Alliance for Qualification

Release Version 1.00

July 2025

A4Q Syllabus

AI Essentials for the EU AI Act

A4Q Alliance for Qualification

Copyright Notice and Disclaimer

Revision History

Version	Date	Remark
1.00	01.07.2025	Intern geprüftes Dokument (V1.000)

Disclaimer

The content of this syllabus has been carefully compiled and is based, among other sources, on the legal regulations available at the time of writing, particularly the General Data Protection Regulation (GDPR) and the EU AI Act.

This syllabus is intended solely for general informational purposes. The relationships and concepts presented may be simplified, potentially misleading, taken out of context, or formulated imprecisely.

This syllabus does not constitute legal advice and cannot replace a thorough legal review tailored to individual circumstances.

For the implementation of specific legal requirements or the assessment of concrete cases, it is strongly recommended to consult a qualified attorney or an appropriately authorized legal advisory service.

The authors expressly assume no liability for the timeliness, completeness, or accuracy of the information provided, nor for any consequences resulting from its use.

Note on AI Assistance

Parts of this syllabus were created or refined with the support of artificial intelligence tools. All content has been reviewed and curated by human experts to ensure relevance and clarity.

Content

Content

Release Version 1.00

This document was officially published by A4Q GmbH on July 1, 2025.

0 Preamble

0.1 Purpose of the Syllabus

This syllabus defines the intended learning objectives and foundational concepts of the introductory AI training course for personnel within organizations that develop, operate, or procure AI systems. The course is specifically designed to meet the requirements of the EU Artificial Intelligence Act—particularly Article 4(4), which mandates the promotion of competence and awareness in the use of AI.

The content is intended for both technical and non-technical employees and serves as a foundational module with a focus on essential principles.

0.2 Examination-Relevant Learning Objectives and Cognitive Knowledge Levels

The learning objectives (LOs) outlined in this syllabus form the basis for the final assessment. Passing the examination confirms successful participation in the training and the achievement of the defined learning objectives. Participants are expected to be able to recall relevant content of each chapter, demonstrate comprehension, and analyze simple scenarios.

The cognitive knowledge levels associated with each learning objective are indicated at the beginning of every chapter and are defined as follows:

- **K1 – Remember:**
  Recall and reproduce facts, terms, and information accurately.
- **K2 – Understand:**
  Grasp, interpret, and explain meanings and relationships.
- **K3 – Apply:**
  Use acquired knowledge in concrete situations or to solve specific tasks.

0.3 The "AI Essentials" Examination

The AI Essentials examination is based on the contents of this syllabus. Answers to examination questions may require the application of knowledge drawn from multiple sections of the syllabus. All chapters are considered examination-relevant, with the exception of the preamble, optional chapters, and appendices. External sources, such as standards, books, and legal texts, may be referenced, but their content is only considered examinable to the extent that it is summarized within this syllabus.

The examination consists of 16 multiple-choice questions, with one question assigned to each learning objective. Each correct answer is awarded one point. A minimum score of 75% (i.e., at least 12 correct answers) is required to pass the exam. The total time allowed for completion is 20 minutes.

0.4 Accreditation

Training providers wishing to offer courses for the Certified AI Essentials program should be accredited by A4Q and can use the official training materials.

0.5 Level of detail

This syllabus consists of the following components:

- **Business objectives**, which define what participants should be able to achieve after completing the learning unit in order to effectively address business challenges,
- **General learning objectives**, which describe the overall goals of the certification,
- Often accompanied by a **list of key terms** that participants are expected to be able to explain,
- **Learning objectives for each area of knowledge**, describing the required level of cognitive achievement,
- A **description of the corresponding core concepts**, including references to authoritative sources such as recognized literature or standards.

The content of this syllabus does not represent an exhaustive overview of the entire domain of AI compliance or the EU AI Act. Rather, it reflects the **level of detail** appropriate for training and certification at the AI Essentials level.

This syllabus does not include specific learning objectives related to individual process models or development methodologies in software or AI engineering.

0.6 Structure of This Syllabus

The syllabus is primarily designed FOR a web-based self-paced training program. However, delivery by a qualified instructor may further enhance learning outcomes.

There are six chapters containing examination-relevant content. For each chapter, an estimated duration is provided. No additional time breakdown is given within the sub-sections of each chapter.

According to the official A4Q AI Essentials curriculum, the minimum total duration of the course is **one hour**, distributed across the six exam-relevant chapters. Since the course is intended to be delivered online, the actual time spent may be extended as needed based on individual learning preferences.

Chapter 1: Introduction to AI – 10 minutes
Chapter 2: Overview of the EU AI Act – 10 minutes
Chapter 3: Transparency and Disclosure – 10 minutes
Chapter 4: GDPR Core Principles – 10 minutes
Chapter 5: Human Oversight – 10 minutes
Chapter 6: Risks and Ethics – 10 minutes
Optional – Chapter 7: Introduction to Prompt Engineering – 15 minutes

0.7 Business Outcomes (BOs)

Chapter 1 – Fundamentals of Artificial Intelligence

BO 1.1

Participants are prepared to understand foundational concepts, general functions, and typical areas of application for artificial intelligence (AI) in professional settings.

BO 1.2

Participants are equipped to distinguish between rule-based and learning-based AI systems and gain awareness of their significance within the EU AI Act's risk-based regulatory framework.

Chapter 2 – Overview of the EU AI Act

BO 2.1

Participants become familiar with the objectives, scope, and intent of the EU AI Act, particularly in relation to its regulatory aims and structure.

BO 2.2

Participants are prepared to classify AI systems into the four risk categories defined in the EU AI Act—prohibited, high-risk, limited-risk, and minimal-risk, and to understand the key requirements that apply to each.

Chapter 3 – Transparency and Disclosure

BO 3.1

Participants are enabled to identify AI use cases within their organization that are subject to transparency requirements and to support their clear and compliant communication, for instance, in chatbots, AI-generated content, or biometric technologies.

BO 3.2

Participants develop an understanding of the legal obligation to disclose the use of AI systems and can contribute to designing transparent interfaces, user guidance, and communication practices in line with regulatory expectations.

Chapter 4 – Core Principles of the GDPR

BO 4.1

Participants are introduced to the core principles of data protection under the General Data Protection Regulation (GDPR) and become aware of categories of data that require heightened protection. They are prepared to support privacy-conscious AI practices, minimize unnecessary data collection, and help reduce compliance and reputational risks.

Chapter 5 – Human Oversight

BO 5.1

Participants recognize the importance of human oversight in the use of AI systems and are prepared to support responsible involvement in the monitoring and supervision of AI applications within their roles.

Chapter 6 – Risk and Ethics

BO 6.1

Participants are sensitized to the potential risks of bias and discrimination in AI systems and are encouraged to support fair, ethical, and legally compliant decision-making within the organization.

Optional Chapter 7 – Introduction to Prompt Engineering (Not subject to examination)

BO 7.1

Participants are introduced to the concept of prompt engineering and are encouraged to formulate effective, context-aware prompts that support meaningful and efficient interaction with AI systems in a business context.

0.8 Abbreviations

The following list provides commonly used abbreviations throughout this syllabus, along with their full definitions and contextual relevance to the EU AI Act, AI governance, and AI applications

Abbreviation	**Definition**
A4Q	Alliance for Qualification. The copyright holder and publisher of this syllabus.
AI	Artificial Intelligence. Explicitly defined as "Artificial Intelligence – abbreviated as AI."
EU AI Act	European Union Artificial Intelligence Act. The world's first comprehensive regulatory framework dedicated solely to AI technologies.
GDPR	General Data Protection Regulation. The primary legal framework for data protection in the European Union since May 2018.
LOs	Learning Objectives. These define the intended learning outcomes of the introductory course.
K1 – Remember	Recall and reproduce facts, terms, and information accurately.
K2 – Understand	Grasp, interpret, and explain meanings and relationships.
K3 – Apply	Apply acquired knowledge in concrete situations or to solve specific tasks.
BOs	Business Outcomes. These define what participants should be able to achieve after completing the learning unit.
ML	Machine Learning. A core category of AI systems that learn from data and recognize patterns.
NLP	Natural Language Processing. A specialized AI technique for processing and understanding human language.
GPAI	General Purpose AI. Regulatory obligations for these systems begin in mid-2025.
SMEs	Small and Medium-Sized Enterprises. The EU AI Act aims to support innovation in these companies.
HITL	Human-in-the-Loop. A human explicitly reviews and approves each AI decision.

