includes AI inventories, risk reviews, transparency measures, testing, monitoring, and incident response. Jurisdiction-specific requirements, such as notice language or audit obligations, can then be incorporated where they apply. This approach promotes consistent governance without creating separate processes for each market. In the next section, we will look at how New York City, Illinois, and Colorado apply these principles to employment and consumer-facing AI systems.

Implementing city and state employment and consumer AI controls

This section focuses on three U.S. regulations that impose obligations on organizations deploying AI in employment and consumer-facing contexts. It explains New York City's bias audit requirements for automated hiring tools, Illinois's transparency, explanation, and data handling requirements for AI-analyzed video interviews, and Colorado's consumer protection framework for high-risk AI systems that addresses algorithmic discrimination. By the end of this section, you will be able to describe the classification, audit, disclosure, and remediation obligations in each jurisdiction and explain how these requirements translate into governance controls and hiring workflows.

Auditing automated hiring tools under NYC local law 144

New York City's **Automated Employment Decision Tool (AEDT)** rules increase accountability for automated hiring tools that score, rank, or recommend candidates for employment decisions. They do not prohibit automation. They reduce the risk of unfair outcomes by requiring organizations to understand how these tools operate, evaluate their performance, and communicate the required information. Hiring AI should be managed as a controlled system because biased tools can affect large numbers of candidates before problems are identified.

Many hiring systems include automated scoring features that organizations may not recognize as AI. A tool that screens resumes, ranks candidates, or recommends who progresses through the hiring process may fall within the scope of the AEDT rules, depending on how it is used. Organizations should map the hiring process, identify where automation influences decisions, and document those activities. Clear classification supports consistent governance across

recruiting teams and allows procurement teams to identify vendor products that include automated decision tools.

Bias audits should form part of an ongoing governance process. Changes in hiring patterns, candidate populations, and model updates can affect previous audit results. Organizations should define audit ownership, audit frequency, data requirements, and review procedures. Effective audits also depend on reliable data, consistent decision records, and a clear understanding of how the tool influences hiring decisions.

The AEDT rules also require transparency about the use of automated hiring tools and the results of bias audits. Organizations should explain what was tested and communicate the results clearly. Public disclosure also encourages consistent governance because hiring practices and supporting documentation remain subject to review. Transparency should be incorporated into recruitment workflows through updated postings, required notices, and standardized procedures.

Organizations should respond when an audit identifies risk. Remediation may include model updates, data improvements, revised decision thresholds, changes to the hiring process, or additional human review. Organizations should document the changes made, the reasons for those changes, and the evidence demonstrating improvement. This documentation supports accountability and reduces the likelihood of recurring issues after model or process changes.

Standardized governance controls allow organizations to manage hiring AI consistently across recruiting teams. Defining automated and human review activities, documenting governance decisions, and monitoring hiring outcomes support predictable workflows and early identification of potential issues. Monitoring candidate outcomes and unusual patterns also allows organizations to investigate concerns before they become systemic.

Applying Illinois disclosure requirements for AI video interviews

Video interviews can capture more than job-related skills. They may also reveal sensitive information such as disability-related characteristics, accents, age, or personal living conditions. When AI analyzes video interviews, it may draw conclusions from information unrelated to job performance or misinterpret factors such as lighting, internet quality, or communication style. Illinois addresses these risks by requiring transparency about the use of AI in video interviews. Applicants should understand when AI is used and how it contributes to the hiring process. These requirements reduce privacy risks and support fairer hiring practices.

Organizations should incorporate notice requirements into the interview process by providing clear information before recording and before AI analysis begins. Notices should remain consistent across job postings, candidate portals, recruiter communications, and vendor

platforms. Consistent messaging reduces confusion, supports compliance, and ensures applicants understand how AI is used during the hiring process.

Applicants should receive clear explanations of the factors evaluated by AI and how those factors relate to the role. Organizations should focus on job-related criteria and avoid unsupported claims about capabilities such as personality prediction or emotion detection. Recruiters and hiring managers should also understand the system's capabilities and limitations so they can apply appropriate human judgment when reviewing AI-assisted recommendations.

Video and audio interviews contain sensitive information and require appropriate controls for storage, access, sharing, and retention. Access should be limited to authorized personnel, and retention periods should align with hiring requirements. AI-generated scores, recommendations, and notes should be governed using the same controls because they influence employment decisions and may be reviewed during disputes or investigations. Documented data handling practices also reduce the risk of unauthorized use or disclosure.