Abbreviation	Definition
HOTL	Human-on-the-Loop. The AI operates autonomously but is continuously monitored by a human.
HOOTL	Human-out-of-the-Loop. The system acts fully autonomously, without human intervention—prohibited for high-risk AI.
HR	Human Resources. Used in contexts like recruitment or HR-related AI tools.
MDR	Medical Devices Regulation. EU regulation relevant to AI in medical contexts.
ISMS	Information Security Management Systems. Full title of ISO/IEC 27001:2022.
PIMS	Privacy Information Management System. Full title of ISO/IEC 27701:2019.
ISO/IEC	International Organization for Standardization / International Electrotechnical Commission. Developers of referenced international standards.
DIN SPEC	German Institute for Standardization – Specification. A type of standard issued by DIN.
TTDSG	Telecommunications and Telemedia Data Protection Act (Germany). A German legal framework for telecommunication data.
BDSG	Federal Data Protection Act (Germany). A national law aligned with the GDPR.
DSG	Data Protection Act (Austria). A national data protection law.
revFADP	Federal Act on Data Protection (Switzerland), revised version. Swiss national data protection legislation.

1 Introduction to the Topic of Artificial Intelligence (AI)

Duration: approx. 10 minutes

- LO 1.1 K1 Recall the definition of Artificial Intelligence and identify differences between AI and traditional software.
- LO 1.2 K2 Differentiate between rule-based and machine learning AI systems based on their characteristics.
- LO 1.3 K2 Explain the importance of this distinction in relation to the EU AI Act's risk-based classification

1.1 What is Artificial Intelligence (AI)?

Artificial Intelligence – abbreviated as AI – refers to the capability of machines and computer systems to perform tasks that typically require human intelligence. These tasks include, but are not limited to, speech recognition, natural language understanding, decision-making, image and pattern recognition, and learning from experience.

In other words, AI systems can acquire, analyze, and interpret information, and then draw conclusions or make recommendations – often faster and at greater scale than humans.

AI is an umbrella term for various technologies – ranging from simple rule-based systems to complex learning systems (e.g., neural networks) trained on vast amounts of data.

1.2 Examples of AI in Everyday Life

AI is already a part of our personal and professional lives. Common applications include:

- **Voice assistants** such as Siri, Alexa, or Google Assistant – for scheduling, information lookup, or dictation at work.
- **Navigation systems** with real-time traffic data – useful for field services, logistics, and route planning.
- **Recommendation systems** on platforms like YouTube, Netflix, or Amazon – similar to personalized learning tools, CRMs, or product suggestions in e-commerce.
- **Chatbots in customer service** – used to answer inquiries, pre-filter support tickets, or enable 24/7 self-service.

- **Facial recognition** – for example, for access control to buildings or secure workplace logins.
- **Credit scoring and risk assessment** – in banking and insurance, to evaluate applications and creditworthiness.
- **AI-powered document classification** – for automated capture, tagging, and processing of invoices, contracts, or emails.
- **Application screening and resume analysis** – to preselect candidates in HR departments.
- **Predictive maintenance** – to anticipate maintenance needs in industrial production.
- **Sales and revenue forecasting** – to support data-driven decision-making in sales and controlling.
- **Text generation via generative AI** – for product descriptions, email drafts, or social media posts in marketing.
- **AI-based translation tools** – for automated text translation in international communication or document processing.

These examples demonstrate how AI already influences decisions – often without users even realizing it. This underscores the importance of building basic understanding.

1.3 Why Should You Understand AI?

Understanding how AI works is essential for using it in a legally compliant, ethical, and secure manner within organizations. This is especially true in the context of the **EU AI Act**, which imposes specific requirements on AI usage.

A foundational understanding helps you:

- Evaluate risks
- Implement appropriate monitoring or documentation measures
- Recognize when external expertise is required

1.4 Two Main Types of AI Systems

In this foundational training, we distinguish between two core types:

1. Rule-Based AI
These systems operate based on predefined if-then rules.
Example:

"If expenses exceed €150, forward for approval."

Characteristics:

- No learning from data
- Transparent and understandable logic
- Low complexity
- Easy to test and document

Advantage: Transparent and reliable when rules are clear

Disadvantage: Inflexible in novel or unforeseen situations

2. Machine Learning (ML)

These systems learn from data – they recognize patterns and derive decisions without needing every rule predefined.

Example:

A system analyzes resumes and learns which characteristics correlate with successful hires.

Types of ML:

- **Supervised learning:**
  Trained with labeled examples, e.g., resumes labeled as "hired" or "rejected."
- **Unsupervised learning:**
  Discovers patterns in unlabeled data, e.g., customer segments based on buying behavior ("bargain hunters," "brand loyalists").
- **Reinforcement learning:**
  Learns through trial and error with rewards, e.g., a robot improving its grip on objects.

Characteristics:

- Capable of learning and high-performing
- Highly flexible
- Often lacks transparency ("black box")
- Dependent on data quality and volume

1.5 Why Is This Distinction Important for the EU AI Act?

Different levels of risk potential

Rule-based systems (e.g., "If A, then B") are transparent, predictable, and testable. Their behavior is deterministic.

Machine learning (ML) systems autonomously detect patterns and make decisions that may not be fully explainable. This introduces risks such as:

- Opaque decision-making logic (black box)
- Bias or discrimination from flawed training data
- Difficulty in auditing or reproducing results

→ **ML systems are typically considered higher risk and more likely to fall under high-risk categories.**

Obligations are often indirectly linked to the type of AI
Although the EU AI Act classifies systems based on use case and risk, not technology, in practice:

- ML-based systems used in areas like HR, lending, healthcare, or law enforcement are often deemed **high-risk**
- Rule-based systems with the same function may be classified as **lower-risk** due to their explainability and controllability

Conclusion: Organizations using ML-based systems must anticipate more regulatory scrutiny, greater documentation burdens, and more rigorous compliance procedures. A solid understanding of these distinctions is key for risk classification, system evaluation, and regulatory compliance.

1.6 Overview of Additional AI Technologies

Beyond the two basic types, there are specialized forms of AI:

- **Deep Learning:** Multi-layer neural networks – especially powerful for speech and image recognition
- **NLP (Natural Language Processing):** Processes and understands human language (e.g., chatbots, translation tools)
- **Computer Vision:** Interprets content in images and videos
- **Generative AI:** Creates new content such as text, images, or music (e.g., ChatGPT, DALL·E)

1.7 What Does This Mean for Organizations?

Organizations must be able to distinguish:

- What kind of AI they are using

- How transparent and controllable its decisions are
- Whether it qualifies as a **high-risk system under the EU AI Act**
- Whether any application falls under **prohibited practices** such as manipulation, social scoring, or unjustified emotional surveillance

Understanding these basics is particularly critical when **developing or procuring** AI systems – to avoid liability, reputational damage, or legal violations.

1.8 Summary

- AI refers to systems that mimic human-like behavior
- There are **rule-based systems** (fixed logic) and **learning systems** (machine learning)
- Machine learning offers more flexibility but less transparency – which is why it is tightly regulated under the EU AI Act
- Knowledge of these fundamentals is key for compliance, risk management, and governance

2 Overview of the EU AI Act

Estimated duration: approx. 10 minutes

Learning Objectives (LOs)

LO 2.1 (K1) Identify the objectives and fundamental principles of the EU AI Act.

LO 2.2 (K2) Describe the scope of application of the EU AI Act for providers, deployers, and international contexts.

LO 2.3 (K2) Explain the definition of an AI system as specified in the EU AI Act.

LO 2.4 (K2) Distinguish between the four AI risk categories defined in the EU AI Act and describe their key differences

2.1 Introduction

The **European Union Artificial Intelligence Act (EU AI Act)** is the **first comprehensive legal framework in the world** dedicated exclusively to the regulation of Artificial Intelligence technologies. It was officially adopted by the European institutions in May 2024 and entered into force in June 2024.

The main intent behind the EU AI Act is to **balance technological innovation with the protection of core European values**, including human rights, safety, transparency, and democratic integrity. By regulating the development and use of AI systems in a proportional and risk-sensitive manner, the EU seeks to ensure that AI remains a force for good in society and the economy.

This regulation introduces a **harmonized set of rules** applicable across all EU Member States, thereby creating a **common legal ground** for both public and private stakeholders who develop, deploy, import, or interact with AI systems. It aims to ensure that AI is used responsibly and that citizens and businesses can trust that AI systems meet high safety and ethical standards.

2.2 Objectives of the EU AI Act

The Act outlines four fundamental objectives that guide its structure and provisions:

1. **Protection of fundamental rights**:
   The EU AI Act seeks to prevent the use of AI in ways that violate privacy, cause discrimination, or result in arbitrary decision-making without human intervention. It is explicitly designed to prevent harm to individuals and groups, particularly in cases where decisions significantly impact legal rights, freedoms, or human dignity.

2. **Fostering trust through transparency**:
   The regulation promotes the responsible and explainable use of AI systems. It requires that users be informed when they are interacting with AI and that high-risk systems be auditable and understandable. This transparency is crucial to building public confidence in AI and encouraging its broader acceptance.

3. **Legal certainty across the EU**:
   By offering a clear and uniform set of obligations, the EU AI Act eliminates the risk of fragmented national laws and establishes a **consistent legal environment** for businesses operating in more than one EU country. This harmonization is essential for fostering a competitive and integrated internal market for trustworthy AI.

4. **Support for innovation and competitiveness**:
   The EU AI Act is not designed to hinder AI development, but rather to promote **responsible innovation**. It seeks to ensure that businesses, including small and medium-sized enterprises (SMEs), can develop AI systems that are both competitive in global markets and aligned with European values. Proportional requirements and regulatory sandboxes are provided to help foster innovation while minimizing risk.

Importantly, the regulation follows a **risk-based approach**, meaning that regulatory obligations increase as the potential risks to individuals and society grow. This ensures a tailored and proportionate legal response to the wide variety of AI use cases.

2.3 Scope – Who Is Affected and When?

The scope of the EU AI Act is intentionally broad and **cross-sectoral**, covering both EU-based and non-EU actors. The regulation applies to all entities that place AI systems on the EU market, use AI within the EU, or affect EU users through AI deployed remotely.

The following actors are within the regulation's scope:

- **Providers** of AI systems, i.e., any natural or legal person who develops an AI system or places it on the EU market under their name or trademark
- **Importers and distributors** of AI-enabled products entering the EU
- **Deployers or users** who use AI systems internally or within their operations
- **Providers of AI services**, including cloud-based or embedded solutions
- **Public institutions**, such as government agencies, schools, universities, and public health authorities
- **Private enterprises** of all sizes, including micro-enterprises and multinational corporations

- **Non-EU companies**, if their AI systems are used in the EU or impact EU citizens or residents

This means that **extraterritorial application** applies. AI systems developed outside the EU must still comply with EU requirements if they are marketed, distributed, or used within the EU.

Implementation Timeline (as of June 2025):

Date	**Regulatory Milestone**
June 2024	Entry into force of the EU AI Act
End of 2024	Immediate prohibition of unlawful AI practices (e.g., social scoring, manipulative systems)
Mid-2025	First obligations for General Purpose AI (GPAI) systems, including transparency and documentation requirements
Mid-2026	Obligations for high-risk AI systems become enforceable: conformity assessment, technical documentation, human oversight
2027	Full enforcement: market surveillance, penalties, reporting obligations, registration in the EU AI database

2.4 What Is an AI System According to the EU AI Act?

The EU AI Act defines an **AI system** as:

"Software that is developed with one or more of the techniques and approaches listed in Annex I and that can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing environments it interacts with."

The regulation is **technology-neutral**, meaning it includes a wide variety of technical approaches such as:

- **Machine learning**, including deep learning, reinforcement learning, and neural networks
- **Logic-based and symbolic approaches**, such as expert systems and rule-based engines
- **Statistical techniques**, including Bayesian estimation and decision trees

This open definition allows the regulation to remain flexible and applicable to **future advancements**, such as hybrid AI systems, large foundation models, and multi-modal architectures. The focus is on functionality and impact, not just the method of implementation.

2.5 The Risk-Based Regulatory Model – Four Risk Categories

The EU AI Act uses a **tiered system of classification** to determine which obligations apply based on the potential harm posed by an AI system. The level of regulation increases with the risk to individuals, public safety, and fundamental rights.

1. Prohibited AI Practices (Unacceptable Risk)

Certain AI practices are completely **banned** in the EU due to their incompatibility with EU values and their threat to democracy, privacy, and personal dignity.

Examples of prohibited systems include:

- **Social scoring** systems that evaluate individuals' behavior for discriminatory purposes, particularly when used by public authorities
- **Manipulative AI** designed to exploit behavioral vulnerabilities of individuals, such as through psychological profiling or addictive design
- **Emotion recognition in workplaces or educational settings**, where it may influence hiring or grading
- **Real-time remote biometric identification** in public spaces, such as live facial recognition, except in strictly defined law enforcement exceptions

These practices are considered unacceptably dangerous and are not allowed under any circumstance, unless a narrowly defined legal exemption applies.

2. High-Risk AI Systems

High-risk AI systems are those that **significantly affect the health, safety, or fundamental rights** of individuals and are used in sensitive societal contexts.

Examples of high-risk applications include:

- AI used in **recruitment and human resources**, including resume screening or automated assessments
- Systems that influence **credit scoring** or access to financial services
- Use in **law enforcement**, including predictive policing or facial recognition in criminal investigations
- **Medical diagnostics and treatment planning**
- AI used in **education**, such as grading, student profiling, or exam monitoring
- Systems used in **immigration, asylum, and border control**

Mandatory obligations for high-risk AI systems include:

- A comprehensive **risk management framework**
- Use of **high-quality, representative, and bias-free datasets**
- Clear **technical documentation** and **user instructions**
- Systems must be **auditable** and **explainable** to users
- **Human oversight** is required to intervene when necessary
- **CE conformity marking** before the system is made available on the EU market
- Ongoing **monitoring, incident reporting, and compliance evaluation**

3. Limited Risk AI Systems

These AI systems interact with individuals but do not pose significant harm. However, they can still **influence human behavior or decision-making**, and thus must meet specific transparency requirements.

Common examples include:

- **Chatbots** and virtual assistants
- **Generative AI tools**, such as image, audio, or text generation systems

Obligations for these systems focus on **transparency**, requiring that users be clearly and unambiguously informed:

- When they are interacting with an AI system
- When content (e.g., text, video, images) has been generated by AI

This allows users to make informed decisions and avoid confusion about whether they are engaging with a human or a machine.

4. Minimal Risk AI Systems

Most everyday AI systems fall into this category and are considered **low-impact** in terms of risk. They require **no legal obligations** under the Act, though ethical use is encouraged.

Examples of minimal-risk systems include:

- **Spam filters** in email systems
- **E-commerce recommendation engines**
- **Autocorrect or grammar suggestion tools**

Although not legally regulated, developers are advised to adhere to **ethical principles** such as safety, non-discrimination, and data protection.

2.6 Interaction with Other Legal Frameworks

The EU AI Act is **complementary to other existing EU regulations**, including but not limited to:

- The **General Data Protection Regulation (GDPR)**, which governs the processing of personal data
- The **Product Safety Directive**, ensuring physical safety of goods
- The **Medical Devices Regulation (MDR)**, for AI used in clinical settings
- Various **sector-specific regulations**, including those in finance, transportation, and cybersecurity

Organizations must ensure that **AI systems comply with all relevant EU laws simultaneously**, especially when personal data is involved. Dual compliance with GDPR and the EU AI Act is expected for many applications.

2.7 International Significance

The EU AI Act has **global implications**. Much like the GDPR, it sets a **global benchmark** for responsible AI regulation and will influence how AI is developed and deployed worldwide.