AI analysis may misinterpret accents, speech patterns, disability-related characteristics, or nervousness that is unrelated to job performance. Organizations should include human review before candidates are excluded on the basis of AI-generated recommendations. Review procedures should identify unreliable indicators, define appropriate decision criteria, and document the reasons for employment decisions. Meaningful human oversight reduces fairness risks and demonstrates that hiring decisions are not based solely on automated processing.

Hiring programs often involve multiple recruiters, job roles, and vendor platforms. Standardized governance practices establish approved AI tools, defined use cases, disclosure requirements, and data handling procedures. Organizations should also assign responsibility for vendor approval, performance monitoring, candidate inquiries, and incident response. Standardized processes support consistent hiring practices as AI use expands across the organization.

Managing high-risk AI and algorithmic discrimination under Colorado law

Colorado's approach classifies certain AI uses as high risk when they affect people's opportunities and outcomes. A central concern is algorithmic discrimination resulting from the way an AI system is designed, trained, or used. Organizations should take reasonable care to prevent harmful outcomes and define clear responsibilities for developers and deployers. This consumer protection approach emphasizes preventing harm, detecting issues early, and responding through documented governance processes. Fairness and accountability remain ongoing governance responsibilities.

Organizations should clearly define an AI system and how it affects decisions before assessing risk. Documentation should describe the system's inputs, outputs, and how those outputs are used, whether they recommend actions, trigger workflows, or influence access to opportunities. Clear system boundaries identify ownership of the model, the business process, and the governance controls that apply. They also prevent systems introduced for low-impact purposes from expanding into high-impact use cases without additional governance.

Fairness risks can change as data, user behavior, and models evolve. Organizations should monitor outcomes, investigate unusual patterns, and reassess systems after retraining, feature updates, or vendor changes. Monitoring activities should be linked to change management so governance reviews occur whenever system changes affect risk. Monitoring should focus on measures that reflect the system's intended use and potential harms.

Organizations should establish response procedures before risks arise. Response plans should identify who investigates incidents, who can pause or restrict system use, how changes are approved and documented, and how information is communicated to leadership, affected teams, and regulators where required. Documented response procedures support consistent decision-making and provide evidence for investigations and corrective actions.

Many high-risk AI systems rely on vendor products or third-party components. Vendor governance should define the system's intended use, data sources, testing practices, update procedures, and support for audits and incident investigations. Procurement processes should require documentation that demonstrates the system can be deployed and governed safely. Vendor updates should also be reviewed through established change management processes so new risks are identified before deployment.

Organizations should apply a common risk framework to identify, evaluate, and manage AI risks across multiple systems. A structured framework defines the system's purpose, identifies potential harms, applies governance controls, validates those controls through testing, monitors performance after deployment, and records improvement activities. NIST AI RMF provides a widely used structure for lifecycle risk management and establishes a common approach for technical, legal, and business teams. In the next section, we will explore the privacy principles that apply across AI development and deployment and the governance responsibilities of providers and deployers.

Embedding privacy principles and governance best practices

This section focuses on the privacy principles that guide responsible AI development and deployment and the governance responsibilities of providers and deployers. It explains privacy by design and default, data minimization, transparency and explainability, security and

integrity, accountability and auditability, and the controls providers and deployers apply throughout the AI lifecycle. By the end of this section, you will be able to apply the core privacy principles across the AI lifecycle, distinguish the governance responsibilities of providers and deployers, and identify the controls appropriate to each role.

Applying core privacy principles across the AI lifecycle

AI systems process large volumes of data and operate across complex development and deployment pipelines, making privacy a core governance consideration throughout the AI lifecycle. Privacy principles provide the link between legal requirements and the technical practices used to design, build, deploy, and maintain AI systems. They establish governance controls that reduce privacy risks throughout the system lifecycle.

Privacy by design and default requires privacy to be incorporated into the architecture of an AI system from the outset. Organizations should identify the minimum personal data needed to achieve the intended purpose and configure systems to use the most privacy-protective settings by default. Privacy requirements should be integrated into system design from the beginning of development.

Data minimization limits the collection and use of personal data to what is necessary for a defined purpose. This principle is particularly important for AI systems because large datasets can increase privacy risks and introduce unnecessary bias. Training and evaluation datasets should contain only relevant data required for the intended use. Techniques such as synthetic data generation and federated learning can further reduce the use of personal data where appropriate.

Transparency and explainability ensure that individuals understand when AI is used and the role it plays in decision-making. Organizations should also be able to explain, in understandable terms, how AI-generated outputs are produced when personal data influences those outcomes. Explainability techniques such as Local Interpretable Model-agnostic Explanations and SHapley Additive exPlanations provide localized explanations that improve understanding of model behavior and support privacy rights.