- **Non-EU companies** exporting AI systems to the EU must comply with its provisions
- **International regulators and lawmakers** view the EU AI Act as a reference for their own frameworks
- Countries such as **Canada, Brazil, and Japan** are already drafting similar laws inspired by the EU's approach

2.8 Criticisms and Implementation Challenges

Despite its ambitions, the EU AI Act has faced several critiques:

- The **administrative burden** on small and medium-sized enterprises (SMEs) may be disproportionately high
- **Determining the correct risk classification** for complex or adaptive systems can be difficult in practice
- There are concerns that **overregulation could stifle innovation**, especially in dynamic sectors
- The Act currently lacks detailed **implementation guidelines**, which are still under development

2.9 Opportunities and Positive Outcomes

On the other hand, the EU AI Act presents several **strategic advantages** for businesses and regulators:

- It creates a **trusted AI ecosystem** that prioritizes user safety and fundamental rights
- **Compliance becomes a competitive differentiator**, opening doors to public procurement and sensitive sectors
- It ensures **legal clarity and consistency**, reducing uncertainty in AI development and procurement
- Organizations benefit from **reduced liability and reputational risk** through structured compliance
- It lays the groundwork for **responsible innovation**, supported by guidance, standards, and best practices

2.10Conclusion

The EU AI Act represents a **historic step forward in global technology governance**. For organizations developing or using AI, it demands:

- **Immediate review of prohibited AI practices**
- **Gradual adaptation to high-risk obligations**, especially by 2026
- **Strong internal governance, documentation, and transparency measures**

Starting early will help minimize compliance costs and reduce long-term legal and operational risks.

3 Transparency and Disclosure

Estimated duration: approx. 10 minutes

Learning Objectives (LOs):

- **LO 3.1 (K1)** Identify the role and importance of transparency as defined in the EU AI Act.
- **LO 3.2 (K1)** List transparency obligations (e.g., interaction, generative AI) under the EU AI Act.
- **LO 3.3 (K2)** Describe the transparency requirements applicable to providers and users of AI systems.

3.1 Introduction

Transparency is one of the core principles of the EU AI Act and a critical condition for the trustworthiness of Artificial Intelligence in both private and professional contexts. Individuals must be able to recognize when an AI system is involved in communication, decision-making, or content creation in order to make informed choices, evaluate risks, and assert their rights.

The EU legislator recognizes that many AI systems operate in ways that are not immediately recognizable as non-human by the average user. As a result, the Act mandates that both **providers** (developers and suppliers) and **deployers** (organizations using AI) disclose the use of AI in a clear, timely, and understandable manner whenever AI plays a functional role in a system.

The purpose of these transparency requirements is not only to prevent deception but also to enable autonomous, informed user behavior. People must be able to understand what the AI system is doing, the basis on which it is operating, and the potential consequences of its outputs.

Transparency is particularly important in systems that:

- Interact directly with humans, such as chatbots, virtual assistants, or automated dialogue systems.
- Generate content that could be mistaken for human-created text, images, or videos (e.g., through generative AI).
- Analyze or classify emotional states or biometric data (e.g., via voice, facial expressions, or body language) – classified as a high-risk system under the EU AI Act.

Without clear disclosure, users may not realize they are interacting with AI, or may ascribe to it a level of authority, empathy, or neutrality that it does not possess.

The overarching objectives of the transparency rules are to:

- **Prevent misleading experiences**, especially where AI mimics human communication or creates indistinguishable content.

- **Protect individual autonomy**, ensuring users can make informed decisions.
- **Promote trust** in AI technologies by fostering openness and accountability on the part of developers and operators.

Transparency under the EU AI Act is not merely a legal compliance issue — it is also a tool for ethical design and socially accepted AI deployment.

3.2 What Does Transparency Mean in the EU AI Act?

In the context of the EU AI Act, transparency refers to the obligation to inform individuals that they are interacting with an AI system — and how that interaction is taking place. This includes AI-generated content, automated decisions, and biometric or emotional analyses.

The use of AI must never be invisible in ways that could compromise users' ability to understand, consent, or respond to the interaction. When AI is not readily apparent, its use must be disclosed.

Specifically, individuals must be able to recognize:

- That they are engaging with an AI system (e.g., during a chatbot conversation).
- That certain content (text, images, videos) was generated by AI.

This duty applies whenever the AI involvement is **not obvious** to a typical user and could **alter their behavior, choices, or perception of neutrality or authority**.

3.3 Transparency Obligations at a Glance

1. Interaction with AI Systems

When is this applicable?
When an AI system is designed to appear human-like through its language, appearance, or interaction model.

Typical examples include:

- Chatbots used in customer service
- Voice assistants with human-sounding speech

Required action:
Users must be explicitly informed, in a timely and comprehensible manner, that they are interacting with an AI system.

Example disclosure:
“This conversation is being conducted by an AI system.”

Purpose:
To avoid confusion or unintentional deception, especially in sensitive contexts such as job interviews, health advice, or legal assistance.

2. Generative AI Systems

When is this applicable?
When AI generates content (text, audio, video, images) and that content is presented to users or published.

Common examples:

- Automatically generated product descriptions in e-commerce
- AI-generated profile images or CV photos
- Synthetic voices or videos used in media (e.g., deepfakes)

Required action:
The content must be **clearly labeled** as AI-generated.

Example disclosure:
“This content was automatically generated by an AI system.”

Best practices:

- Use **technical markers** like watermarks or metadata.
- Display **visible notices** on the user interface (e.g., labels, banners, tooltips).

Purpose:
To prevent deception, particularly in contexts involving opinion formation, misinformation, or content authenticity (e.g., journalism, advertising, or academic writing).

3.4 Limitations and Exceptions

In certain well-defined cases, transparency obligations may be waived:

- When **legal confidentiality** applies, such as in undercover investigations or national security operations.
- When **disclosure would undermine the intended function**, such as during certain security protocols.

- When the use of AI is **obviously recognizable**, such as with clearly animated avatars or stylized robotic agents.

General principle:
When in doubt, transparency should be prioritized to avoid misunderstanding, user complaints, or legal non-compliance.

3.5 Relationship to the GDPR

The transparency obligations in the EU AI Act **complement** the existing information duties under the **General Data Protection Regulation (GDPR)**, particularly Articles 12 to 14.

While the GDPR focuses primarily on the **protection of personal data**, the AI Act expands the requirement to include:

- **Disclosure of the technological origin** of content and decisions.
- **Clarity on AI involvement** in user interactions.
- **Support for meaningful consent** and informed decision-making, even outside traditional data processing scenarios.

Example:
If a credit application is rejected by an AI system, the user must be informed not only of the decision rationale under GDPR, but also that the decision was AI-generated, per the AI Act.

3.6 Responsibilities of Providers and Deployers

Both **providers** (those who develop or market AI systems) and **deployers** (those who implement or operate them) are responsible for ensuring transparency compliance.

- **Providers** are responsible for system design, baseline documentation, and embedded disclosures.
- **Deployers** must ensure that transparency obligations are fulfilled in real-world use cases — including technical integration and user communication.

Practical implications include:

- Designing user interfaces with clear, visible disclosure messages
- Implementing notification elements such as tooltips, audio signals, or on-screen messages
- Training staff to communicate AI usage properly
- Documenting the disclosure mechanisms and their implementation

3.7 Best Practices for Effective Transparency

To meet the EU AI Act's transparency standards, disclosures must be **accessible, understandable, and adapted to the target audience**.

Recommended measures include:

- **Clear, plain language:** Avoid legal or technical jargon. Make the disclosure easy to grasp.
- **Timing:** Disclose AI use **before** the user interacts or before analysis begins. Late notification does not fulfill the requirement.
- **Visual aids:** Use consistent visual cues (e.g., color coding, badges, banners) to support user awareness, especially in non-text environments.
- **Accessibility:** Ensure disclosures are inclusive — compatible with screen readers, high-contrast modes, and other assistive technologies.

Transparency is only effective if it leads to **awareness and understanding**, not merely legal formality.

3.8 Relevance for Different Risk Categories

Transparency requirements vary based on the **risk classification** of the AI system:

- **Limited-risk systems:**
  Transparency is **mandatory** under Article 52. These systems may only be used if users are properly and promptly informed.
- **High-risk systems:**
  In addition to basic transparency, further obligations apply under Articles 13 and 14, including explainability, user training, and risk communication.
- **Minimal-risk systems:**
  No legal requirement exists, but **voluntary transparency** is encouraged — especially to build trust and demonstrate corporate responsibility.