Security and integrity extend beyond protecting stored and transmitted data. AI systems also require controls that protect training pipelines, model artifacts, and inference processes from attacks that compromise privacy or system reliability. Data poisoning attacks manipulate training data to influence model behavior, while model inversion attacks attempt to reconstruct sensitive training data from model outputs. Governance controls should protect the AI system throughout its lifecycle.

Accountability and auditability require organizations to assign responsibility for AI systems and maintain records throughout their lifecycle. Documentation should record how models are developed, tested, deployed, monitored, and updated, together with the data used and the

governance controls applied. These records support compliance activities, incident investigations, and continual improvement.

Together, privacy by design and default, data minimization, transparency and explainability, security and integrity, and accountability and auditability translate privacy principles into governance practices that can be applied throughout the AI lifecycle.

Implementing practical recommendations for providers and deployers

Providers and deployers have different governance responsibilities throughout the AI lifecycle. Providers develop AI systems and place them on the market, while deployers integrate those systems into operational environments. Although both roles are responsible for AI governance, their obligations differ according to how they develop, deploy, and manage AI systems.

Providers are responsible for designing AI systems with appropriate governance controls before deployment. Their responsibilities include defining the system's intended purpose, documenting capabilities and limitations, implementing safety measures, and demonstrating compliance with applicable requirements. Deployers are responsible for using AI systems within their intended scope, maintaining human oversight, monitoring system performance, managing operational risks, and responding to issues that arise during deployment. Together, these responsibilities establish governance throughout the AI lifecycle.

Embedding trustworthiness and documentation for AI providers

Providers should implement trustworthiness by design by incorporating ethical and safety controls from the earliest stages of development. Governance controls should be established before deployment and maintained as systems evolve. These controls include mitigating bias in training data, designing for robustness against adversarial behavior, and supporting explainability where AI systems influence decisions or outcomes. Trustworthiness should be demonstrated through system design and governance controls.

Documentation is a core provider responsibility because deployers depend on it to understand how an AI system should be used. Providers should supply a model card or technical documentation describing the intended use, system limitations, training and testing data considerations, known bias risks, performance metrics, including performance across relevant groups where appropriate, operating requirements, and known failure modes. Comprehensive documentation supports safer deployment, auditing, and incident response.

Providers should also establish update and maintenance policies that explain how patches are released, vulnerabilities are addressed, changes are classified, and end-of-support is communicated. Deployers need this information to assess the operational impact of updates

and maintain governance throughout the system lifecycle. Planned updates, documented changes, and clear communication support safe deployment and ongoing compliance.

Establishing oversight and monitoring for AI deployers

Deployers should complete an AI impact assessment before placing AI systems into operational use, particularly in high-risk or high-impact contexts. The assessment should identify the individuals or groups affected, the potential harms and expected benefits, and the governance controls required before deployment. It should also evaluate risks to fundamental rights, safety, and discrimination and be reviewed whenever the AI system or its operating context changes. Documented impact assessments provide evidence to support deployment decisions.

Deployers should establish meaningful human oversight for AI systems used in high-impact decisions such as hiring, lending, or medical triage. Individuals responsible for oversight should understand the system's capabilities and limitations, have the authority to override AI-assisted decisions, and be able to explain those decisions where required. Governance controls should also define escalation procedures, assign responsibility, and reduce over-reliance on AI outputs.

Deployers should implement continuous monitoring and periodic audits throughout the system lifecycle. Monitoring activities should identify model drift, data drift, and other changes that affect system performance. Regular audits confirm that governance controls remain effective under operational conditions. Organizations should also provide role-based AI governance training for legal, compliance, data science, product, and business teams so governance responsibilities are understood and applied consistently.

In the next section, we will explore international AI governance frameworks, beginning with South Korea's national AI law before introducing the **ISO/IEC** standards that establish common terminology, management systems, and AI impact assessment methodologies.

Applying international AI frameworks and ISO standards

This section focuses on international AI governance frameworks that complement the jurisdiction-specific requirements introduced earlier in the chapter. It explains South Korea's national AI law and its promotion-first approach to AI governance, together with **ISO/IEC 22989** for AI terminology, **ISO/IEC 42001** for AI management systems, and **ISO/IEC 42005** for AI impact assessments. By the end of this section, you will be able to describe South Korea's AI governance model and its key provisions, explain how the three ISO/IEC standards relate to one another, and apply their guidance to an auditable AI governance program.