3.9 Summary

- Transparency means clearly informing users when and how AI is involved.
- Obligations apply **when AI use is not obvious**, and are shared by both developers and system operators.

- Transparency is both a **legal requirement** and an **ethical imperative** to promote user trust, autonomy, and accountability.

4 Connection to GDPR Core Principles

Estimated duration: approx. 10 minutes

Learning Objectives (LOs)

- **LO 4.1 (K2)** Identify what constitutes personal data and describe its particular sensitivity under data protection regulations.
- **LO 4.2 (K1)** Name the principle of data minimization in data protection and provide simple examples of its application.

4.1 Introduction

The **General Data Protection Regulation (GDPR)** has been the core legal framework for data protection in the European Union since May 2018. It governs how personal data may be collected, processed, stored, and shared, and it establishes clear rights for data subjects.

In the context of **Artificial Intelligence (AI)**, the GDPR plays a **critical role**, especially when AI systems process **personal data**—that is, any information that relates to an identified or identifiable individual. If an AI system operates using such data, the GDPR fully applies, regardless of whether the AI system was developed or deployed inside or outside the EU.

This means that **organizations developing, deploying, or offering AI systems in the EU** must comply with both the **EU AI Act** and the **GDPR**. Ensuring dual compliance is essential to minimize legal risks and maintain public trust.

4.2 What Is Personal Data?

Personal data refers to any information that relates to an identified or identifiable natural person (“data subject”). A person is considered identifiable if they can be directly or indirectly recognized by reference to identifiers such as:

- Name, postal address, or date of birth
- Email address or telephone number
- IP addresses, geolocation data
- Health or biometric data
- Online identifiers and behavioral patterns

Example:
If an AI-driven recruitment tool analyzes résumés and uses personal identifiers like names, contact information, work history, or photographs, it is processing personal data. As such, the system falls within the full scope of the GDPR.

The nature of personal data becomes especially significant when AI models learn from or make predictions about individuals based on such information, often without direct human involvement.

4.3 Core Principles of the GDPR

Article 5 of the GDPR outlines **seven foundational principles** that apply to all processing of personal data—including by AI systems. These principles ensure that data processing remains lawful, fair, and transparent.

1. Lawfulness, Fairness, and Transparency

- Data processing must have a **valid legal basis** (e.g., consent, contract, legal obligation).
- Processing must be **fair**, meaning not deceptive or overly intrusive.
- Data subjects must be **informed** about what data is being collected and for what purpose.

2. Purpose Limitation

- Personal data must be collected **for specified, explicit, and legitimate purposes**.
- Use of the data for new or secondary purposes is only allowed under strict conditions.

3. Data Minimization

- Only the **minimum necessary** personal data should be collected.
- Irrelevant or excessive data should be avoided entirely.

4. Accuracy

- Data must be **accurate**, complete, and kept up to date.
- Inaccurate data must be corrected or deleted promptly.

5. Storage Limitation

- Personal data must be kept **only for as long as necessary** for the defined purpose.
- Organizations must implement **deletion policies** or **anonymization** strategies.

6. Integrity and Confidentiality

- Data must be protected against **unauthorized access, loss, or tampering**.

- Technical and organizational safeguards (e.g., encryption, access control) are required.

7. Accountability

- Organizations must **demonstrate compliance** with all principles.
- This includes documentation, role definitions, audit logs, and staff training.

These principles are not optional—they are legally binding and apply to **all personal data processing**, whether automated or manual, and regardless of system complexity.

4.4 Data Minimization in the AI Context

The principle of **data minimization** poses particular challenges in AI systems, which often depend on large-scale data for training and optimization. However, under the GDPR, collecting "as much as possible" is **not legally justified** unless each data element is strictly necessary for the stated purpose.

Examples of responsible application:

- A customer service chatbot does **not** need to collect religious affiliation or ethnic origin to function.
- A recommendation algorithm can be designed to use **anonymized behavior patterns** instead of personal identifiers like names or email addresses.

Technical and organizational measures to enforce data minimization in AI systems:

- Limit input fields in forms to what is essential.
- **Pseudonymize or anonymize** training and operational data whenever feasible.
- Activate "**privacy by default**" settings in all interfaces and services.
- Consider using **synthetic data** or **aggregated statistics** for model development.

These actions ensure not only legal compliance but also reduce the risk of reputational damage or discriminatory outcomes.

4.5 Legal Grounds for AI Data Processing

To process personal data lawfully with AI, the organization must identify a valid **legal basis** under **Article 6(1) of the GDPR**. Commonly used legal bases in AI contexts include:

- **Consent (Art. 6(1)(a))**: The data subject has voluntarily agreed to the processing.
- **Contractual necessity (Art. 6(1)(b))**: Data is required to fulfill a service contract.

- **Legitimate interests (Art. 6(1)(f))**: The controller's interest (e.g., fraud detection) outweighs the data subject's rights, subject to a balancing test.
- **Other lawful bases:** Legal obligations (c), protection of vital interests (d), or tasks in the public interest (e).

Special categories of data, such as **health, biometric, genetic, or racial/ethnic data**, are subject to **additional protections** under **Article 9 of the GDPR**. These may only be processed under strict exceptions, such as **explicit consent** or **substantial public interest**.

In AI systems, where profiling, prediction, or inference may occur without direct user input, organizations must assess the **applicability and sufficiency** of their chosen legal ground—particularly when the data subject may not be fully aware of the processing.

4.6 Conclusion

The GDPR applies in full to any AI system that processes personal data. Organizations must take proactive steps to ensure compliance with both the **AI-specific obligations under the EU AI Act** and the **data protection principles of the GDPR**.

To achieve this, organizations should:

- Implement and regularly review the **seven GDPR principles**
- **Justify each use of personal data** with a valid legal basis under Article 6
- Apply **data minimization techniques** to avoid over-collection
- Ensure **accountability and transparency** through documentation and user communication

Failure to comply with the GDPR in AI contexts may lead not only to **regulatory penalties** (e.g., fines up to 4% of global turnover), but also to **lasting reputational harm** and loss of user trust.

5 Human Oversight

Estimated duration: approx. 10 minutes

Learning Objectives (LOs)

- **LO 5.1 (K2)** Describe the purpose of human oversight as defined in the EU AI Act.
- **LO 5.2 (K2)** Explain the intervention options and control responsibilities required for AI systems under human oversight.

5.1 Introduction

One of the central principles of the **EU AI Act** is that **AI systems must not be allowed to make fully autonomous decisions about human beings**, particularly when such decisions significantly affect fundamental rights, legal obligations, or life opportunities.

For this reason, **human oversight** is a core requirement, especially for **high-risk AI systems**. Human oversight ensures that AI operates within clear boundaries and that critical decisions remain subject to human control, judgment, and accountability.

The law aims to prevent the delegation of power to AI systems in areas where human dignity, fairness, or personal freedoms are at stake. Oversight mechanisms must be designed so that individuals can monitor, intervene in, and override AI operations when needed.

5.2 Why Human Oversight Matters

While AI systems can process data faster, more objectively, and at greater scale than humans, they are **not infallible**. Risks arise from:

- **Incorrect or biased training data**
- **Algorithmic bias** and drift
- **Unanticipated use cases** or operating conditions
- **Security vulnerabilities** or misuse

Human oversight serves as a **safeguard layer** against these risks. It is intended to:

- **Prevent harmful or unintended outcomes**
- **Protect fundamental rights and freedoms**
- **Build and maintain public trust** in the use of AI systems

This requirement is particularly relevant in domains such as recruitment, healthcare, law enforcement, finance, and education—where decisions carry serious personal consequences.

5.3 Legal Basis in the EU AI Act

Article 14 of the EU AI Act requires that **high-risk AI systems** be designed and developed in such a way that they can be **effectively overseen by natural persons**.

The regulation obligates providers to implement mechanisms that allow human operators to:

- **Monitor the AI system's performance**
- **Detect errors or unexpected behavior**
- **Intervene, modify, or shut down the system when necessary**

The **level and form of oversight** depend on:

- The **intended purpose** of the AI system
- The **importance and criticality** of the decisions being made
- The **role and responsibilities** of the human user in the operational process

The system design must ensure that humans can actually exercise their oversight duties—not just in theory, but **in practice and under real-world conditions**.