Examining South Korea's AI basic law

South Korea's **Artificial Intelligence Development and Trust-Based Ecosystem Act**, commonly referred to as the **AI Basic Law**, illustrates a national approach to AI governance that combines AI development with governance measures designed to build public trust. Although the European Union's AI Act receives international attention, South Korea has adopted a comprehensive national framework that promotes AI innovation alongside governance and oversight.

The AI Basic Law is often described as adopting a promotion-first approach. It strengthens national competitiveness in AI and establishes a coordinated national strategy for AI development. Centralized governance supports the growth of the AI ecosystem before broader regulatory oversight is introduced.

A key feature of the framework is the establishment of national coordination structures that recognize AI as a strategic policy priority. Coordinated governance aligns policy development, industry support, and oversight across government and reduces fragmented approaches to AI governance.

The law also establishes expectations for trustworthy AI by emphasizing respect for human dignity, fundamental rights, transparency, fairness, and accountability. These principles align with governance themes that appear across many international AI frameworks.

The AI Basic Law requires disclosure and watermarking of AI-generated content. These measures address deepfakes and misinformation by identifying content that has been created or substantially modified using AI, allowing individuals to distinguish AI-generated content from human-created content.

The framework also introduces additional governance expectations for high-impact AI systems that affect public services, fundamental rights, or critical sectors. Although it does not adopt the European Union's tiered risk classification, organizations deploying AI in sensitive domains such as public safety, finance, and healthcare should expect stronger governance controls, risk assessments, and documentation requirements.

The law also promotes digital accessibility and the protection of vulnerable populations, reflecting the objective of ensuring that AI benefits are broadly shared and do not reinforce existing inequalities.

Establishing shared AI terminology with ISO/IEC 22989

ISO/IEC 22989, Artificial Intelligence – Concepts and Terminology, establishes the common vocabulary that underpins AI governance. Consistent terminology allows organizations to apply governance controls, conduct audits, assign responsibilities, and interpret regulatory requirements using shared definitions.

AI evolves rapidly, introducing new terminology and inconsistent usage across technical, legal, regulatory, and business contexts. Shared definitions ensure that terms such as AI agent, bias, and trustworthiness have consistent meaning across organizations. Common terminology also reduces inconsistent interpretation of governance requirements, overlooked risks, and conflicting audit outcomes.

ISO/IEC 22989 is a foundational standard within the ISO/IEC AI standards portfolio. It establishes a globally recognized vocabulary and conceptual framework that organizations use to build consistent AI governance programs across departments, vendors, and regulators.

The standard defines key AI concepts and terminology that establish the scope of AI governance. Definitions determine what constitutes an AI system, identify the concepts included within the AI lifecycle, and provide the basis for applying governance standards, developing documentation, and aligning organizational practices.

A clear definition of an AI system supports governance and compliance by identifying the systems that fall within governance processes and regulatory obligations. Consistent terminology also supports the implementation of AI management systems under ISO/IEC 42001 by ensuring that organizations apply governance controls to the same categories of AI systems.

The standard also defines trustworthiness as a set of characteristics that support AI governance. These characteristics include reliability, resilience, transparency, and accountability. Defining trustworthiness through measurable characteristics allows organizations to establish governance controls, testing activities, documentation requirements, and monitoring processes.

ISO/IEC 22989 also introduces terminology for AI lifecycle concepts, data and dataset concepts, and system components such as AI agents. A common vocabulary supports system classification, risk analysis, design documentation, and the assignment of governance responsibilities across products, vendors, and business units.

ISO/IEC 22989 provides the terminology that underpins the broader ISO/IEC AI governance standards. Shared definitions establish a common basis for implementing AI management systems under ISO/IEC 42001 and conducting AI impact assessments under ISO/IEC 42005.

Building an AI management system under ISO/IEC 42001

ISO/IEC 42001, Artificial Intelligence – Management System, establishes the requirements for an **Artificial Intelligence Management System (AIMS)**. Where ISO/IEC 22989 provides a common vocabulary for AI governance, ISO/IEC 42001 provides the management system used to apply those concepts throughout an organization. It is the first international standard developed specifically for AI management systems and can be adopted by organizations of any size.

ISO/IEC 42001 defines the requirements for establishing, implementing, maintaining, and continually improving an AI management system. The standard applies to organizations regardless of their size, structure, or use of AI. It requires documented processes, defined responsibilities, governance controls, and measurable activities that demonstrate how AI risks and opportunities are managed. Its alignment with the common high-level structure used by other ISO management system standards also supports integration with existing management systems.

ISO/IEC 42001 follows the **Plan-Do-Check-Act (PDCA)** cycle to support continual improvement. AI systems change over time as models are updated, data changes, user behavior evolves, and regulatory requirements develop. The PDCA cycle establishes a structured process for planning governance controls, implementing them, evaluating their effectiveness, and improving them throughout the AI lifecycle.

Defining context, scope, and leadership responsibilities

Organizations should first understand their operating context, including the needs and expectations of interested parties such as customers, regulators, and the public, and how those expectations relate to the use of AI. They should then define the scope of the AIMS by determining whether it applies to all AI systems or only those used in specific business functions or high-risk contexts. Scope defines which systems, governance activities, and documentation fall within the AI management system.

Top management is responsible for demonstrating leadership and commitment to the AIMS. This includes establishing an AI policy that defines the organization's objectives and direction for responsible AI and assigning roles, responsibilities, and authorities for the AI management system. Clear governance responsibilities establish ownership for approvals, monitoring, incident response, and accountability throughout the AI lifecycle.

Planning, operating, and continuously improving the AIMS

During the planning phase, organizations identify AI risks and opportunities through a structured risk assessment and define measurable objectives for responsible AI. Planning also establishes the governance controls needed to manage identified risks. These decisions determine what will be tested, documented, restricted, and verified before deployment.

The support and operation clauses govern the day-to-day operation of the AIMS. Organizations should provide the resources, competence, and awareness needed to implement governance effectively. Operational activities apply the controls defined during planning, including controls for data quality and acquisition, AI system development and testing, transparency and explainability, and meaningful human oversight.

Organizations should also monitor, measure, analyze, and evaluate the performance of the AIMS through internal audits and management reviews. These activities confirm that

governance controls remain effective and support continual improvement. ISO/IEC 42001 provides an internationally recognized management system for demonstrating responsible AI through documented governance processes and auditable evidence. The next section introduces ISO/IEC 42005 and its role in AI impact assessment.

Conducting AI system impact assessments with ISO/IEC 42005

ISO/IEC 42005, Guidance for the Assessment of the Impacts of AI Systems on Individuals and Society, provides guidance for conducting AI system impact assessments. Building on the common terminology established by ISO/IEC 22989 and the management system requirements of ISO/IEC 42001, it provides a structured process for assessing the impact of AI systems on individuals and society before and during deployment.

An **AI system impact assessment (AIA)** identifies, analyzes, and evaluates the potential risks and benefits of an AI system throughout its lifecycle. Organizations use the assessment to identify foreseeable impacts before harm occurs and to determine the governance controls needed to manage those risks.

The standard provides guidance for conducting impact assessments consistently across a wide range of AI applications. The assessment process can be applied to systems with different purposes, stakeholders, and risk profiles while maintaining a consistent governance methodology.

The assessment begins by defining the AI system's intended purpose, intended use, and operating context. Organizations should also identify the individuals, groups, and organizations that may be affected by the system. Defining the assessment scope ensures that the evaluation captures the full range of potential impacts. For example, an AI system used in hiring may affect job applicants, employees, and the human resources teams responsible for employment decisions.

Organizations should identify potential positive and negative impacts across multiple dimensions, including fundamental rights, societal impacts, and technical risks. Each impact should be evaluated according to its likelihood and severity to produce structured risk assessments that identify who may be affected and the consequences of those risks.

Organizations should then determine whether identified risks are acceptable and implement governance controls where additional risk treatment is required. Risk treatment may include changes to the model, training data, deployment approach, human oversight, or the decision not to deploy the AI system. The selected controls should reflect the nature and level of the identified risks.

The AI system impact assessment should be fully documented to provide evidence supporting governance decisions and compliance activities. Organizations should also review and update the assessment whenever the AI system, its operating context, or its risk profile changes. Impact assessment therefore becomes an ongoing governance activity throughout the AI lifecycle.

ISO/IEC 42005 provides the structured methodology for assessing AI impacts within the broader ISO/IEC AI governance framework. ISO/IEC 22989 establishes common terminology, ISO/IEC 42001 defines the AI management system, and ISO/IEC 42005 provides the methodology for assessing impacts on individuals and society. Together, these standards establish a consistent and auditable approach to AI governance.