5.4 Three Forms of Human Oversight

Human oversight can be implemented in different ways, depending on the context of use, system design, and risk level. The EU AI Act recognizes **three general models**, each with distinct implications for control and accountability:

1. Human-in-the-Loop (HITL)

- A human explicitly reviews and approves each AI decision **before it is acted upon**.
- Offers **maximum control** but may **slow down decision-making**.
- **Example:** A credit application is evaluated by AI, but the final approval is made by a human loan officer.

2. Human-on-the-Loop (HOTL)

- The AI system makes decisions **autonomously**, but a human continuously monitors operations and has the ability to **intervene if necessary**.

- Balances autonomy and control but depends on the human operator’s awareness and availability.
- **Example:** A traffic control AI suggests light sequences, while a human operator supervises and can override the system during anomalies.

3. Human-out-of-the-Loop (HOOTL)

- No human intervention is possible; the AI system operates fully autonomously.
- **Prohibited for high-risk AI systems** under the EU AI Act due to the lack of accountability and transparency.
- **Example:** An AI screening tool automatically rejects job applications without any human review—this is not permissible in high-risk domains.

Only HITL and HOTL models are considered **legally acceptable** for high-risk AI, provided the human role is meaningful, informed, and feasible in practice.

5.5 Requirements for Oversight and Intervention

Human oversight must be **real, actionable, and built into system design**. It is not sufficient to state that oversight is possible if no tools or workflows are in place to support it.

Key requirements include:

- **User interfaces** that allow intervention, including pause, override, and shutdown functions
- **Logging and traceability** of AI decisions and human interventions
- **Training** for staff responsible for monitoring and control, including risk awareness and decision review
- **Clear delegation** of roles and responsibilities to ensure someone is accountable at every stage

These elements are essential to ensure that oversight is **effective** rather than symbolic.

5.6 Practical Examples of Oversight Models

Example 1: Recruitment System with HITL

An AI system ranks job applicants based on relevance. A human HR professional then **reviews the ranking and makes the final hiring decision**. The AI supports efficiency, but the human retains decision authority.

Example 2: Traffic Management with HOTL

A municipal traffic control system uses AI to manage traffic flow. The system operates autonomously but is monitored by human traffic engineers who can **intervene in case of emergencies or system errors**.

Example 3: Insurance Claim Rejection with HOOTL

An insurance company uses AI to automatically **deny claims without any human review**. This use of HOOTL would **not comply** with the EU AI Act if the system is considered high-risk, as it removes human judgment from potentially life-altering decisions.

These examples show how oversight can and must be tailored to the context and potential impact of the AI system.

5.7 Practical Limitations of Human Oversight

Even when oversight mechanisms exist, they may not be effective due to **real-world constraints**, such as:

- **High levels of automation**, which reduce opportunities for meaningful intervention
- **Time pressure or workload**, making it difficult for humans to monitor AI systems effectively
- **Lack of clarity** about who is responsible for oversight
- **Insufficient technical understanding** of the AI system (lack of explainability)
- **Opaque system design** (so-called “black box” models) that prevent meaningful review

To address these challenges, **technical, procedural, and organizational safeguards** must be implemented to ensure that oversight is **practical and enforceable**.

5.8 Conclusion

Human oversight is a **non-negotiable element of trustworthy AI**. It ensures that AI systems remain subordinate to human values, ethical standards, and legal norms. The goal is not to eliminate AI autonomy, but to ensure that it is always **subject to human accountability**.

Conditions for effective oversight include:

- Clear assignment of roles and responsibilities
- Technically feasible and timely intervention options
- Transparent system design that supports understanding and control
- Continuous training and awareness programs for personnel

The EU AI Act underscores that **AI must serve people—not replace them**. Systems must be designed to support, not supplant, human decision-making.

6 Risks and Ethics

Estimated duration: approx. 10 minutes

Learning Objectives (LOs)

- **LO 6.1 (K1)** Recall the concept of bias in the context of AI systems.
- **LO 6.2 (K1)** Name examples of unintended discrimination resulting from the use of AI systems.

6.1 Introduction

Artificial Intelligence (AI) has transformative potential for innovation, efficiency, and automation. Yet, alongside these benefits, AI systems also pose significant **ethical and legal risks**—especially when used in ways that affect people's rights or well-being.

These risks are not merely technical. They often arise from **biased data**, **non-transparent logic**, or **overly automated decision-making**, leading to unfair or discriminatory outcomes.

For this reason, the **EU AI Act** requires that AI systems—particularly **high-risk systems**—not only meet technical safety standards but also adhere to **ethical responsibilities** and **non-discrimination principles**.

This chapter focuses on the most common and critical risks: **algorithmic bias**, **discrimination**, and the lack of fairness in AI systems. It also highlights the importance of corporate responsibility and ethical design.

6.2 What Does Fairness in AI Mean?

Fairness in AI means that systems should not systematically disadvantage any individual or group based on personal characteristics such as:

- Age
- Gender
- Ethnicity
- Religion
- Disability
- Socioeconomic status

A **fair AI system** should:

- Avoid outcomes that disproportionately harm or exclude any group.
- Incorporate multiple perspectives during development.
- Identify and mitigate **pre-existing inequalities** that could be reinforced by training data or algorithms.

AI systems make decisions based on data. If that data is biased, incomplete, or unbalanced, the system may unintentionally amplify societal biases rather than correct them.

6.3 What Is Bias?

Bias refers to distortion or systematic error in how a system interprets data and makes decisions. In the context of AI, bias can emerge in several ways:

- **Data Bias:** The training data is not representative of the real-world population. For example, if women are underrepresented in the dataset, the AI system may underperform on female users.
- **Label Bias:** Human-generated labels (e.g., “successful candidate”) reflect subjective or stereotyped judgments.
- **Algorithmic Bias:** The model’s logic, weighting, or optimization goals favor certain groups over others, often unintentionally.

Bias is rarely introduced on purpose. It typically results from:

- Historical data reflecting societal inequalities
- Lack of diversity in development teams
- Technical decisions during model design and feature selection

The presence of bias does not necessarily mean a system is unlawful—but **it must be detected and** addressed to meet ethical and legal standards.

6.4 Examples of Discrimination by AI

Real-world examples illustrate how AI systems—if unchecked—can perpetuate discrimination:

- A **resume screening tool** favors male applicants because historical hiring data mostly reflects male career paths.
- A **speech recognition system** performs poorly on dialects or accented speech and rates these users as “less competent.”

- A **credit scoring model** assigns lower ratings to people with non-European names, reflecting past rejection patterns.

These cases show that even so-called **“neutral” algorithms** can have discriminatory effects if fairness is not systematically validated.

6.5 Legal Requirements

EU AI Act

The EU AI Act explicitly mandates anti-discrimination measures, especially for **high-risk AI systems**. These systems must:

- Be **tested and validated for discrimination risks**
- Use **high-quality, representative, and unbiased training data**
- Implement **fairness testing procedures**
- Ensure **explainability** and enable **human oversight** (Article 10 and Article 14)

These requirements help ensure that AI systems do not cause unlawful or unethical harm through data-driven decision-making.

General Data Protection Regulation (GDPR)

The GDPR also prohibits discriminatory outcomes, especially when decisions are made **solely by** automated **means** (Article 22). Key articles include:

- **Article 5** – Lawfulness, fairness, and transparency
- **Article 9** – Additional protections for **special categories of personal data**, such as race, religion, or health status
- **Article 22** – Right to human intervention in decisions that significantly affect individuals

AI systems that use or infer sensitive data must be designed to respect these rights and legal constraints.

6.6 Ethical Guidelines for Trustworthy AI

In addition to legal rules, the EU promotes **ethical frameworks** to ensure responsible and trustworthy AI development. The **EU Commission’s Guidelines for Trustworthy AI** outline five core ethical values:

1. **Human agency and oversight:** AI must support human decision-making, not replace it.
2. **Transparency:** AI operations should be explainable and understandable.

3. **Fairness:** Systems should prevent unjustified bias, discrimination, or exclusion.
4. **Accountability:** Roles and responsibilities must be clearly defined across the AI lifecycle.
5. **Technical robustness and safety:** Systems must be secure, reliable, and resilient to misuse or failure.