Summary

In this chapter, we explored the global AI regulatory environment and the practical implications of regulatory alignment and dissonance. We discussed why many jurisdictions share common concerns about safety, bias, privacy, and transparency while adopting different approaches to defining scope, assigning responsibility, and enforcing obligations. We introduced the baseline-plus-local-add-ons governance model as a practical approach to managing AI across multiple jurisdictions and covered the U.S. governance landscape, from federal executive guidance and the NIST AI RMF to state privacy requirements that translate governance expectations into product controls and reusable compliance programs.

We then looked at three city and state examples that show how jurisdiction-specific requirements become operational governance controls. We covered New York City's bias audit requirements for automated hiring tools, Illinois's transparency, explanation, and data handling requirements for AI-analyzed video interviews, and Colorado's consumer protection framework for high-risk AI systems, where governance focuses on preventing algorithmic discrimination through clear system boundaries, continuous monitoring, incident response, and vendor accountability.

We also covered the privacy principles that apply across AI development and deployment, including privacy by design and default, data minimization, transparency and explainability, security and integrity, and accountability and auditability. We discussed how these principles translate privacy requirements into engineering and governance controls, including protections against threats such as data poisoning and model inversion attacks. We also distinguished the responsibilities of AI providers and deployers, showing how providers establish trustworthiness through system design, documentation, and maintenance, while deployers are responsible for impact assessments, human oversight, continuous monitoring, and operational governance.

Finally, we explored international AI governance frameworks. We covered South Korea's AI Basic Law and its promotion-first approach to AI governance, together with the ISO/IEC standards that support AI governance in practice. We discussed how ISO/IEC 22989 establishes a common vocabulary for AI, how ISO/IEC 42001 provides a management system for governing AI throughout its lifecycle, and how ISO/IEC 42005 provides a structured methodology for assessing the impacts of AI systems on individuals and society.

This chapter brings the book to a close. Across the previous chapters, we have explored the foundations of AI governance, responsible AI principles, AI risk management, AI development and deployment governance, the EU AI Act, and the broader regulatory and standards landscape that shapes AI governance in practice. We hope you have found this book valuable and wish you every success as you apply these principles and frameworks in your work and your preparation for the AIGP certification.

8

Unlock Your Exclusive Benefits

Your copy of this book includes the following exclusive benefits:







Follow the guide below to unlock them. The process takes only a few minutes and needs to be completed once.

Unlock this Book's Free Benefits in 3 Easy Steps

Step 1

Keep 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/en-us/unlock?step=1`.

> **Note**
>
> **Note**: If you bought this book directly from Packt, 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 8.1* on desktop), search for this book by name and select the correct edition.

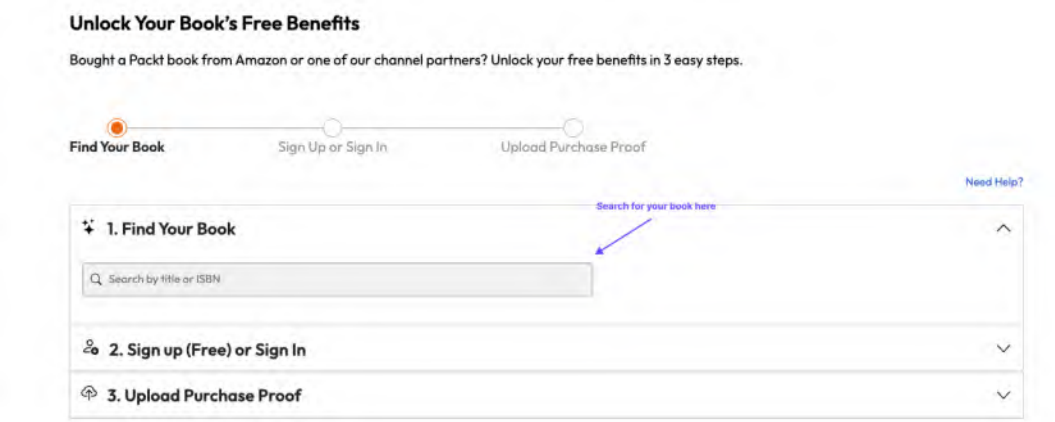


Figure 8.1: Packt unlock landing page on desktop

Step 3

After selecting your 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 invoices 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:



Driving Data Quality with Data Contracts

Andrew Jones

ISBN: 9781837636242

- Gain insights into the intricacies and shortcomings of today's data architectures
- Understand exactly how data contracts can solve prevalent data challenges
- Drive a fundamental transformation of your data culture by implementing data contracts
- Discover what goes into a data contract and why it's important
- Design a modern data architecture that leverages the power of data contracts
- Explore sample implementations to get practical knowledge of using data contracts
- Embrace best practices for the successful deployment of data contracts



Data Governance Handbook

Wendy S. Batchelder

ISBN: 9781803242477

- Comprehend data governance from ideation to delivery and beyond
- Position data governance to obtain executive buy-in
- Launch a governance program at scale with a measurable impact
- Understand real-world use cases to drive swift and effective action
- Obtain support for data governance-led digital transformation
- Launch your data governance program with confidence

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 *Artificial Intelligence Governance Professional (AIGP) Certification Guide*, 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/1808083318`

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 Risk Management Framework (AI RMF) 1.0 **10**

AI development lifecycle
- business problem, defining 64, 65
- governing 63, 64
- impact assessment, conducting 66 – 68
- navigating 64
- scope, determining 65, 66
- stakeholders, engaging 68
- use cases, defining 64, 65

AI governance
- comprehensive approach, applying 11
- defining 9, 10
- principles, distinguishing from frameworks 10

AI governance infrastructure
- building 44
- leadership support, securing 49
- policies, building 46, 47
- roles and accountability 45, 46
- roles and responsibility 47, 48
- structures and models 49, 50
- tailoring 44, 45

AI harms **25**
- bias and discrimination 28, 29
- identifying 26
- impact 27, 28
- mapping, with taxonomy approach 27
- organizational harm, assessing 29, 30

AI infrastructure
- managing 19

AI literacy and risk classification principles
- AI systems, classifying with four-tier risk framework 118, 119
- applying 117
- role-based AI literacy for safe deployment, developing 117, 118

AI management system
- AIMS improvement 146
- building, in ISO/IEC 42001 145, 146
- responsibilities, defining 146

AI model **13**
- decision tree models 13
- language models 13
- neural networks 13
- statistical models 13
- types 12

AI operational monitoring
- managing 19

AI pipeline data
- data lineage and provenance, tracing 77, 78
- governance questions, applying 76, 77
- governing 75
- with five V's 80, 81
- with quality and format controls 78 – 80

AI platforms
- managing 19
- use cases 19

AI providers oversight and monitoring
- defining 143

AI providers trustworthiness and documentation
- embedding 142

AI risk assessments
- conducting 54
- designing 54
- scalable AI processes, building 55
- scoring, with probability and severity 57, 58
- via context and outcomes 56, 57

AI risk management
NIST ARIA program for model evaluation, using 60, 61
NIST frameworks, applying 58
AI risks organizations
assessing 30
business risks 32, 33
privacy risks, addressing in AI systems 31, 32
recognizing 33, 34
security and operational risks, managing 30, 31
AI system impact assessment (AIA) **147**
AI systems
autonomy 3
baseline performance, assessing 70
communication plan, developing 74, 75
documentation, maintaining 73
fairness and safety, testing 70
objective 2
operational readiness, validating 71
output 3
post-launch, managing 71, 72
post-launch, monitoring 71, 72
robustness, assessing 70
stability, assessing 70
technology 2
testing 70
training, with machine learning fundamentals 15
validating 70
AI technology stack **20**
Artificial General Intelligence (AGI) **7**
Artificial Intelligence Management System (AIMS) **145**
Artificial Narrow Intelligence (ANI) **7**
Artificial Superintelligence (ASI) **7, 8**
Automated Employment Decision Tool (AEDT) **137**
accountability and transparency
AI systems, auditing 104, 105
AI, managing 102, 103
disclosure and transparency obligations 101, 102
ensuring 101
aggregation **82**
artificial intelligence (AI) **12**
as sociotechnical system 3, 4
classifying, with OECD dimensions 4 – 6
defining 2, 3
modern drivers, recognizing 6, 7
use cases 8

C

classification **14**
cloud computing **6**
compute infrastructure **21**
conformity assessment **127**
conformity documentation **84**

D

data drift **22**
data integrity **22**
data lineage **77**
data minimization **76**
data provenance **77**
decision tree models **13**
deep learning **13**
deployer **116**
deployment architectures and model patterns
agentic AI architectures, govering 109
audit trails and recovery plans, building 111
deployment environment, selecting 105
human oversight and confirmation gates, implementing 110
model use pattern, selecting 107, 108

prompt injection, defending 110
retrieval-augmented generation (RAG), deploying 108, 109
selecting 105
tool access and permissions, controlling 110

deployment environment
cloud deployment 106
edge deployment 106, 107
on-premises deployment 106