These principles help organizations go beyond formal compliance and **build public trust** in AI systems.

6.7 Corporate Responsibility in Fairness

Fairness is not just a technical feature—it is a **strategic and cultural responsibility**. Companies developing or using AI should:

- **Address ethical risks** proactively during design and deployment.
- Promote **diversity** in training data, development teams, and testing strategies.
- Establish **mechanisms for user feedback, redress, and review**.
- Raise awareness among staff regarding **social and ethical impacts** of AI.

These responsibilities must be embedded in **governance structures**, not delegated only to technical teams.

6.8 Practical Measures for Fair AI

Organizations can implement specific actions to ensure fairness and compliance:

- **Bias audits** and **fairness assessments** before launching AI products
- Involvement of **interdisciplinary teams** (e.g., legal, technical, ethics, domain experts)
- Development of **guidelines and internal standards** for data quality, labeling, and model fairness
- **Employee training programs** focused on responsible AI
- **Continuous monitoring and documentation** of system behavior
- Inclusion of **external stakeholders**, such as ethics boards, domain experts, and affected user groups

These measures demonstrate a company's commitment to building AI systems that are **not only legal but** ethically **defensible**.

6.9 Conclusion

AI systems are **not neutral**. They process data that is shaped by history, social context, and human decisions. Therefore, bias and discrimination are not theoretical risks—they are **real, measurable, and addressable**.

Key takeaways:

- Bias is common and often unintended, but must be identified and mitigated.
- Fairness can be measured technically, but must be evaluated ethically.
- Organizations are **accountable** for ensuring fair and equitable outcomes from their AI systems.

Guiding questions for responsible AI deployment:

- Is the goal of the system ethically justifiable?
- Are the data sources complete and inclusive?
- Is discrimination being tested for regularly?
- Are there mechanisms for users to appeal or challenge outcomes?

Trustworthy AI is not only about performance—it is about **responsibility, fairness, and human dignity**.

7 Introduction to Designing Effective Prompts for AI Systems (Prompt Engineering)

Note: This chapter is not subject to examination. It is provided as an **optional bonus module** to deepen practical understanding of AI systems and improve workplace application.

Estimated Duration: Approximately 20 minutes

Learning Objectives (LOs)

- **LO 7.1 (K1)** Recall the concept of prompt engineering and explain its relevance in the use of AI systems.
- **LO 7.2 (K1)** Identify different types of prompts commonly used in AI interactions.
- **LO 7.3 (K2)** Analyze the characteristics of effective prompts using examples, and identify potential areas for improvement.

7.1 Introduction – Why Is Prompt Engineering Important?

Modern generative AI systems, such as ChatGPT, Claude, and Gemini, are capable of handling a wide range of complex tasks, including:

- Writing and rewriting text,
- Analyzing and summarizing documents,
- Answering structured or open-ended questions,
- Supporting workplace decisions and project planning.

In order to perform these tasks effectively, such systems require **clear, specific, and well-structured instructions**, known as **prompts**.

A **prompt** is the input (question, task, or instruction) given to the AI. It triggers the response mechanism and guides how the system interprets and responds to the request. The practice of **prompt engineering** refers to the skill of formulating prompts that yield **relevant, high-quality, and contextually appropriate results**.

Example 1: Weak vs. Strong Prompt

- **Weak prompt:** “Tell me something about data protection.”
- **Strong prompt:** “Explain what data protection means under the GDPR, using plain language, and provide three typical workplace examples.”

The strong prompt improves output by narrowing the topic, indicating the legal context (GDPR), defining the tone (plain language), and requesting specific content (examples).

Example 2: Ambiguous vs. Clear Prompt

- **Ambiguous prompt:** "What is a contract?"
- **Clear prompt:** "Explain to a 16-year-old student what an employment contract is, and mention two rights and two obligations of the employee."

The clear version reduces interpretation errors and focuses the response on a specific audience and legal topic.

In the context of the **EU AI Act**, where explainability and user control are core principles (especially for general-purpose AI tools), prompt engineering helps ensure that **AI responses are aligned with user expectations, ethical use, and risk-mitigation goals**.

7.2 Core Principles of Effective Prompts

The quality of a prompt greatly affects the quality of the AI system's response. The following principles are essential:

1. **Clarity**
   - Use precise and unambiguous language. Avoid vague terms and generalities.
2. **Context**
   - Include background information (target audience, legal framework, industry use case). This helps the AI tailor the response appropriately.
3. **Structure**
   - Use bullet points, step-by-step instructions, or headings to guide the AI's output.
4. **Role Assignment**
   - Instruct the AI to answer from a specific role.
     Example: "You are a certified data protection officer. Explain the key risks of AI under GDPR."
5. **Constraints**
   - Limit the format or scope of the response.
     Example: "Answer in no more than 100 words." or "Provide your answer as a table."
6. **Iteration**

- Prompt engineering is a cycle: test a prompt, assess the result, refine it, and repeat until the result meets the goal.

These principles are especially useful in organizational settings where **output reliability, compliance, and reproducibility** are critical.

7.3 Typical Prompt Types with Examples

Below are common categories of prompts, along with examples:

1. **Information Query**
    - "What are the legal requirements for high-risk AI systems under the EU AI Act?"
2. **Summarization**
    - "Summarize this article in three key bullet points focused on data security risks."
3. **Role-Based Prompting**
    - "You are an HR manager. Draft a short internal announcement introducing new AI usage guidelines."
4. **Comparison Prompts**
    - "Compare the compliance obligations for limited-risk vs. high-risk AI systems under the EU AI Act. Display the result as a two-column table."
5. **Creativity and Ideation**
    - "List five innovative ways to integrate AI in public schools while ensuring data protection and fairness."
6. **Question Generation**
    - "Write five multiple-choice questions on the GDPR's rules for automated decision-making. Include answers."
7. **Stylistic Rewriting**
    - "Rewrite this legal clause using accessible language for non-lawyers."
8. **Translation and Simplification**
    - "Translate this section of the EU AI Act into plain English and explain its implications for small businesses."

These types allow prompt engineering to be used across departments—legal, HR, compliance, marketing, customer support, etc.

7.4 Common Prompting Mistakes and How to Avoid Them

- **Common pitfalls include:**
- **Overly open prompts:**
  "Tell me something interesting." → Likely to yield random or irrelevant information.
- **Ambiguous language:**
  Terms like "it", "this", or "good" without context may confuse the AI and produce inconsistent answers.
- **Lack of background:**
  Without context, the AI has to make assumptions—often incorrectly.
- **Overloading the prompt:**
  Asking for too many tasks at once (e.g., summarize, analyze, rephrase, and compare) results in poor performance.
- **No refinement:**
  Accepting poor answers without adjusting the prompt leads to inefficient or misleading outputs.

Best practice: Always review the result and iterate on your prompt.

7.5 The Prompt Feedback Cycle (4 Steps)

Prompt engineering follows a structured improvement loop:

1. **Write the initial prompt** based on your task.
2. **Review the AI's output** critically for accuracy, tone, and relevance.
3. **Refine the prompt** to remove ambiguity or adjust structure.
4. **Evaluate the revised output**, and repeat the process if needed.

This process is especially useful in **regulated domains**, where the quality of information may impact legal or business decisions.

7.6 Advanced Prompt Engineering Techniques

As users become more experienced, they can use advanced methods for more precise control over AI behavior.

1. **Chain-of-Thought Prompting**
    - Example: “First explain the problem, then outline three root causes, and finally propose a policy-level solution.”
2. **Few-Shot Prompting**
    - Provide 1–3 examples of desired output.
      Example: “Here are two model privacy notices. Now write a similar one for a company using AI in hiring.”
3. **Zero-Shot Prompting**
    - Ask a question directly, without prior examples.
      Example: “Explain what algorithmic bias is and how it can affect automated decision-making.”
4. **Meta-Prompting**
    - Ask the AI to evaluate its own output.
      Example: “From the perspective of a compliance auditor, is the previous explanation of GDPR Article 22 accurate and sufficient?”

These methods are especially useful for **knowledge workers**, **AI trainers**, and **compliance officers** seeking more consistent and credible results.