deployment options
versus model relationships 17

deployment readiness 88
privacy and security policies, evaluating 88, 89
privacy and security policies, updating 88, 89
requirements 89, 90

differential privacy 82

distributor 116, 121

E

EU AI Act, scope and actors 113
deployer, identifying 115, 116
distributor, identifying 115, 116
exemptions, defining to remove regulatory obligations 115
importer, identifying 115, 116
provider, identifying 115, 116
to governance practice 114

employment and consumer AI controls
automated hiring tools auditing, under NYC local law 144 137
high-risk AI and algorithmic discrimination, managing under Colorado law 139, 140
Illinois disclosure requirements, applying for AI video interviews 138, 139
implementing 137

enforcement 129, 130

ethical principles and frameworks
applying 34
ethical issues, identifying 36, 37
fair information practice principles, applying 34, 35
OECD AI principles, working with 35, 36

expert systems 18
exploring 16

G

general-purpose AI (GPAI) models
assessments, conducting 126
governing 124
obligations and systemic risk, managing 125, 126
technical documentation, producing 126

generative AI 13

global AI regulatory landscape
AI governance, building 135, 136
alignment and dissonance 134, 135
navigating 133

H

high-risk AI obligations
Article 50 transparency obligations, for limited-risk AI systems 122, 123
for importer and distributor 121, 122

high-risk AI systems
governing, as deployer 120
governing, as provider 120

human oversight
designing 128, 129

I

International Electrotechnical Commission (IEC) 133

International Organization for Standardization (ISO) 133

impact 68

impact assessment 127

importer 116, 121

influence 68

international AI governance frameworks and ISO standards
AI management system, building in ISO/IEC 42001 145, 146
AI system impact assessments, conducting with ISO/IEC 42005 147, 148
applying 143
shared AI terminology, with ISO/IEC 22989 144, 145
South Korea's AI law 144

L

labeled data **15**
language models **13**

M

machine learning (ML) **12**
fundamentals 12
model types, exploring 14
machine learning models
reinforcement learning 14
semi-supervised learning 14
supervised learning 14
unsupervised learning 14
minimal-risk AI systems
governance maintenance 123, 124
model card **84**
model cards and conformity records
creating 85
features, designing for responsible model development 83, 84
governing 83
model relationships
versus deployment options 17
model use pattern
selecting 107, 108
monitoring **22**
multimodal models **17**
multimodal systems
exploring 16

N

NIST AI Risk Management Framework (AI RMF 1.0) **136**
state privacy rules, translating into product controls 136, 137
NIST AI Risk Management Framework (AI RMF) **58**
applying 58 – 60
NIST frameworks
applying, for AI risk management 58
National Institute of Standards and Technology (NIST) **10**
natural language processing (NLP) **16**
neural networks **13**

O

Organization for Economic Co-operation and Development (OECD) **2**
observability **22**
operational compliance
building 125
operational risk
incident response and downstream consequences, managing 100, 101
monitoring 96 – 99
periodic performance and safety assessments, conducting 97, 98
responding 96

P

Plan-Do-Check-Act (PDCA) **146**
penalty structures **129, 130**
privacy principles and governance best practices
applying, in AI lifecycle 141
embedding 140

providers and deployers recommendations, implementing 142
provider 116
pseudonymization 82
purpose limitation 76

Q

quality management
designing 128, 129

R

record-keeping for traceability and auditing
implementing 127
regression 14
reinforcement learning 14
responsible AI
building 38
ethical AI culture, building 39
ethical risks, mitigating with foundational controls 38
operationalizing 38, 41
trustworthy AI, achieving 40
trustworthy AI, defining 40
retrieval-augmented generation (RAG)
deploying 108, 109
risk tiers
compliance obligations, navigating 119
prohibited AI systems, recognizing 119, 120
prohibited AI systems, responding 119, 120

S

Strong AI 7
secure access controls 82
semi-supervised learning 14
stakeholders
AI and risk strategy, aligning 53, 54
aligning, with AI lifecycle 51
engaging 51, 52
engaging, with AI lifecycle 51
identifying 51, 52
training and programs, designing 52, 53
statistical models 13
supervised learning 14
synthetic data 82
systemic risk 125

T

technical documentation
maintaining 126
preparing 126
transformer models
exploring 16
transformer-based models 16
transformers 16
transparency
designing 128, 129

U

unsupervised learning 14

V

vendor and third-party dependencies
deployment governance, applying 95, 96
governing 91
proprietary AI models, governing 91, 92
third-party risk, managing 92, 93
vendor and open-source agreements, evaluating 94

W

Weak AI 7