7.7 Final Practical Tips

To use prompt engineering effectively in real-world settings:

- Clearly define your **desired outcome** before writing the prompt.
- Consider the **audience** (legal team, executive, general staff).
- Test multiple prompt **versions** and choose the most effective one.
- Keep a **library of successful prompts** for reuse across teams.
- Use **structured guidance** (roles, length, tone) to maintain consistency.

Organizations should consider prompt libraries part of their **AI governance model**—especially in light of upcoming compliance requirements under the EU AI Act.

7.8 Limitations of Prompt-Based AI Systems

While prompt engineering can significantly improve the quality and relevance of AI responses, it is equally important to understand the **inherent limitations of prompt-based AI systems**. These limitations reflect the underlying architecture of current AI technologies, such as large language models, and are particularly relevant in professional and regulated environments where **accuracy, transparency, and accountability** are essential.

1. Risk of Inaccurate or Fabricated Information (Hallucination)

Even well-structured prompts can lead to incorrect or invented facts. AI systems may produce responses that sound plausible but are **factually wrong**, **outdated**, or **completely fabricated**.

- **Example:** An AI might reference a non-existent EU regulation or misstate GDPR requirements.
- **Implication:** In legal, medical, or compliance contexts, such outputs can result in real-world harm or regulatory violations.

2. Lack of True Understanding or Judgment

AI does not understand content in the human sense. It generates responses based on statistical patterns in training data, not logical reasoning or common sense.

- **Example:** The AI might rephrase a discriminatory policy without recognizing its ethical implications.
- **Implication:** AI should not be used as a source of final judgment in high-stakes decisions.

3. Sensitivity to Prompt Variations

Small changes in wording can result in dramatically different outputs. This makes consistency difficult and limits the reliability of automated workflows.

- **Example:** “List three risks of using AI in HR” may yield different results than “What risks should be avoided when using AI in HR?”
- **Implication:** Organizations must validate prompt design and standardize templates to maintain quality.

4. Limited Memory and Context Awareness

Most AI systems cannot retain long-term memory or knowledge of previous interactions unless integrated into a specialized platform.

- **Example:** The AI may contradict earlier statements in the same session or forget the user’s instructions.
- **Implication:** Users should restate key context within each prompt and avoid assuming continuity across interactions.

5. Opaque Internal Logic (Black Box Effect)

AI systems often cannot explain why a particular answer was generated. This limits their suitability in contexts that require **traceability**, **reasoning**, or **legal justification**.

- **Example:** A summary of a legal case may omit critical facts without indicating the reasoning behind the omission.
- **Implication:** The EU AI Act emphasizes transparency and explainability—limitations in this area increase the need for human oversight.

6. Ethical and Social Blind Spots

AI reflects biases in training data and does not apply ethical reasoning. Without explicit constraints, it may generate harmful, discriminatory, or misleading content.

- **Example:** The AI might unintentionally perpetuate stereotypes when generating examples involving age, gender, or nationality.
- **Implication:** Ethical review and fairness auditing are essential when using AI for public-facing content or decision support.

7.9 Conclusion

Prompt engineering is a core digital communication skill, not a purely technical or developer task. It is central to ensuring AI systems remain explainable, controllable, and context-sensitive.

Well-designed prompts empower AI to support tasks more efficiently, ethically, and transparently. For organizations that work with general-purpose AI, prompt engineering should be:

- Taught in structured training formats,
- Supported with shared internal resources, and
- Recognized as a key competence in responsible AI use.

Final takeaway:
The quality of the AI's answer is only as good as the quality of your prompt.

8 Supplementary Reading Material

Bücher

[1] Ethem Alpaydin – *Introduction to Machine Learning* (MIT Press, 2020)

[2] Virginia Dignum – *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way* (Springer, 2019)

[3] Hannah Fry – *Hello World: Being Human in the Age of Algorithms* (W. W. Norton, 2018)

[4] Kate Crawford – *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence* (Yale University Press, 2021)

[5] Stuart Russell – *Human Compatible: Artificial Intelligence and the Problem of Control* (Viking, 2019)

[6] Dan McCreary & Ann Kelly – *Making Sense of AI: A Beginner's Guide to Artificial Intelligence* (O'Reilly, 2023)

[7] Luciano Floridi – *The Ethics of Artificial Intelligence* (Oxford University Press, 2023)

[8] Pascal Finette & Jeffrey Rogers – *The Disruption Map: AI and the Future of Business* (Singularity University Press, 2022)

Reports and Studie

[9] OECD – *OECD Framework for the Classification of AI Systems*, 2022

[10] European Commission – *Impact Assessment Accompanying the Proposal for the EU AI Act*, 2021

[11] McKinsey & Company – *The State of AI in 2023: Generative AI's Breakout Year*, 2023

[12] Deloitte – *State of AI in the Enterprise: Third Edition*, 2022

[13] Boston Consulting Group – *Mind the Gap: AI Adoption in Europe's Businesses*, 2022

[14] Fraunhofer Institute – *AI in Practice: Guidelines for Introducing Artificial Intelligence in Companies*, 2022

[15] Accenture – *AI: Built to Scale – Unlocking the Value of AI Across the Enterprise*, 2021

[16] German Economic Institute – *AI Strategies in German Businesses: Success Factors and Barriers*, 2021

[17] EY – *How to Scale Artificial Intelligence Sustainably in Business*, 2023

[18] Capgemini Research Institute – *AI and the Ethical Conundrum: Managing AI Responsibly*, 2020

[19] PwC – *AI and the Economy: Strategy, Opportunities, and Risk in a European Context*, 2022

[20] SAP – *Integrating AI into Enterprise Processes: Practical Approaches and Lessons Learned*, 2021

[21] IBM – *Global AI Adoption Index 2023*, 2023

[22] Swiss Digital Initiative – *Trusted AI: Requirements for Companies in Switzerland*, 2022

[23] ETH Zurich – *Artificial Intelligence in Swiss Companies: Current State and Challenges*, 2021

[24] Austrian Federal Ministry – *Fields of Application for AI in the Austrian Economy*, 2021

[25] University of Oxford – *AI Governance: A Research Agenda*, Center for the Governance of AI, 2018

[26] Zukunftsinstitut – *Smart Business: How AI Is Redefining Value Creation*, 2022

[27] German AI Association – *Position Paper: Requirements for a Future-Oriented AI Policy for Companies*, 2022

Legal Regulations and Strategy Documents

[28] European Union – *Regulation (EU) 2024/XXXX on Artificial Intelligence (EU AI Act)*, adopted by the European Parliament on March 13, 2024; publication in the Official Journal of the EU pending
[29] European Union – *General Data Protection Regulation (GDPR), Regulation (EU) 2016/679*, in force since May 25, 2018
[30] Federal Ministry for Economic Affairs and Climate Action (Germany) – *German AI Strategy*, updated version 2020
[31] Federal Department of Economic Affairs, Education and Research (Switzerland) – *Swiss National AI Strategy*, 2020
[32] Federal Ministry for Digital and Economic Affairs (Austria) – *Austrian AI Strategy*, 2021
[33] German Federal Ministry of Justice – *Telecommunications and Telemedia Data Protection Act (TTDSG)*, in effect since December 1, 2021
[34] Germany – *Federal Data Protection Act (BDSG)*, in alignment with the GDPR, effective since May 25, 2018 (revised 2021)
[35] Austria – *Data Protection Act (DSG)*, Federal Law on the Protection of Personal Data, most recent amendment in 2018
[36] Switzerland – *Federal Act on Data Protection (revFADP)*, new version in force since September 1, 2023

Standards and Norms

[37] ISO/IEC 22989:2022 – *Artificial Intelligence – Concepts and Terminology*
[38] ISO/IEC 23053:2022 – *Framework for Machine Learning*
[39] ISO/IEC TR 24028:2020 – *Artificial Intelligence – Overview of Trustworthiness in AI*
[40] ISO/IEC 23894:2023 – *Artificial Intelligence – Risk Management*
[41] DIN SPEC 92001-1:2020 – *Quality Requirements for AI Systems – Basic Principles*
[42] DIN SPEC 92001-2:2021 – *Quality Requirements for AI Systems – Testing Criteria*
[43] ISO/IEC 27001:2022 – *Information Security Management Systems (ISMS)*
[44] ISO/IEC 27701:2019 – *Privacy Information Management System (PIMS), Extension to ISO/IEC 27001*